Functional Inference from Orthology and Domain Architecture
Mateusz Kaduk

Abstract
Proteins are the basic building blocks of all living organisms. They play a central role in determining the structure of living beings and are required for essential chemical reactions. One of the main challenges in bioinformatics is to characterize the function of all proteins. The problem of understanding protein function can be approached by understanding their evolutionary history. Orthology analysis plays an important role in studying the evolutionary relation of proteins. Proteins are termed orthologs if they derive from a single gene in the species’ last common ancestor, i.e. if they were separated by a speciation event. Orthologs are useful because they retain their function more often than other homologs.

Inference of a complete set of orthologs for many species is computationally intensive. Currently, the fastest algorithms rely on graph-based approaches, which compare all-vs-all sequences and then cluster top hits into groups of orthologs. The initial step of performing all-vs-all comparisons is usually the primary computational challenge as it scales quadratically with the number of species.

A new, more scalable and less computationally demanding method was developed to solve this problem without sacrificing accuracy. The Hieranoid 2 algorithm reduces computational complexity to almost linear by overcoming the necessity to perform all-vs-all similarity searches. The algorithm progresses along a known species tree, from leaves to root. Starting at the leaves, ortholog groups are predicted conventionally and then summarized at internal nodes to form pseudo-species. These pseudo-species are then re-used to search against other (pseudo-)species higher in the tree. This way the algorithm aggregates new ortholog groups hierarchically. The hierarchy is a natural structure to store and view large multi-species ortholog groups, and provides a complete picture of inferred evolutionary events.

To facilitate explorative analysis of hierarchical groups of orthologs, a new online tool was created. The HieranoiDB website provides precomputed hierarchical groups of orthologs for a set of 66 species. It allows the user to search for orthology assignments using protein description, protein sequence, or species. Evolutionary events and meta information is added to the hierarchical groups of orthologs, which are shown graphically as interactive trees. This representation allows exploring, searching, and easier visual inspection of multi-species ortholog groups.

The majority of orthology prediction methods focus on treating the whole protein sequence as a single evolutionary unit. However, proteins are often composed of individual units, called protein domains, that can have different evolutionary histories. To extend the full sequence based methodology to a domain-aware method, a new approach called Domainoid is proposed. Here, domains are extracted from full-length sequences and subjected to orthology inference. This allows Domainoid to find orthology that would be missed by a full sequence approach.

Networks are a convenient graphical representation for showing a large number of functional associations between genes or proteins. They allow various analyses of graph properties, and can help visualize complex relationships. A framework for inferring comprehensive functional association networks was developed, called FunCoup. A major difference compared to other networks is FunCoup's extensive use of orthology relationships between species, which significantly boosts its coverage. Using naive Bayesian classifiers to integrate 10 different evidence types and orthology transfer, FunCoup captures functional associations of many types, and provides comprehensive networks for 17 species across five gold-standards.

Keywords: Orthology, Functional coupling networks, Association networks, Hierarchical groups of orthologs.

Stockholm 2018
http://urn.kb.se/resolve?urn=urn:nbn:se:su:diva-155096


Department of Biochemistry and Biophysics
Stockholm University, 106 91 Stockholm
FUNCTIONAL INFERENCE FROM ORTHOLOGY AND DOMAIN ARCHITECTURE

Mateusz Kaduk
Functional Inference from Orthology and Domain Architecture

Mateusz Kaduk
Abstract

Proteins are the basic building blocks of all living organisms. They play a central role in determining the structure of living beings and are required for essential chemical reactions. One of the main challenges in bioinformatics is to characterize the function of all proteins. The problem of understanding protein function can be approached by understanding their evolutionary history. Orthology analysis plays an important role in studying the evolutionary relation of proteins. Proteins are termed orthologs if they derive from a single gene in the species’ last common ancestor, i.e. if they were separated by a speciation event. Orthologs are useful because they retain their function more often than other homologs.

Inference of a complete set of orthologs for many species is computationally intensive. Currently, the fastest algorithms rely on graph-based approaches, which compare all-vs-all sequences and then cluster top hits into groups of orthologs. The initial step of performing all-vs-all comparisons is usually the primary computational challenge as it scales quadratically with the number of species.

A new, more scalable and less computationally demanding method was developed to solve this problem without sacrificing accuracy. The Hieranoid 2 algorithm reduces computational complexity to almost linear by overcoming the necessity to perform all-vs-all similarity searches. The algorithm progresses along a known species tree, from leaves to root. Starting at the leaves, ortholog groups are predicted conventionally and then summarized at internal nodes to form pseudo-species. These pseudo-species are then re-used to search against other (pseudo-)species higher in the tree. This way the algorithm aggregates new ortholog groups hierarchically. The hierarchy is a natural structure to store and view large multi-species ortholog groups, and provides a complete picture of inferred evolutionary events.

To facilitate explorative analysis of hierarchical groups of orthologs, a new online tool was created. The HieranoiDB website provides precomputed hierarchical groups of orthologs for a set of 66 species. It allows the user to search for orthology assignments using protein description, protein sequence, or species. Evolutionary events and meta information is added to the hierarchical groups of orthologs, which are shown graphically as interactive trees. This representation allows exploring, searching, and easier visual inspection.
of multi-species ortholog groups.

The majority of orthology prediction methods focus on treating the whole protein sequence as a single evolutionary unit. However, proteins are often composed of individual units, called protein domains, that can have different evolutionary histories. To extend the full sequence based methodology to a domain-aware method, a new approach called Domainoid is proposed. Here, domains are extracted from full-length sequences and subjected to orthology inference. This allows Domainoid to find orthology that would be missed by a full sequence approach.

Networks are a convenient graphical representation for showing a large number of functional associations between genes or proteins. They allow various analyses of graph properties, and can help visualize complex relationships. A framework for inferring comprehensive functional association networks was developed, called FunCoup. A major difference compared to other networks is FunCoup’s extensive use of orthology relationships between species, which significantly boosts its coverage. Using naïve Bayesian classifiers to integrate 10 different evidence types and orthology transfer, FunCoup captures functional associations of many types, and provides comprehensive networks for 17 species across five gold-standards.
This thesis is dedicated to my grandfather
Franciszek Dmitrowski.
List of Papers

The following papers, referred to in the text by their Roman numerals, are included in this thesis.

PAPER I: **Improved orthology inference with Hieranoid 2**
DOI: https://doi.org/10.1093/bioinformatics/btw774

PAPER II: **HieranoiDB: a database of orthologs inferred by Hieranoid**
DOI: https://doi.org/10.1093/nar/gkw923

PAPER III: **Domainoid: Domain-oriented orthology inference**

PAPER IV: **FunCoup 4: new species, data, and visualization**
DOI: https://doi.org/10.1093/nar/gkx1138

† Contributed equally.

