Computational prediction of gene function under the Open World Assumption

A dissertation submitted to attain the degree of

Doctor of sciences of ETH Zurich
Dr. sc. ETH Zurich

presented by

Nives Škunca
Faculty of Food Technology and Biotechnology, Zagreb,
Croatia
Master of Bioprocess Engineering

born on February 16, 1983
citizen of Croatia

accepted on the recommendation of
Prof. Dr. Gaston H. Gonnet, examiner
Prof. Dr. Christine Orengo, co-examiner
  Prof. Dr. Jörg Stelling, co-examiner
Prof. Dr. Amos Bairoch, co-examiner
Dr. Christophe Dessimoz, co-examiner

2014
This thesis is lovingly dedicated to my parents, Snježana and Živko. Their support and encouragement made the pursuit of my PhD possible.
Hodaj! Nebo strpljive voli. Hodaj! Možda se ipak sve u dobro pretvori. —Zlatan Stipić Gibonni
Abstract

The abundance of newly sequenced genomes comes with a challenge: unlike sequencing, discovering the function(s) of a gene remains painstaking work that is largely restricted to a handful of model species. In fact, less than 1% of the available function annotations are based on direct experimental evidence [38]. It is the remaining 99% that is the focus of my thesis: computational predictions of function, often the only function annotations available for non-model species.

For computational predictions to be useful, the first prerequisite is establishing their quality; normally, by comparing predictions to a selected subset of existing information in function annotation databases. However, incompleteness of the annotation databases hinders evaluation: databases capture only a subset of the information available in the literature; the literature itself is constantly being amended with new information. Therefore, absence of evidence of function in the database does not indicate an erroneous computational prediction.

It is the influence of the incompleteness of the annotation databases that I explored in my first thesis contribution [32]. I showed that when the computational prediction is not known to be true, assuming it is wrong might have significant consequences on the ranking of different methods used in computational prediction of function.

In the second of my thesis contributions [132], I avoided the pitfalls of the incompleteness of the annotation databases by using an experimental validation of predictions. First, I created a computational function prediction method based on phylogenetic profiling that includes both orthologs and paralogs—homologs separated by a speciation and a duplication event, respectively. I showed that the phylogenetic profiling-based model that includes both orthologs and paralogs provides more annotations at the same average Precision than the model that includes only orthologs. Second, experimental assays in the model organism Escherichia coli showed that my function prediction method provides a realistic assessment of Precision for the predicted annotations: a growth phenotype screen on E. coli knockout mutants indicated an overall Precision of 66%—out of 38 tested, 25 confirmed predictions—agreeing with the expected Precision of 60%.

Although experimental verification is arguably the most direct mode of validating predictions, it is prohibitively expensive even for a small subset of the available computational annotations—there are over 200 million computational annotations in the November 2013 release of the Gene Ontology Annotation database alone. Instead, in the third of my thesis contributions [115], I sought to exploit the existing, but newly available experimental annotations to evaluate computationally predicted annotations.
As a surrogate for the intuitive notion of correctness, I defined Reliability as the ratio of confirmed computational annotations to confirmed and rejected/removed ones. One computational annotation is deemed confirmed or rejected, depending on whether a new, corresponding experimental annotation supports or contradicts it. Furthermore, if a computational annotation is removed, the annotation is deemed implicitly rejected and thus contributes negatively to the reliability measure. As a surrogate for the intuitive notion of sensitivity, I defined Coverage as the proportion of newly added experimental annotations that had been correctly predicted by computational annotations in a previous release.

Overall, I found that electronic annotations are more reliable than generally believed, to an extent that they are competitive with annotations inferred by curators when they use evidence other than experiments from primary literature. But I also reported significant variations among inference methods, types of annotations, and organisms, thereby providing guidance for GO users and laying the foundations to further improve computational approaches of GO function prediction.

The successes of computational function prediction methods can be attributed to a combination of methodological refinements, an increase in the number of sequenced genomes, and an increase in the number of functional annotations. Whereas most of the previous literature on function prediction methods has focused on methodological refinements, relatively little is known about the contribution of more genomes and more function annotations.

In the fourth of my thesis contributions (currently under review), I explored the extent to which newly sequenced genomes and new information in the annotation databases influence a well-established approach for function prediction, phylogenetic profiling. The intuition behind phylogenetic profiling is that genes co-occurring in different genomes could be involved in a common function by 1) being involved in the same biological pathway which is therefore incomplete without all its members in a given genome, and/or 2) being beneficial for the phenotype in a particular environment.