Reprints were made with permission from the publishers.
**Contents**

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>1</td>
</tr>
<tr>
<td>List of Papers</td>
<td>5</td>
</tr>
<tr>
<td>List of Figures</td>
<td>9</td>
</tr>
<tr>
<td>1 Introduction</td>
<td>11</td>
</tr>
<tr>
<td>2 Orthology prediction</td>
<td>13</td>
</tr>
<tr>
<td>2.1 Evolutionary relationships</td>
<td>13</td>
</tr>
<tr>
<td>2.2 Pairwise similarity</td>
<td>14</td>
</tr>
<tr>
<td>2.2.1 Smith-Waterman</td>
<td>14</td>
</tr>
<tr>
<td>2.2.2 Basic local alignment tool (BLAST)</td>
<td>15</td>
</tr>
<tr>
<td>2.2.3 Hidden Markov Models</td>
<td>15</td>
</tr>
<tr>
<td>2.2.4 Other search tools</td>
<td>16</td>
</tr>
<tr>
<td>2.3 Multiple sequence alignment</td>
<td>16</td>
</tr>
<tr>
<td>2.4 Orthology prediction</td>
<td>17</td>
</tr>
<tr>
<td>2.4.1 Tree based methods</td>
<td>17</td>
</tr>
<tr>
<td>2.4.2 Graph based methods</td>
<td>18</td>
</tr>
<tr>
<td>2.4.3 Hierarchical methods</td>
<td>19</td>
</tr>
<tr>
<td>2.5 Other orthology methods</td>
<td>20</td>
</tr>
<tr>
<td>2.6 Quest for Orthologs</td>
<td>20</td>
</tr>
<tr>
<td>2.7 Protein domains</td>
<td>21</td>
</tr>
<tr>
<td>2.7.1 Domain architecture</td>
<td>22</td>
</tr>
<tr>
<td>2.7.2 Domain similarity</td>
<td>22</td>
</tr>
<tr>
<td>2.8 Domain orthology</td>
<td>23</td>
</tr>
<tr>
<td>2.9 Methods and databases</td>
<td>24</td>
</tr>
<tr>
<td>3 Association networks</td>
<td>27</td>
</tr>
<tr>
<td>3.1 Networks definition</td>
<td>27</td>
</tr>
<tr>
<td>3.2 Network methods</td>
<td>27</td>
</tr>
<tr>
<td>3.2.1 Bayes’ theorem</td>
<td>28</td>
</tr>
</tbody>
</table>
List of Figures

2.1 Example showing orthologs, inparalogs and outparalogs. . . . 14
2.2 Profile hidden Markov model of sequence which models emission probabilities for match (M), insertion (I) and deletion (D) states, as well as transition probabilities. . . . . . . . . . . . . . . 16
2.3 The distance between best matching sequences from two species is used to find initial seed orthologs and the distance between them $S$ is to expand the group with additional in-paralogs that are more similar to the main pair than this distance. . . . . . . . . . 19
2.4 Example showing domain architecture difference for the same species tree. After insertion, H1 gene has two domains leading to confusion, as H1 becomes orthologues to both M1, Z1 and M2, Z2 genes. . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . . 23
3.1 FunCoup pipeline showing: metric calculation for evidence type, model training with LLRs obtained from the gold-standard and random samples and network creation for each gold-standard. 30
1. Introduction

Recent advances in the genome sequencing technology contributed to accelerated growth in the number of sequenced genomes. Public online databases continue to grow with every additional genome sequenced. However, the function of many genes remains undefined. Many bioinformatics methods have been developed to find the function of such uncharacterized genes. Notably, gene orthologs play a vital role in this process.

Orthology refers to the pair of genes assembled from two different species that emerged in an evolutionary event called speciation. These genes tend to conserve their function more often than any other homologs. Orthologous relations can be used to shed light on a possible role of uncharacterized genes, given that there exists an orthologous gene with a known function in another, related species. Furthermore, ortholog tree by definition follow species tree and can therefore help to classify new species in phylogenetics. The growing number of uncharacterized genes and usefulness of orthologs justifies the importance of faster orthology prediction methods.

Most of the orthology prediction methods can be divided into two major categories: tree or graph-based. Tree-based methods focus on finding the evolutionary events in trees which define orthologous relationships. Graph-based methods use an estimated distance between sequences and aggregate most similar genes into so-called groups of orthologs. Many graph-based methods achieve comparable accuracy but at reduced computational complexity. Furthermore, graph-based methods do not require many conditions which need to be fulfilled when inferring an evolutionary tree.

Despite the advantages of graph-based methodology, many existing algorithms rely on all-vs-all sequence comparisons to estimate evolutionary distance between sequences. Given that the number of sequenced species is growing, the all-vs-all methods will meet a computational bottleneck and hence the need for faster methods.

The work described here aims to provide algorithms, as well as ready to use orthology assignments, that are accurate but also easy to compute using a hierarchical approach to reduce computational time. This work focuses on the importance of domain architecture in predicting orthologous relationships. By presenting an algorithm that finds orthology in the presence of different domain architectures. Results are contrasted to full sequence approach to get insight
into possible orthology relations otherwise missed in full sequence methods.

Finally, in the context of the functional coupling networks, an application of orthology is presented. Such systems strive to provide comprehensive information about interactions between genes based on different evidence types. Functional coupling networks are essential tools which, by integrating various evidence sources, help biologists to plan experiments and find potential novel interactions. Orthology assignments are used in two ways. First, to calculate phylogenetic profiles that serve as a separate evidence type. Second, to transfer experimental data, between species, where orthology mapping is possible. Together these works enable the functional identification of related genes or proteins.
2. Orthology prediction

2.1 Evolutionary relationships

Homology

The classical definition of homology refers to the shared physical trait that is conserved evolutionary and passed from a common ancestor. This meaning of homology in the phenotypic context can be accredited to Owen [1]. In the era of genomics this term has been re-appropriated to the narrower context of phylogenetics where homologs now differentiate themselves by sharing a common ancestor in their evolutionary history evidenced through sufficient sequence similarity.

Homoplasy

The classic converse of homology is called homoplasy, which initially represented similarity in phenotypic traits, arising from convergent evolution, but not present in common ancestry. Similarly, genomics-era homoplasy can refer to genes in which independent mutation events in separate lineages lead to similarity in the biological sequence [2].

Orthology and paralogy

Orthologs are genes that emerge from a speciation event in a last common ancestor. Therefore orthologous relationships in a gene tree will follow species tree. Opposite to this are paralogs, which are merely the result of gene duplication [3]. Subdivision among paralogs concerns the time series of duplication events. Pairs of genes which duplicate before speciation are referred as out-paralogs, while genes which duplicate after are called in-paralogs. Given the nature of the later, in-paralogs can also have orthologs within other species. Finding orthologs can therefore be simplified to looking for in-paralogs [4]. An example of orthologous relationships is depicted in Figure 2.1.

An example of orthologous relationships is depicted in Figure 2.1

13
2.2 Pairwise similarity

A prerequisite for reconstructing the evolutionary history of genes from sequence data can be generalized to the initial task of finding the distance between all sequences. This evolutionary distance can be approximated first by estimating how similar sequences are using various bioinformatic sequence alignment tools. Many bioinformatics methods rely on the information contained in the biological sequences. Therefore many comparison methods exist and can be used for finding orthologs. The most popular methods include global sequence alignment Needleman-Wunch [5], local Smith-Waterman [6], Basic Local Alignment Search Tool (BLAST) [7] and probabilistic HMMER [8].