What I found is that phylogenetic profiling generally benefits from an increased amount of input data. However, by decomposing this improvement in performance in terms of the contribution of additional genomes and of additional annotations, I observed diminishing returns in adding more than \( \sim 100 \) genomes, whereas increasing the number of annotations remained strongly beneficial throughout. I also observed that maximising phylogenetic diversity within a clade of interest improves performance, but the effect is small compared to changes in the number of genomes under comparison. Finally, I showed that these findings are supported in light of the OWA.

Overall, in my thesis I focused on computational methods to infer gene function. In addition to presenting my own method for computational function prediction, I introduced a novel metric to benchmark the quality of an established database hosting function predictions. I quantified the contribution of the growing set of sequenced genomes, as well as the new annotations, all the while tackling the issue of the Open World Assumption, which posits that functional annotation databases are inherently incomplete.
Zusammenfassung

Mit der Fülle an neu sequenzierten Genomen stellt sich eine neue Herausforderung: Anders als die Sequenzierung bleibt die Entdeckung der Funktion(en) eines Gens mühsame Arbeit, welche sich grösstenteils auf eine Handvoll Modellspezies beschränkt. Tatsächlich basieren weniger als 1% der verfügbaren Funktionsannotationen auf direkter experimenteller Evidenz. Der Schwerpunkt meiner Arbeit sind die restlichen 99%: Computervorhersagen, welche von unschätzbarem Wert sind um die Flut von Sequenzdaten zu verstehen, mit der wir konfrontiert sind.


Die Erforschung des Einflusses der Unvollständigkeit von Annotations-Datenbanken ist der erste Beitrag dieser Arbeit. Ich konnte zeigen, dass, wenn nicht bekannt ist, ob eine rechnerische Vorhersage wahr ist, die Annahme, sie sei falsch (z.B. die Anwendung der Closed World Assumption (CWA)), erhebliche Auswirkungen auf die Rangfolge verschiedener Methoden zur rechnergestützten Vorhersage haben kann.

E. coli ergab eine Genauigkeit von 66 Prozent—von 38 getesteten Vorhersagen wurden 25 bestätigt—which mit der erwarteten Genauigkeit von 60 Prozent übereinstimmt.

Auch wenn die experimentelle Prüfung wohl der direkteste Weg ist, um Vorhersagen zu bestätigen, ist sie bereits für ein kleines Subset der verfügbaren rechnergestützten Annotationen — alleine die Gene Ontology Annotation Datenbank von November 2013 enthält über 200 Millionen rechnergestützte Annotationen — unerschwinglich teuer. Deshalb versuche ich im dritten Beitrag meiner Arbeit bestehende, jedoch neu verfügbare experimentelle Annotationen auszunutzen, um Annotationen, die auf rechnergestützten Vorhersagen basieren, zu beurteilen.

Als Ersatz für die intuitive Idee von Genauigkeit ('correctness') habe ich die Vertrauenswürdigkeit ('Reliability') definiert als das Verhältnis von bestätigten berechneten und verworfenen oder gelöschten Annotationen. Eine berechnete Annotation wird als bestätigt oder als verworfen gewertet, je nachdem ob eine neue experimentelle Annotation eine bestehende berechnete Annotation unterstützt oder dieser widerspricht. Wenn eine berechnete Annotation gelöscht wird, wird sie als implizit verworfen gewertet und trägt damit negativ zur Messgröße der Vertrauenswürdigkeit bei. Als Ersatz für die intuitive Idee von Empfindlichkeit ('sensitivity') definiere ich die Abdeckung ('coverage') als Anteil von neu hinzugefügten experimentellen Annotationen, welche in einer früheren Version der Datenbank von einer berechneten Annotation korrekt vorhergesagt worden sind.


Ich konnte zeigen, dass eine Vergrößerung der Datenmenge generell die phylogenetische Profilerstellung verbessert. Die Unterteilung der Performanceverbesserung in Beiträge aus zusätzlichen Genomen und zusätzlichen Annotationen zeigte allerdings, dass das Hinzufügen von mehr als ~ 100 Genomen kaum weitere Verbesserungen bringt, wohingegen die Güte der phylogenetischen Profilerstellung beständig von zusätzlichen Annotationen profitiert. Weiter konnte ich beobachten, dass eine hohe phylogenetische Diversität innerhalb einer monophyletischen Gruppe die Performance verbessert, aber dass dieser Effekt gering ist im Vergleich zum Einfluss der Zahl der verglichenen Genome. Zu guter letzt konnte ich zeigen, dass diese
Ergebnisse im Lichte der Open World Assumption gültig sind.

Acknowledgements

I am using this opportunity to express my gratitude to everyone who supported me while pursuing my PhD.

I would like to thank Prof. Gaston H. Gonnet, my thesis supervisor. Joining the Computational Biochemistry Research Group (CBRG) was a turning point in my career, giving me the opportunity to work at one of the most prestigious universities in the world on what I consider the most interesting bioinformatics problems. I am particularly thankful for the scientific freedom he allowed me. While pursuing my own research projects, I found our discussions helpful and stimulating: his illuminating views often sharpened my thinking and certainly improved the quality of my research.

Dr. sc. Christophe Dessimoz played a key role at every stage of my thesis. Meeting him at the ISMB conference in 2009, I could not have dreamed that the first friendly encounter would lead to a successfully defended PhD thesis at ETH Zurich. As a scientist, he is what I aspire to being. In particular, I admire his grasp of a wide range of scientific topics, as well as his ability to distil the essentials from a piece of work. He is generous with praise and pointed with criticism. It is because of his professional, as well as personal treats that I am proud to have worked under his guidance and I hope that our collaboration so far is only the beginning.

Additionally, I would like to thank the remaining members of my committee: Prof. dr. Christine Orengo, Prof. dr. Jörg Stelling, and Prof. dr. Amos Bairoch. It was a pleasure to engage in a discussion with scientists I deeply respect during my thesis defence.

I am grateful to have an admirable set of co-authors. I consider Dr. sc. Adrian Altenhoff both a colleague and a friend. His optimism, kindness, wit, and willingness to engage in every adrenalin-raising activity known to man continue to impress. Dr. sc. Fran Supek was there when I was first introduced to the field of bioinformatics; he helped to cement my admiration for this interdisciplinary scientific field and is among the most brilliant young scientists I have the pleasure of knowing. Dr. sc. Tomislav Šmuc supported and mentored me during my time at the Rudjer Bošković Institute, for which I am immensely grateful. I further extend my gratitude to Matko Bošnjak, Dr. sc. Sašo Džeroski, Dr. sc. Panče Panov, Dr. sc. Anita Kriško, and Louis du Plessis. Their ideas, experiments, and text contributed to making my thesis a work I can be proud of.

I extend my gratitude to the members of the CBRG at ETH Zurich, both past and present, as well as the members of the Dessimoz Lab at UCL. I am thankful for their help with all things technical, their constructive criticism, friendship, laughs, conversation, understanding, and patience. Naming everyone, I would surely miss someone. But if I have ever made a cake for
you, you know you belong on this list.

I am grateful for the wonderful Lean in ladies: Anwar, Elke, Fiona, Julia, Katarina, Tahmineh, Aparna, and Franziska. They are a collection of young ambitious women, taking the proverbial bull by the horns, and they will always be my inspiration.

Vincent is the most wonderful serendipity of my coming to Zürich, as this is where we met. He is my rock, my inspiration, and the source of much of my joy. The PhD life has become much easier and much more pleasurable since we met. Our relationship gives me a first-row seat to explore the clashes of cultures; these clashes are more often a source of laughter than of pain, and they are what makes us... us.

Finally, I would like to thank my family, my parents Snježana and Živko and my sister Monika. There were many difficult moments during my PhD that were made much easier with the support of my family.
## Contents