2.2.1 Smith-Waterman

The Smith-Waterman is a dynamic programming algorithm which finds the most similar local region in two character sequences. It is commonly used to detect the best local alignment in protein or nucleic acid sequences.

This algorithm guarantees the best local alignment between two sequences given the scoring matrix for characters. The procedure starts by choosing the desired substitution matrix, which is used to assign a score between two characters in the alignment. The choice of substitution matrix is usually motivated by how distant homologs are to be found. Also, the gap penalty scheme is selected when opening and extending a gap. Given these parameters, two sequences are scored by filling up the scoring matrix. The last step involves tracing back the optimal route in the scoring matrix, from the highest scoring cell, while constructing the alignment, to determine the most similar local match.

Despite its optimal performance, the Smith-Waterman algorithm is computationally expensive and therefore many heuristics based alternatives have
been developed.

Substitution matrices

The BLock SUbstituion Matrix (BLOSUM) [9] or Point Acceptable Mutation (PAM) [10] are often used substitution matrices for protein sequences. BLOSUM matrices are constructed from closely related protein families, by aligning them and calculating frequency based scores which determine how likely it is for a given residue to mutate into another. Different versions of BLOSUM or PAM matrices are available, and typically suffixed with a number for the purpose of indicating on the sequence distance it is based on. High BLOSUM matrices are based on sequences with high similarity and should be used for aligning close homologs. Conversely, high PAM numbers correspond to the high rate of accepted mutations, and therefore such PAM matrices should be used for aligning more distant homologs.

2.2.2 Basic local alignment tool (BLAST)

Nowadays, one of the most popular sequence similarity search tools is BLAST. The BLAST algorithm directly approximates the local sequence alignment methods by applying heuristics, and is an order of magnitude faster than full local alignment tools while maintaining comparable sensitivity. Additionally, under the assumption that scores of local alignments follow Gumbel extreme value distribution, the statistical significance of aligned sequences is provided [7].

The algorithm starts by masking the low complexity regions, which contain repetitive elements. Pre-processed sequences are compared with a simple heuristic approach of finding the highest number of shared words of the fixed size (i.e k-length). The top scoring words in BLAST are then used to seed the full local alignment. Locally extended fragments are referred to as high scoring segments (HSP), and their statistical significance is assessed.

2.2.3 Hidden Markov Models

The next milestone in the development of sequence similarity search methods is achieved by the use of Hidden Markov Models (HMMs). Hidden Markov models can be used to model position specific frequencies of the characters in a biological sequence, as well as the position-specific insertions and deletions. Since HMMs are probabilistic models, statistical testing of potentially identified similarities is straightforward. A popular package - HMMER provides a tool - phmmer for searching a single protein sequence against the database.
of the sequences in a similar way to BLAST. Because there is no multiple-alignment involved and input is a single sequence, the tool builds a profile from a position-independent scoring system, a BLOSUM substitution matrix, and converts it to probabilities. Additionally, fixed gap opening and extension probabilities are added [11].

The profiles are linear sets of matches (M) which can emit a residue with a specific emission probability. Frequencies mainly determine the emission probabilities. Additionally, such profiles contain deletion states (D) which do not emit any residue and insertion states (I), which can emit residues. Each state can be repeated multiple times, transition probabilities model transitions between states, which correspond to position (column) in the aligned sequences, shown in Figure 2.2.

The objective is to compare HMM-profile to a sequence or a database of sequences. Sequences that score better compared to null model are considered to be homologs to the query profile. The query profile can be constructed either from multiple sequence alignment or single sequence with the help of substitution matrices and some assumptions about gap opening and extension.

2.2.4 Other search tools

**MMseqs2** (Many-against-Many searching) is a tool developed to search and cluster large number of proteins [12]. It is designed to run in a distributed computing environment and is 10000x faster than BLAST maintaining similar sensitivity.

**DIAMOND** It is a sensitive and fast sequence alignment tool, similar to BLAST, but with improved indexing. It has been shown to be 20000 times faster than BLAST and maintains similar sensitivity [13].

2.3 Multiple sequence alignment

The phylogenetic methods for reconstructing trees often starts with building a multiple sequence alignment (MSA). Multiple sequence alignment is a way
to optimally arrange the residues of biological sequences according to their similarity, so similar residues are placed one under another with gaps between non-matching regions. Such alignment allows for identification of the conserved areas that might have an essential function for the whole protein.

Methods to construct multiple alignments can be roughly divided into three categories: dynamic programming, progressive and iterative. Dynamic programming is a method that extends pairwise alignment to multiple sequences by defining scoring functions and optimizing alignment by building multidimensional scoring matrix for all the sequences simultaneously. However, this approach is computationally very demanding and requires a lot of computing memory. The progressive alignment is the most commonly used method, repeatedly performing pairwise alignment. It starts with the most similar pair of sequences and progresses to the most distantly related. Examples of progressive alignment methods include software such as Kalix [14], Clustal [15], T-Coffee [16] and MAFFT [17]. Progressive alignments are sensitive to the choice of starting the pair of sequences, and this is mitigated by iterative alignment methods, which repeatedly re-aligns the first pair of sequences while adding new sequences to the alignment for example Muscle [18].

Given the multiple alignments, distance matrix can be computed between every pair of aligned sequences. Numerous methods are available to estimate that distance with the aim to provide more realistic measures of true evolutionary relationship by applying a correction to observed substitution rates. One example of such correction methods is scoredist [19].

2.4 Orthology prediction

Both orthologs and paralogs are pair-wise relations defined in the context of evolutionary events. The better the inference of evolutionary history between sequences, the better the orthology predictions. Full evolutionary events identification gives an advantage to the methods which focus on analyzing evolutionary trees. However, reconstructing gene phylogenies with modern probabilistic models is computationally demanding, and it is not feasible for extensive datasets. Building trees might also depend on the correct multiple alignments. An alternative graph-based methods exist, relying on similarity searches estimated between pairs of sequences as a proxy for evolutionary distance.

2.4.1 Tree based methods

The first step in tree-based ortholog identification usually is a reconstruction of phylogenetic tree. The most straightforward methods use a distance matrix, obtained from multiple sequence alignment assumed to be correct. Given
the distance matrix, a hierarchical clustering method such as unweighted pair group method with arithmetic mean (UPGMA) [20] or neighbor-joining (NJ) [21] can be used. UPGMA assumes that all lineages evolve at the same rate, while NJ does not but instead require its trees be rooted, a universal common ancestor. Both methods lack consideration for any evolutionary model but are relatively fast and accurate given enough data. After rooting the tree, the gene and species tree reconciliation can be performed to identify the evolutionary events and infer potential orthologous relations [22].

Another approach for estimating phylogenetic trees is maximum parsimony. Maximum parsimony produces a topological tree by minimizing the number of substitutions that are needed to explain the differences in positions of the alignment. Such methods, similarly to the distance-based ones, do not assume any explicit model of the evolution. Under maximum parsimony criterion, the best tree is the shortest tree, which does not guarantee the correct tree. Additionally, finding the most parsimonious trees for a large number of nodes becomes computationally infeasible and requires heuristics.

Probabilistic methods are the most rigorous and provide the possibility of using explicit evolutionary models as well as assigning probabilities to both tree branches and picking the most likely trees in the case of the maximum likelihood (ML) [23] approach, or finding a distribution of possible trees via Bayesian inference [24]. Probabilistic methods are the most computationally demanding and require careful evolutionary model selection [25].

Although tree-based methods seem ideal for identifying orthologs, their sophistication does not scale well to more massive datasets. Furthermore, some organisms are likely to deviate from a tree like the model of evolution (i.e. horizontal gene transfer), which can negate meaningful orthology inference via conventional tree-based methods [26].

2.4.2 Graph based methods

Graph-based methods usually require finding all-vs-all similarities between sequences and species. Differences in methods often lay in the type of similarity search algorithm and the way orthologs pairs are identified or clustered.

Sequence similarities serve as proxy for the evolutionary distance between species. Graph-based methods build groups of orthologs from similarities and might differ by the clustering approach. The simplest way is the best reciprocal hit (RBH) [27]. Given two proteins $i$ and $j$ and their corresponding genomes $i \in I$ and $j \in J$ respectively, the RBH is defined as the best match of $i$ in genome $J$ and at the same time $j$ being the best match in genome $I$. Although this method is very fast, it has its caveats, one being that it is not very accurate and another that it only considers best matches, ignoring other in-paralogs.
A more elaborate method which considers other in-paralogs is InParanoid [4]. InParanoid also applies a low complexity filter and uses an overlap threshold to avoid spurious matches due to repeats and low complexity regions. The underlying assumption of InParanoid is that two sequences from the same species are in-paralogs if they are more similar than the closest match from any other species (Figure 2.3). At the time of writing InParanoid 8 [28] provides the orthology predictions for 273 species.

![Figure 2.3: The distance between best matching sequences from two species is used to find initial seed orthologs and the distance between them S is to expand the group with additional in-paralogs that are more similar to the main pair than this distance.](image)

Many variations of the graph based approaches have been developed. The Orthologous MAtrix (OMA) uses full Smith-Waterman sequence alignment for all-versus-all comparisons. The algorithm applies a greedy algorithm to find the maximum weight edge cliques for identification of the groups of orthologous sequences [29].

An alternative approach relies on the Markov Cluster algorithm [30] and is used by the OrthoMCL [31] algorithm. As with the aforementioned methods, here the relationships are inferred from all-vs-all similarities. Authors report similar results to InParanoid with an advantage of clustering more than two species simultaneously. At the time of writing, no comprehensive benchmark for more than 7 species was performed.

### 2.4.3 Hierarchical methods

The strength of the graph based approaches is speed, although the computational time required to perform all-vs-all searches grows quadratic with the number of species. This can be circumvented by restricting the search space only to the species along the fully bifurcated guide-tree. An example of such method is Hieranoid [32; 33]. Hieranoid heavily depends on the correctly bifurcated species tree.

The algorithm starts with a conventional similarity search only now this is between the species at the leaves of the tree. It proceeds to the parent nodes,
representing them with orthologous groups, which are collapsed either into consensus sequences or profiles. Parent nodes containing representative sequences or profiles are used further up the branches to perform the similarity searches. This strategy reduces the time required for all-vs-all comparisons, although a limitation is that not all comparisons can be parallelized in such a way.

2.5 Other orthology methods

Some of the methods predict orthologous genes by applying transitivity property, which implies that if A-B and B-C are pairs of orthologs, then A and C are also orthologs [34]. Such approach dramatically reduces the need for performing all-vs-all comparisons, as some relations can be inferred indirectly.

In addition to the pure sequence similarity-based methods, approaches using synteny can be found. In those ones, the conservation of gene order is used to improve the orthology assignments [35]. In previous study, a local synteny between two genes is defined as the maximum number of homologous matches between six genes in the genomic neighborhood. Homologous neighboring genes are found by BLAST, thus accounting for the genomic conservation.

An accessible and extended approach that combines sequence similarity with the genomic location (synteny) is known as MOSAR [36]. MOSAR calculates all-vs-all sequence similarities and assigns orthologs based on rearrangement/duplication (RD) distance, finding the minimum common partition followed by the maximum graph decomposition and detecting inparalogs.

2.6 Quest for Orthologs

Quest for Orthologs is a joint community effort that tries to improve orthology predictions by collaboration. Examples of the resources are standard reference proteomes, standard file formats and a list of databases [37]. An up to date list of orthology databases is maintained at https://questfororthologs.org/orthology_databases.

Reference proteomes

The first set of reference proteomes consisted of 66 representative species. Sequences are manually curated; usually, the most extended transcript is taken from both UniProt [38], Ensembl and Ensembl Genomes [39]. The latest reference proteomes contain 78 species. In addition to protein sequences, resources include the mapping between gene identifiers and protein identifiers.
Standard formats

Many orthology prediction methods output their predictions in various formats, which makes for difficult comparison. Furthermore, those methods often rely on various versions of FASTA files to produce their results. The QfO consortium established standard formats for input - SeqXML and output - OrthoXML data [40] to assure more consistency between collaborating groups.

Standard benchmarking

Further collaborations within QfO defined a standard orthology benchmarking service [41]. This benchmark performs a set of phylogenetic and functional tests [42]. Orthology is defined by speciation-event; therefore it is expected that a gene tree would follow a species tree. Based on this premise, gene trees are reconstructed when two submitted benchmark pairs are present in a known species tree. Inferred trees are compared to the known species tree and the Robinson-Foulds (RF) distance is computed as a proxy for false positives, then the average value for multiple sampled trees is reported by the benchmark.

2.7 Protein domains

The fragments of the protein sequence or their structural counterparts which show evolutionary conservation can be referred as protein domains. Protein domains are motifs which can evolve independently of the remaining part of the whole sequence. Different definitions of protein domain exist and it also can be viewed as a conserved three-dimensional structure, such as those archived in SCOP database [43], but usually it is refers to the sequence conservation based as domains collected in Pfam database [44] in the form of profiles.

Pfam domains database is particularly useful for domain based orthology prediction methods, which are based on information from protein sequences. Pfam domains are built from curated seed alignments, which are represented by the Hidden Markov Model (HMM) profiles, summarizing the variation within the domain families. This profile is useful for finding distant domain homologs, as well as for identifying domains in the new protein sequences. The use of profiles also allows domain based orthology methods to descend into a domain realm by predicting domain coordinates in sequences provided by the end-user and analyzing them separately.
2.7.1 Domain architecture

Although the majority of proteins for the full range of species collected in Pfam are the single domain proteins [45], a protein with multiple domains appear as more frequent in eukaryotes than in prokaryotes [46]. The particular order and the content of numerous domains in proteins are often referred as a domain architecture. Many of the proteins gain a new function due to evolutionary events leading to the domain fusion, fission and rearrangements in multi-domain architectures [47]. Because different domain architectures can determine the protein function, tools that predict protein function from the information of domain composition has been developed [48; 49].