<table>
<thead>
<tr>
<th>Section</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Abstract</td>
<td>iii</td>
</tr>
<tr>
<td>Zusammenfassung</td>
<td>v</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>ix</td>
</tr>
<tr>
<td>Introduction</td>
<td>1</td>
</tr>
<tr>
<td>1 General Background</td>
<td>5</td>
</tr>
<tr>
<td>1.1 Annotating an avalanche of sequences</td>
<td>5</td>
</tr>
<tr>
<td>1.2 Gene Ontology (GO): a vocabulary for function annotation</td>
<td>6</td>
</tr>
<tr>
<td>1.2.1 What is the GO?</td>
<td>7</td>
</tr>
<tr>
<td>1.3 Where do annotations come from?</td>
<td>8</td>
</tr>
<tr>
<td>1.3.1 Inferred from experiment</td>
<td>9</td>
</tr>
<tr>
<td>1.3.2 Inferred from author statement</td>
<td>9</td>
</tr>
<tr>
<td>1.3.3 Curator statement evidence codes</td>
<td>10</td>
</tr>
<tr>
<td>1.3.4 Inferred using a computational method</td>
<td>10</td>
</tr>
<tr>
<td>1.3.5 Importance of evidence and qualifiers</td>
<td>11</td>
</tr>
<tr>
<td>1.4 Computational methods for function prediction</td>
<td>12</td>
</tr>
<tr>
<td>1.4.1 Inferring function from sequence or structure</td>
<td>12</td>
</tr>
<tr>
<td>1.4.2 Inferring function from orthology</td>
<td>13</td>
</tr>
<tr>
<td>1.4.3 Inferring function from ancestors or descendants</td>
<td>14</td>
</tr>
<tr>
<td>1.4.4 Inferring function from sequence models</td>
<td>14</td>
</tr>
<tr>
<td>1.4.5 Inferring function from genomic context</td>
<td>15</td>
</tr>
<tr>
<td>1.4.6 Why use the GO?</td>
<td>15</td>
</tr>
<tr>
<td>1.5 Conclusion</td>
<td>17</td>
</tr>
<tr>
<td>2 The open world of protein function prediction</td>
<td>19</td>
</tr>
<tr>
<td>3 Quality of computationally inferred Gene Ontology annotations</td>
<td>23</td>
</tr>
<tr>
<td>3.1 Results</td>
<td>24</td>
</tr>
<tr>
<td>3.1.1 Electronic annotations in subsequent UniProt-GOA releases are increasing in quality</td>
<td>25</td>
</tr>
<tr>
<td>3.1.2 GO term’s specificity is only partially indicative of the reliability of electronic annotations</td>
<td>26</td>
</tr>
<tr>
<td>3.1.3 The three ontologies have similar reliability, but different coverage</td>
<td>27</td>
</tr>
<tr>
<td>3.1.4 Different sources provide annotations of different quality</td>
<td>29</td>
</tr>
</tbody>
</table>
Introduction

New genomes are sequenced at a high pace and there are close to 19 000 complete genome projects currently available [87]. This abundance of sequence data allows insight into the genetics of a wide range of organisms, ranging from eukaryotes to cultured microbes, as well as metagenomic collections.

Unlike sequencing however, discovering the function(s) of a gene remains painstaking work that is largely concentrated in a handful of model species. To understand this abundance of sequence data, computational approaches for function prediction are invaluable. This fact is reflected in the annotation databases: over 98% of the currently available Gene Ontology (GO) annotations were inferred computationally [38]. It is therefore critical to understand the strengths and weaknesses of different function prediction methods through benchmarking efforts.

The goal of benchmarking is to estimate the performance of different predictors when applied to unannotated genes. This estimation is often done using a validation set, using either a fraction of annotations hidden from the predictors or new annotations that arrive to the database after the predictions had been made.

However, despite being independent, the validation set is not complete: the annotation databases capture only a subset of information available in the literature; the literature itself is continuously being updated with new information. Indeed, most proteins are revisited with annotations after they had been first added to the database [32]. Therefore, absence of evidence of function in the database does not imply evidence of absence of function [123]. In formal logic, this view on the database is called the Open World Assumption (OWA).

As a consequence of the OWA, we can confirm function prediction when the annotation is available in our validation set, and we can reject a prediction when its negation is available, e.g., via the NOT qualifier in the Gene Ontology database. If negative annotations are sparse, as is often the case, to reject a prediction, we often take the opposite view on the database, the Closed World Assumption (CWA): when the predicted annotation is not available in our validation set, we assume it is wrong. Despite being valid for certain types of databases—e.g., a company’s employee database can provide information of a person not being employed there—CWA is not valid for biological databases. The possible extent of the benchmarking error associated with the CWA is shown in one of my thesis contributions: by assuming the CWA, we might be rejecting as many as 14% of predictions that later prove to be correct [32].

A straightforward but onerous way of avoiding the pitfalls of the CWA is
to experimentally evaluate the predictions: a well-constructed experiment will either confirm or reject the computational prediction.