2.7.2 Domain similarity

Quite a few measures that can be used to compare multiple domain architectures can be found in the literature. The first one defines the domain distance (DD) as the number of nonmatching domains, counted from a query against shorter or equal length domain architectures [50]. Another study explored the applicability of the domain similarity measures to compare domain content, such as Jaccard index and Goodman-Kruskal $\gamma$ index [51], which were benchmarked against Cluster of Orthologous Groups (COG) database [52]. Further studies summarize and compare different domain similarity measures and their suitability for the homology identification [53]. Based on previous studies, network topology based approaches exploring Neighborhood Correlation (NC) have been developed for finding multi-domain homologs that may have different domain architectures [54]. Later studies use a combination of NC and synteny to improve the partitioning of homologous genes [55] and at the same time increasing computational complexity of the algorithm.

Methods as mentioned earlier either focus on the estimation of a distance-similarity between sequences or assume that proteins with the most similar domain architecture are good candidates for homologs. However, homology is not the equivalent to the orthology. Orthology distinguishes itself from homology by not only relating genes which follow speciation events but also emphasizes stronger functional conservation between genes related through orthologs relationship [56]. Distinguishment of orthologs from homologs genes has been performed either from evolutionary trees or by the faster method from graphs by different clustering techniques [4; 29; 31].

22
2.8 Domain orthology

Many of the currently available orthology prediction methods assume that a full protein sequence is the most basic evolutionary unit. The long-standing history of protein classification by conserved subunits in Pfam database shows that sequences can be dissected into smaller parts called protein domains [44]. Evolutionary trees obtained by the distance methods from full sequences and contrasted to the trees from protein domains alone may indicate different evolutionary histories [50]. Even close concerning sequence similarity homologs - orthologs have been shown to sometimes have different domain compositions [57]. Changes in the domain architecture of proteins may confuse full-sequence based orthology prediction methods. A hypothetical scenario may involve the domain insertion, depicted in Figure 2.4.

![Figure 2.4: Example showing domain architecture difference for the same species tree. After insertion, H1 gene has two domains leading to confusion, as H1 becomes orthologous to both M1, Z1 and M2, Z2 genes.](image)

This highlights the importance of the development of algorithms capable of distinguishing orthologs that are supported by all domains or a fraction of domains. Local alignment methods generally focus on a single similarity region or the most similar single domain, hence miss the neighborhood of different domain architectures. On the other hand, methods relying on the construction of the multiple alignments to build trees can have difficulty in aligning such sequences with a very different domain content, leading to an unreliable estimate of evolutionary distances and incorrect trees.

Different tools exploit the concept of the domain architecture to improve orthology predictions. A computationally efficient approach known as DODO focuses on performing reverse PSI-BLAST against Pfam profiles and clusters domains based on the simple reciprocal best hit (RBH) method [58]. The advantage of this method is its speed compared to conventional InParanoid. Additionally, the approach for orthology prediction relies on multiple sequence alignment of the best matching single domains and tree-building [59], but it ignores the remaining domain context and no automated pipeline was published, making it difficult to apply for large-scale analysis. Finally, software
porthoDom, uses domain similarity metrics to cluster input sequences to subgroups. Orthology detection is performed within each sub-group, greatly reducing computational time as compared to all-vs-all approaches [60].

2.9 Methods and databases

A selection of the orthology databases with their short description explaining notable differences is presented in this theses. For more complete and up to date list one should refer to the Quest for Orthologs website.

COG

Clusters of orthologous group (COG) is an online resource containing orthology predictions obtained by all-vs-all blast search and clustering based on best hit (BeT) in each of the other genomes. Currently it is limited to 66 organisms [61].

InParanoid

InParanoid is an online database aggregating pair-wise predictions for 273 organisms including 246 eukaryotes, 20 bacteria and 7 archaea. Predictions are performed with InParanoid 4.1 algorithm and input sequences are taken from both Ensembl and Uniprot [28].

OMA

The Orthologous MAtrix is a database of orthologs for a collection of proprietary and complete public genomes. OMA delivers pair-wise orthology assignments from all-vs-all Smith-Waterman alignments, collapsed and represented by the hierarchical orthologous groups [62].

OrthoMCL-DB

OrthoMCL-DB is an online database consisting of predictions for 55 species, including 16 bacterial and 4 archael genomes. OrthoMCL-DB relies on all-vs-all comparisons with BLAST and Markov clustering (MCL) algorithm to derive the orthologous groups from multiple species [63].

PhylomeDB

PhylomeDB is a database containing maximum-likelihood trees and multiple-alignments as well as phylogeny based predictions across 1059 species [64].
EggNOG

EggNOG is a collection of orthologous groups (OGs) combined with functional annotations from gene ontology terms, pathways and protein domains. The database covers orthology assignments for 2031 organisms. HMMER finds similarities from all-vs-all comparisons and Hidden Markov models (HMM) profiles are available for each orthologous group [65].

EnsemblCompara

Another orthology resource is EnsemblCompara, which contains gene trees obtained from the phylogenetic approach for orthology. An automated pipeline runs BLAST and RBH clustering, multiple alignment and tree generation while being capable of handling large gene families [66].

HieranoiDB

HieranoiDB is a database that contains predictions of the hierarchical groups of orthologs for 66 species. Instead of all-vs-all similarity searches, Hieranoid uses a guide tree to reduce the computational complexity by the indirect comparisons. The hierarchical approach produces trees which can be viewed in an interactive tool [67].

TreeFam

The TreeFam is a database of phylogenetic trees for gene families in animals [68]. TreeFam is based on the multiple sequence seed alignments similarly to Pfam [44], but instead of classifying domains it uses full-length gene sequences. The gene family trees are built using the constrained version of neighbor-joining algorithm [21].
3. Association networks

To understand the function of one protein, it is necessary to understand its role in the whole system and therefore the function of proteins associated with it. A network graph can conveniently represent such a relation. Association in this context means either physical interaction, belonging to the same protein complex, participation in the same signaling or metabolic pathway or same share the same subcellular localization. Shreds of evidence of such associations can be collected from a different types of experimental data and sources. Integrating that knowledge under a framework of functional association networks can help identifying functional modules essential in performing a specific biological functions.

3.1 Networks definition

Nodes connected by edges appear in the association networks. Commonly nodes represent constituents of the observed system and edges represent relations between them. Association networks can be directed determining the direction of the influence or undirected in the case when the causality is not known or omitted for generality.

Analysing association networks focuses on studying collective behavior and interplay between nodes. A graphical representation of a network gives an opportunity to see the system from more global perspective. The existence of modules or hubs becomes more apparent, potentially revealing distinct patterns. Such comprehensive approach distinguishes network-based approaches from studying properties of isolated nodes.

Another feature of the association networks is the strength of the links which represent the propensity of the connection to relate two nodes or the degree of belief that such interaction can occur given the collected evidence.

3.2 Network methods

Different methods for learning association networks from the multiple data types have been developed. Initially, the problem of inferring such networks was attempted by an approach in which links are categorized into high and
low confidence by an agreement in the underlying data [69]. Alternatively, individual networks have been inferred independently for each evidence type and combined into consensus [70]. The more advanced approaches involve the application of Support Vector Machines (SVMs) [71], which turned out to be computationally too complex for bigger problems. Another approach used Gaussian field propagation algorithm for binary classification, predicting the presence or lack of association [72]. Additionally, the evidence has been weighed by the number of contributing samples using ridge regression [73].

The common denominator of all the above methods is a binary classification task, in which presence or lack of the link between two nodes is to be predicted. Solving these problems involves the integration of heterogeneous data sets, with limited overlap, missing values and varying signal-to-noise ratios. A well-suited prediction method that works with such data calls for a Bayesian approach [74], typically the Naïve Bayes classifier [75].

3.2.1 Bayes’ theorem

The Naïve Bayes classifier has its root in Bayes’ theorem. The concept was introduced by Thomas Bayes, who viewed the probability as the degree of belief and expressed it in equation 3.1.

$$P(\theta|x) = P(\theta) \frac{P(x|\theta)}{P(x)} \quad (3.1)$$

Bayes’ theorem can be used to calculate the posterior, probability of a link \( \theta \) given the evidence expressed as \( P(\theta|x) \) from the initial probability of observing a link \( P(\theta) \) known as prior and the ratio between likelihood \( P(x|\theta) \) probability of observing evidence if the link exists and the probability of evidence \( P(x) \) known as marginal probability.

3.3 Classification

The presence or lack of the link can be considered as two mutually exclusive hypotheses \( H_0 \) or \( H_1 \) and some evidence data \( x \). Expressing two hypotheses using Bayes’ theorem leads to

$$\frac{P(H_0|x)}{P(H_1|x)} = \frac{P(x|H_0) P(H_0)}{P(x|H_1) P(H_1)} \quad (3.2)$$

That expression is sometimes called the Bayes factor, and it is a measure of how the evidence favors either of two hypotheses proportional to the degree that either of hypotheses predicts the observed data better than the other.
In order to apply the Bayes factor to multiple evidences, under the assumption of independence, it can be expressed as a product of ratios

\[
P(H_0|x) / P(H_1|x) = \prod_i P(x_i|H_0) / P(x_i|H_1) \prod_i P(H_0) / P(H_1)
\]  

(3.3)

where \(x_i\) corresponds to different independent evidence types.

Because of the limited machine precision, multiplication can be replaced by summation of natural logarithms.

\[
\ln P(H_0|x) / P(H_1|x) = \sum_i \ln P(x_i|H_0) / P(x_i|H_1) \ln P(H_0) / P(H_1)
\]  

(3.4)

The likelihood of association \(P(x_i|H_0)\) can be calculated by the occurrence of links in the training set (gold standard), while the lack of association \(P(x_i|H_1)\) can be calculated from the randomized pairs. The right-hand side of the equation 3.4 contains expression for prior which is constant and can be left out.

\[
LLR_i = \sum_i \ln P(x_i|H_0) / P(x_i|H_1)
\]  

(3.5)

For each piece of the evidence data, specific metrics can be calculated, for instance Pearson’s correlation coefficient between genes for mRNA expression data. For some genes present in the gold-standard (positive samples) and present in the random pairs (negative samples) loglikelihood ratios (LLRs) can be computed (see Figure 3.1).

Metrics computed for each evidence type take usually continues values, but as they are not easy to work with, since it would require difficult estimation of joint probability densities. Instead, scores are binned across data specific metric range, where bin sizes are estimated with adaptive method [76] calibrating model. Then LLRs are computed within each bin. For the remaining genes that are not presented in the gold-standard, depending on to which bin pair of genes falls into, calibrated LLRs are assigned.

### 3.4 Gold-standards

To calibrate the Bayesian classifier, the gold standard with known positivize and negative interactions is required. Part of that gold-standard must overlap with genes present in the evidence data. For instance in FunCoup [77], the gold-standard is divided into five categories: protein-protein interactions (PPI), same complex co-membership, same metabolic or signaling pathway and shared operon. Negative samples are obtained by reshuffling gene pairs.
Figure 3.1: FunCoup pipeline showing: metric calculation for evidence type, model training with LLRs obtained from the gold-standard and random samples and network creation for each gold-standard.

to get a random coupling which should not interact. For each gold-standard, a separate network is inferred and later each link is summarized by taking the highest scoring link, but downloadable flat files contain information on the link strength for each gold-standard.

3.5 Evidences

Phylogenetic profiles

One of the first methods for predicting associations using phylogenetic profiles [78] is based on the assumption that two related genes more often co-occur or are absent from set of species than unrelated genes. Therefore, classically, the phylogenetic profiles are constructed for each protein, one entry for each genome, encoding absence or presence of a homolog. Correlated profiles are used as the indicators of the possible functional link. This approach, however, does not take into account evolutionary events. More advanced strategies consider correlated loss and gain of genes obtained from maximum likelihood (ML) estimated trees [79]. However, ML trees method is computationally expensive and hard to apply for large-scale problems. An alternative heuristic method for finding similarities and estimation of phylogenetic profiles was introduced in the functional coupling network FunCoup 3.0 [80]. Given the
pre-computer neighbour-joining (NJ) tree, it is rooted in the species for which link between two genes is predicted. The profile for two genes is populated with a positive score if two species contain a pair of genes up to the common ancestor. The negative evidence score is calculated for species which do not share a pair of genes. Scores are normalized and their sum is bound between 0 and 1. Those scores are binne and log-likelihood ratios are computed from the ratio between positive and negative evidence samples in each bin.

mRNA co-expression

The mRNA co-expression (MEX) consist of distinct datasets mainly downloaded from Gene Expression Omnibus (GEO) [81]. A correlation coefficient is calculated across experimental conditions and genes with similar co-expression are considered to be related.

Protein interaction

Physical protein interaction (PPI) is obtained from iRefIndex [82]. Interactions are weighted according to how often they appear to be confirmed in iRefIndex and based on the type of experiment.

Protein co-expression

Levels of transcripts do not always correlate with protein levels. To complement mRNA co-expression data, ranked expression of protein abundance in tissues is taken from Human Protein Atlas (HPA) [83].

Shared transcription factors

Genes are often regulated by transcription factors (TFs). Commonly genes regulated by the same set of TFs happen to perform a similar function. Similarities in shared transcription factors measured by Jaccard index are indicators of potential functional couplings.

Shared miRNA

Similarly, to shared transcription factors, miRNAs are another level of regulation that often applies in case of co-regulated genes. Similarities in shared miRNAs are therefore good indicators of functional coupling.
Domain interactions

For a pair of potentially interacting domains, a confidence score is computed by summing individual confidence scores for each domain from predicted domain interactions in UniDomInt [84].

Quantitative mass spectrometry

The absolute abundances of proteins are downloaded from PaxDB database for multiple species [85]. Protein abundances are sorted into highest and lowest categories, and Jaccard index is used to measure similarity in abundance across different conditions.

Sub-cellular co-localization

Sub-cellular localizations are extracted from gene ontology (GO) terms [86]. Similar and dissimilar co-localization are indicators of the presence or lack of functional coupling. Terms in gene ontology are organized in the form of the tree from leaves that are more specific towards more general terms at the root. Less general terms are down-weighted in this metric.

Genetic interaction profile

Genetic interactions obtained under different experimental conditions (phenotype readout) can be used to calculate a similarity with simple Pearson’s correlation. Two genes that interact with the same set of genes are more likely to be functionally related [87].