My second main thesis contribution contains just such an evaluation: first, I created a function prediction method based on phylogenetic profiling that includes both orthologs and paralogs—homologs separated by a speciation and a duplication event, respectively [104]. I showed that the phylogenetic profiling-based model that includes both orthologs and paralogs provides more annotations at the same average Precision than the model that includes only inferred orthologs. With this model, I obtained a large number of computational annotations for microbial genomes with high predictive accuracy in train/test evaluations.

To evaluate how realistic the train/test evaluations were, I chose 38 predicted annotations for genes in *Escherichia coli* K-12. The predictions had at least 60% expected Precision for the three GO terms that were straightforward to investigate experimentally using readily available antibiotics. Experimental results indicated that my prediction method could be used to generate new biological hypotheses: out of the 38 predictions obtained at the reported Precision of 60%, the experiments confirmed 25 predictions. The observed experimental Precision of 66% thus suggests that the prediction method’s confidence estimates can be used to make informed decisions on experimental validation.

Even though experimental verification is a direct mode of validating predictions, experimental verification is prohibitively expensive even for a small subset of the available computational annotations; there are over 200 million computational annotations in the November 2013 release of the Gene Ontology Annotation database [55]. Instead, in the third of my thesis contributions, I sought to exploit experimental annotations newly added to functional databases to evaluate computational annotations.

As a surrogate for the intuitive notion of correctness, I defined *Reliability* as the ratio of confirmed computational annotations to confirmed and either explicitly contradicted or removed ones. One computational annotation is deemed confirmed or rejected, depending on whether a new, corresponding experimental annotation supports or contradicts it. Furthermore, if a computational annotation is removed, the annotation is deemed implicitly rejected and thus contributes negatively to the reliability measure. As a surrogate for the intuitive notion of sensitivity, I defined *Coverage* as the proportion of newly added experimental annotations that had been correctly predicted by a computational annotation in a previous release.

Overall, I found that electronic annotations are more reliable than generally believed, to an extent that they can compete with annotations inferred by curators when they use evidence other than experiments from primary literature. But I also reported significant variations among inference methods, types of annotations, and organisms. This work provides guidance for GO users and lays the foundations for improving computational approaches to GO function inference.

The successes of the above analysed and many other function prediction methods can be attributed to a combination of methodological refinements, [e.g., 28, 138, 26, 115], an increase in the number of sequenced genomes, and an increase in the number of function annotations—GO annotations in particular [38]. However, whereas most of the previous literature on
function prediction methods has focused on methodological refinements, relatively little is known about the contribution of more genomes and more functional annotations.

In the fourth of my thesis contributions, I explored the extent to which newly sequenced genomes and new information in the annotation databases influence a well-established approach for function prediction, phylogenetic profiling. Phylogenetic profiling is based on patterns of gene presence and absence across species; it has been successfully used to discover protein-protein interactions and to annotate new genomes, particularly in prokaryotes [reviewed in 63].

I focused on the phylogenetic profiling method presented in my third thesis contribution and found that phylogenetic profiling generally benefits from an increased amount of input data. However, by decomposing this improvement in performance in terms of the contribution of additional genomes and of additional annotations, I observed diminishing returns in adding more than \( \sim 100 \) genomes, whereas increasing the number of annotations remained strongly beneficial throughout. I also observed that maximising phylogenetic diversity within a clade of interest improves performance, but the effect is small compared to changes in the number of genomes under comparison. Finally, I showed that these findings are supported in light of the OWA.

In my thesis I focus on computational methods to infer gene function. In addition to presenting my own method for gene function prediction, I introduced a novel metric to benchmark the quality of an established database hosting predicted annotations. I quantified the contribution of the growing set of sequenced genomes, as well as the new annotations, all the while tackling the issue of the Open World Assumption, which posits that functional annotation databases are inherently incomplete.

**Outline of the Thesis**

Chapter 1 starts by reviewing basic concepts that are relevant for function annotation in general and computational annotation in particular. I discuss the current state of the literature and emphasize arguably the most important vocabulary for functional annotation, Gene Ontology, as well as the state-of-the-art methods for computational functional annotation.

In chapter 2, I introduce the concepts of the Open and the Closed World Assumptions with particular emphasis on the significance of each of these assumptions in biological databases and the effect they have on the evaluation of computational annotations.

In chapter 3, I introduce a novel metric to measure the quality of computational annotations. I show the results of my evaluation and discuss the implications of my findings on the future use and the development of computational annotations.