3.6 Orthology transfer

Orthologous genes have been shown to be functionally more conserved [56] than any other paralogues. Therefore, the evidence of the functional association in one species can be used to indicate the same interaction in another species [88]. A different approach is to transfer the data for orthologous genes to the new species [89]. This approach requires a database with ortholog pairs, and the amount of transferable data heavily depends on the mapping between external orthology database and the internal identifiers.

3.7 Functional coupling networks

Starting with FunCoup, the selection of other similar network resources is given in this section. Short description is provided which highlights main fea-
tures of databases along with relevant citations.

FunCoup

FunCoup is a functional coupling network integrating multiple evidence types into a single network by Naïve Bayes approach. Orthology information is used to both construct phylogenetic profiles for predicting couplings between links and to transfer data for the orthologous gene pairs between species. The latest version 4.0 [77] includes 17 species, and its improvements are the subject of the last paper in this thesis.

GeneMania

GeneMania uses a real-time, multiple association network integration algorithms to predict the functional links between genes [90]. First, the functional association network is constructed from the data sources with weights for each link. Weights correspond to the usefulness of evidences to predict functional link and they are obtained using a linear regression-based algorithm that calculates composite association network from multiple data sources. Secondly, an adopted variant of Gaussian field propagation algorithm is performed to assign a score to each node in the network. This score defines the strength of the association which node has to the list of functions it defines.

STRING

The Search Tool for the Retrieval of Interacting Genes-Proteins (STRING) database aggregates the data on associations from various sources, like predictions based on conserved gene neighborhood, gene fusion events, phylogenetic gene co-occurrence across multiple genomes, mRNA co-expression data or co-appearance of gene names in literature abstracts. STRING allows the user to select either of two modes for transferring associations between organisms, orthology based or simpler sequence homology-based. Each evidence is benchmarked separately against KEGG and scores are combined in naive Bayes way [91].

GIANT

The Genome-Scale Integrated Analysis of Networks in Tissues (GIANT) provides another Naïve Bayes integration framework used to predict 144 human tissues specific networks. Associations are benchmarked against the tissue-specific gold-standards [92]. The web-based front-end offers a tool for comparative analysis of tissue-specific re-wiring of genes.
IMP

The Integrative multi-species prediction (IMP) database integrates functional associations from high-throughput data [93; 94] for 7 organisms using regularized for redundancy between datasets, Naïve Bayes algorithm. The evidence types include protein-protein interactions, co-regulation by transcription factors (TFs), mRNA co-expression and presence of similar Pfam domains. Information on associations is supplemented from other species by orthology transfer using the TreeFam orthology database. Additionally, IMP uses the predicted functional networks as an input to SVM to classify genes in the network to specific biological processes from GO terms or disease genes. Diseases genes predictions for human are obtained from Online Mendeline Inheritance in Man (OMIM) database [95].
4. Present investigations

4.1 Paper 1

Schreiber et. al introduced the first version of Hieranoid [32] for efficiently finding orthologs at the reduced computational time. However, the previous implementation suffers from low coverages compared to the existing methods, some variability and could not run in parallel computation mode. Most of the graph based orthology prediction methods require all-vs-all comparisons and their computational time grows quadratically with the number of species. The idea behind Hieranoid is to utilize the known fully bifurcated species tree as a guide tree to perform a full comparison only at the leaf nodes and to build consensus sequences that represent pseudo-species at intermediate nodes while the algorithm proceeds from leaves to the root. Hieranoid was further developed by improving its reproducibility, extending it to allow distributed computing on compute cluster. Additionally, the accuracy and coverage were improved by changing the way how multiple alignments and consensus sequences are constructed. The final results were validated against the standard benchmark from Quests for Orthologs (QfO). Predicted orthologs were compared Hieranoid 2 to InParanoid per species pair for the same set of species. We found that the most prominent differences could be attributed to large protein families in which few different genes lead to substantial differences in the number of orthologs.

Future prospects

In future Hieranoid, could be improved by including new faster tools similarity search tools, such as MMSeq2 or Diamond [12; 13]. Currently, Hieranoid 2 produces results in the form of OGTrees, which is an adapted Newick format for storing phylogenetic trees. To get orthologous pairs, one needs to parse all the trees and supply additional information about species. New Hieranoid, could be redesigned to produce orthoXML files automatically, which are easier to parse and contain all necessary information. The new version of Hieranoid should also be recomputed with newly released reference proteomes and uploaded to the community benchmark.
4.2 Paper 2

In the second paper, the new web interface is implemented that allows the user to search through the results of Hieranoid 2, pre-computed for 66 species from reference proteomes provided by Quest for Orthologs (QfO) consortium. The web interface simplifies the use of Hieranoid 2 predictions by eliminating the necessity for setting up the pipeline or running any computations. Additionally, features allow searching the database by protein description or gene symbols and finding relevant trees. Large trees can be manipulated, they are searchable and can be exported in standard formats, like the community standard OrthoXML. For the sequences which are not covered in 66 reference proteomes and since reference proteomes contain only the most extended transcript BLAST feature has been implemented, which allows to search for the most similar sequence and its orthologs.

Future prospects

In the current release, HieranoiDB contains only 66 model organisms. Orthology predictions were obtained already for 273 species, which is the same set of species as in InParanoid 8 [28]. The future development would involve extending the database to more species. The other potential improvements could include presenting domain architecture for HieranoiDB ortholog groups and multiple intermediate alignments obtained from Hieranoid 2 algorithm.

4.3 Paper 3

The third paper explores the possibility of finding the domain based orthologs with the Domainoid algorithm. The motivation for this approach is rooted in the need to extend the full length orthology predictions by InParanoid algorithm to include also the sequences which could undergo domain rearrangements. First, domains are extracted from the full sequences and the conventional InParanoid run is performed on the domain level. Resulting clusters of domains are used to calculate the ratio between co-clustering domain and all domains that pair of proteins are detected with. Next, the threshold is applied to extract pairs with sufficient number of matching domains. Pairs that meet threshold are compared to results obtained from full length InParanoid run. The investigation focuses on the new pairs that are uniquely found by domain based approach. The domain-level approach inevitably leads to failure to detect some orthologs, usually due to insufficient similarity in short domains, but these are readily recovered from InParanoid.
Future prospects

Currently, Domainoid aims to extend the InParanoid algorithm to orthologs that have different domain architectures. It can be used as a protein-level orthology tool by applying a threshold on the fraction of domains that are orthologous between two proteins. In the current version, long unannotated regions are considered domains as they could potentially be unidentified. These regions often tend not to have orthologs thus lowering the fraction orthologous domains and lowering the threshold required to obtain protein-level orthology. By identifying a large number of domain-based orthologs missed by conventional methods, exploratory analysis can be carried out on Domainoid results. An interesting question is how often a protein possesses domains that are orthologous on the domain level to domains in different proteins. Moreover, such phenomena and their frequency can be compared across different taxonomic levels. Domain-based orthologs can also be categorized into different types of domain rearrangements, such as domain gain, domain loss, domain order change and it can be estimated how frequent such events are. For the current version of Domainoid, the default InParanoid parameters are used. These parameters are optimized specifically for full-sequence based orthology predictions, hence the influence of adjusting them for a domain based approach should be further explored.