Chapter 4 introduces my own method of computational functional annotation. The main novelty of the method lies in the construction of the dataset: contrary to the common practice of including only orthologs in phylogenetic profiling, I show that using both orthologs and paralogs provides superior performance in train/test evaluations. Importantly, I show
that the train/test evaluation of the method shows realistic estimates of predictive accuracy.

Finally, in chapter 5, I evaluate the influence of, not methodological improvements, but new data on the performance of phylogenetic profiling. I separately evaluate the influence of newly sequenced genomes and new annotations and show the potential of computational functional annotation with the increase in the available data.
Chapter 1

General Background

Parts of this chapter were derived from the peer-reviewed publication: du Plessis L, Skunca N, Dessimoz C. The what, where, how and why of gene ontology—a primer for bioinformaticians. Briefings in Bioinformatics. 2011

1.1 Annotating an avalanche of sequences

The sophistication of DNA sequencing technologies is driving the elucidation of a vast amount of biological sequences. Because the genome carries the information for (virtually) everything a cell does, decoding the genome—discovering its sequence—could in principle provide the key to knowing all the processes that take place in a cell. With the reduced costs associated with sequencing, we have high expectations for the potentially useful information in this data deluge.

However, the bulk of the cost now is not in the sequencing, but in the analysis: annotation, curation, visualization, and verification. Because the genomic databases are growing, not only in their size, but also in their diversity, the situation we now face is illustrated with a famous quip: we have the $1,000 genome and the $100,000 analysis [79].

Bridging the gap between sequencing and analysis is the daily work of scientists; making this information easily accessible to the scientific community is the work of biocurators [45], experts that sift through the scientific literature in search of evidence of gene product function for it to be available in the public databases such as SwissProt/UniProt [8]. However, as high quality the curators’ assignments are, their efforts must be supplemented with computational methods: it is the computational methods that can cope with the volume of the new sequences being produced, thereby helping to form new biological hypotheses.

In this chapter I survey methods for computational functional annotation. I start with the detailed description of arguably the most important vocabulary for functional annotation, the Gene Ontology (GO). Next, I describe the main methods used to connect terms from the GO to gene products. Finally, I focus on various methods used for computational functional annotation.
1. General Background

Figure 1.1: Increase in the number of experimentally verified GO term assignments available for the respective organism between September 2002 and September 2010. The GO consortium was initially focused on Eukaryotes, a fact reflected in the distribution and increase of annotations available in the GO database. Contrast for instance the steady growth of experimentally verified annotations for *A. thaliana*, *S. cerevisiae* or *M. musculus* with the sharp increase in the number of experimentally verified annotations available for *E. coli*: from 33 in 2002 to 1852 in 2010.

1.2 Gene Ontology (GO): a vocabulary for function annotation

One of the first steps in extracting knowledge from genome sequencing data is often done through discerning the function of the gene product. However, it is not only the functional annotation that is difficult; defining the vocabulary for function annotation is far from trivial.

The first attempts at classifying gene functions made use of natural language annotations in databases. Early on it was found that natural language by itself is too vague and unspecific to accurately capture the function of genes [16], as it is difficult to perform searches and establish relationships with natural language annotations. The first efforts towards a structured and controlled annotation of genes were schemes such as the Enzyme Classification (EC) system representing the function of an enzyme using a four digit sequence of numbers [125]. Such classification schemes are still widely used but were found to be insufficient to accurately describe gene function. This motivated the introduction of the Gene Ontology (GO) [11], which has grown to be the largest resource of its kind.

The “GO Consortium” consists of a number of large databases working together to define standardized ontologies and provide annotations to the GO. The three ontologies it encompasses are non-redundant and share a common space of identifiers and a well-specified syntax. Apart from
1.2. Gene Ontology (GO): a vocabulary for function annotation

Figure 1.2: The structure of the GO is illustrated on some of the paths of term GO:0060491 to its root term. Note that it is possible for a term to have multiple parents.

providing a standardized vocabulary for describing gene and gene product functions, one key motivation behind the GO was the observation that similar genes often have conserved functions in different organisms. The combination of information from all organisms in one central repository makes it possible to integrate knowledge from different databases and to infer the functionality of newly discovered genes. Originally, the GO was developed for a general eukaryotic cell [11]. The initial GO vocabulary, as well as the available GO term annotations present in the first years of its existence reflect this fact (Figure 1.1). However, the GO Consortium now includes several annotation groups that focus on prokaryotes [52], further contributing to the expansion of the vocabulary and annotations.