4.4 Paper 4

The last paper focuses on finding the functional couplings between genes, by integrating different evidence types for which orthology and phylogenetic profiles play a crucial role. Orthology predictions are obtained for the established database InParanoid 8. Orthology is used both to transfer data between species and to compute phylogenetic profiles which are used as evidence type itself. Not all evidence types are transferred between the species. For instance, subcellular localization or phylogenetic profiles are excluded from the transfer. In the new release of FunCoup 4.0 [77] a new evidence type Quantitative mass spectrometry (QMS) is included by adding data sets obtained from PaxDB database [96]. The website interface is refreshed with a new JavaScript-based interactive network viewer which replaced previous jSquid applet [97]. Previous experimental data were updated, as well as new data sets were added. Finally, FunCoup is extended by six new model organisms. Previous versions focused only on eukaryotic organisms, while with FunCoup 4.0 two new prokaryotic organisms are included to open the door for finding functional couplings in bacteria.
Future prospects

Currently, mRNA expression evidence type obtained from Gene Expression Omnibus (GEO) database is the largest source of experimental data in FunCoup. To compute the score for the metric used to assess if two genes are co-expressed, Pearson product-moment correlation coefficient can be used. Many of correlations in this simple approach might be spurious or indirect, being correlated with another variable. One way to improve the quality of predictions for directly related targets would remove spurious associations by calculating first the order partial correlation with recursive formula

\[
\rho_{x|z,y} = \frac{\rho_{x,y} - \rho_{x,z}\rho_{z,y}}{\sqrt{1 - \rho^2_{x,z}}\sqrt{1 - \rho^2_{z,y}}}
\]

(4.1)

where \( x, y \) is the pair for which first-order partial correlation is computed and \( x, z \) and \( y, z \) are the pairs identifying \( z \) variable with the highest correlation to both \( x \) and \( y \).

Computing first-order partial correlations would still be computationally feasible for high-throughput data integration such as one used in FunCoup, but the effect of some of the indirect associations would be reduced. This could also reduce the overall amount of links from mRNA co-expression (MEX), but the availability of new mRNA expression data is not a limiting factor, in fact, it was necessary to avoid introducing too much of it, not to make network too biased towards that particular evidence type. The smaller number of the direct links could be compensated by adding more MEX data.

Further iterations of FunCoup could also include disease-related genes in the network-viewer. Furthermore, currently, the final Bayesian score for each link is the highest one picked from all gold-standard networks, whereas for future FunCoup versions, an alternative approach could be developed, where all gold-standards are integrated, contributing to final network topology.

Finally, current training pipeline and web-application is based on core technology, which is mostly outdated. Improvements could be made in the network-training pipeline, leveraging it in the form of scripts using a more performant programming language (i.e., Scala, Julia) that is easier to parallelize and supports distributed computing. Besides new JavaScript network viewer, a large part of FunCoup website still relies on the older technological solutions, which could be replaced for instance by Python Django framework, to allow similar functionality at reduced costs. A new command-line based pipeline could produce XML or JSON files with networks, while web-application could accept this new standard format.
Sammanfattning


Likheter mellan två sekvenser kan berätta om deras evolutionära historia, som i sin tur kan avslöja deras funktion. De evolutionärt närbesläktade protei- ner, som har snarlika aminosyrasekvenser har oftast liknande funktion. Detta är viktigt, då det tillåter upptäckten av funktioner hos okända proteiner genom att studera funktionen hos liknande proteiner hos andra arter.

Tack vare de snabba tekniska framstegen har det blivit enklare att få fram nya sekvenser för nya organismer. Dock återstår fortfarande utmaningen att bestämma proteinernas funktion antingen genom experiment eller genom användandet av bioinformatiska verktyg.

Ortologer är genpar som uppstår då en art utvecklas till två. Ortologer är kända för sin förmåga att behålla samma funktion trots att generna i genparet hamnar i nya arter. Detta faktum används ofta i bioinformatik för att förutsäga funktionen av ännu icke karakteriserade gener utifrån gener med känd funktion i en annan art.


Sådana komplexa bioinformatiska verktyg kan för en stor analys fortfar- rande kräva åtkomst till ett kraftfullt datorkluster, därför utvecklades en webbaserad databas i det andra projektet. Med det verktyget kan andra forskare leta efter redan identifierade ortologer direkt via webbläsaren. Dessutom för-
bätttras resultaten genom att visa proteiner direkt i ett evolutionärt träd med proteinbeskrivningar och arttillhörigheter. Grafisk representation i form av träd tillåter enklare tolkning av resultaten.


FunCoup är ett exempel på ett proteinnätverk, byggt från olika experimentella datakällor, där länkar representerar hur säker man är på att en funktionell koppling mellan två proteiner existerar. FunCoup fokuserar på att integrera rent experimentella data och förutsäga nya funktionella kopplingar.
Acknowledgements

I would like to thank my supervisor Erik Sonnhammer, for giving me the opportunity to contribute to the field of orthology throughout my PhD studies and also for being available and helpful as well as for his positive attitude. I would further like to thank him for introducing me to Quest for Orthologs (QfO) community and amazing group retreats. I would like to thank my co-supervisor Andrej Alekseenko, who was always available and ready to discuss scientific progress. Additionally, I would like to thank Gunnar von Heijne for our talks during the checkpoints and the enjoyable qualifying exam.

I thank staff at DBB for their administrative help throughout my whole PhD studies. I would also like to thank the former dean Stefan Nordlund and the current dean Pia Ådelroth for being helpful and support with the courses. Finally, I would like to thank my collaborator Sophia K. Forslund for suggestions and inspirational enthusiasm about domain aware orthology prediction.

I would also like to acknowledge my former colleagues at SciLifeLab, Christoph Ogris for inspiring discussions while we were sharing an office, Andreas Tjärnberg for hackathon and generally talking nerdy trends and introducing me to Emacs, Torbjörn Nordling for helping me out with my initial accommodation, Matthew Studham for his relaxed attitude, Gabriel Östlund for help with InParanoid and Fabian Schreiber for his help with grasping the Hieranoid code base.

Many thanks go to my to current colleagues Dimitri Guala for shared interests regarding photography and movies. Big thanks to Daniel Morgan for his friendship as well as shared interested in books and time spend together plotting how to conquer the world. To Stefanie Friedrich for her interesting perspectives during coffee break discussions, as well as to more recent colleagues Deniz Secilmis and Miguel Castresana for their friendship and youthful spirit.

Finally, my deepest gratitude goes to my grandparents Franciszek and Zofia as well as Bronislaw and Maria who mean a lot to me. I am also grateful to my parents Elżbieta and Krzysztof, who raised me and supported me through my whole education. I am grateful to my brother, Tobiasz and my cousin Tomasz for being wonderful family. Of course, none of these would be possible without patience and love from my fiancee Katarzyna, who gave me the confidence and space in proceeding with my research, and also supported me in all ups and downs. I would like to comfort her that this is my last PhD.
References


[56] Adrian M. Altenhoff, Romain A. Studer, Marc Robinson-Rechavi, and Christophe Dessimoz. Resolving the Ortholog Conjecture: Orthologs Tend to Be Weakly, but Significantly, More Similar in Function than Paralogs. PLOS Computational Biology, 8(5):1–10, 05 2012. 22, 32