1.2.1 What is the GO?

The GO is a structured and controlled vocabulary of terms. The terms are subdivided in three non-overlapping ontologies: Molecular Function (MF), Biological Process (BP) and Cellular Component (CC) [47]. Each ontology describes a particular aspect of a gene or gene product functionality, as well as the relations between the terms. These relations are either is_a, part_of, has_part, or regulates. There are two subclasses of the regulates relationships: positively_regulates and negatively_regulates. The is_a relationship is not used to imply that a term is an instance of another term; instead, it connects a subtype to its more general counterpart (Figure 1.2). The part_of and has_part re-
1. General Background

Relationships are logical complements of each other [48]. The relationships form the edges of a Directed Acyclic Graph (DAG), where the terms are the nodes (Figure 1.2). This allows for more flexibility than a hierarchy, since each term can have multiple relationships to broader parent terms and more specific child terms. Any path from a term towards the root becomes more general as terms are subsumed by parent terms.

Each gene is associated with the most specific set of terms that describe its functionality. By definition, if a gene is associated with a term, it is also associated with all the parents of that term. The process of assigning a GO term to a gene product—annotation—is discussed in more detail in the next section.

The GO undergoes frequent revisions to add new relationships and terms or remove obsolete ones. If a term is deleted from the ontology, the identifier for the term stays valid, but is labelled as obsolete and all relationships to the term are removed [100]. Changes to the relationships do not affect annotations because annotations always refer to specific terms, not their location within the GO.

It is clear that relationships between the three ontologies exist. For example, an instance of a BP ontology is the execution of one or more instances in the MF ontology [51]. Similarly, relationships exist between the MF and CC ontologies. Recently, these relationships have been integrated into the GO by introducing some inter-ontology links [48]. It should be noted that for the moment there are two concurrent versions of the GO, the filtered and the full GO. The main difference is that the filtered GO does not contain any has_part or inter-ontology relationships. Many of the analysis tools can only use the filtered GO. Thus, the full expressiveness of the GO structure is not always available.

1.3 Where do annotations come from?

Annotations connect genes and gene products to GO terms. Each annotation in the GO has a source and a database entry attributed to it. The source can be a literature reference, a database reference, or computational evidence [48, 116]. In addition, there are three qualifiers used to modify the interpretation of an annotation—contributes_to, colocalizes_with, and NOT, making them an integral part of the annotation [100].

Perhaps the most important attribute of an annotation is the evidence code. The 22 evidence codes available describe the basis for the annotation (Figure 1.3). These evidence codes are divided into four categories. General guidelines for deciding which evidence code to use are given in the GO evidence code decision tree (http://www.geneontology.org/GO.evidence.tree.shtml), and are briefly described below. It should be kept in mind that one gene can be annotated to the same term with more than one evidence code and that multiple annotations to the same term for the same gene could even share the same reference. This makes it possible to see whether an annotation is supported by more than one type of evidence. However, if the gene is annotated with more than one evidence code and one evidence code is a superclass of another, the annotation with the more general evidence code does not need to be specified explicitly.
1.3. Where do annotations come from?

1.3.1 Inferred from experiment

Annotations considered most reliable are those inferred directly from experimental evidence: the parent evidence code EXP (Inferred from Experiment) and its child terms IDA (Inferred from Direct Assay), IPI (Inferred from Physical Interaction), IMP (Inferred from Mutant Phenotype), IGI (Inferred from Genetic Interaction), and IEP (Inferred from Expression Pattern). Such annotations are important to seed the ontology so that the gene function of related genes can be inferred by computational methods [98]. The largest fraction of manual annotations are made by professional curators examining the literature [75]. Arguably, researchers directly annotating genes they themselves characterized would be more efficient, but this practice has not yet caught on because annotation is time consuming and annotation guidelines are complex [76]. The efforts underway to make it compulsory for authors to submit GO term suggestions with article manuscripts [71] will surely benefit from various overviews of how annotations are made [51].

1.3.2 Inferred from author statement

Annotations in this group fall into two categories: on the one hand, for a Traceable Author Statement (TAS) curators derive annotations from papers having cited the evidence for annotation without presenting the original evidence, e.g., review papers. On the other hand, a Non-traceable Author Statement (NAS) refers to a statement derived from a database entry or a statement in a paper that cannot be traced to another paper.
1. **General Background**

1.3.3 **Curator statement evidence codes**

Two evidence codes fall in this category: Inferred by Curator (IC) and No biological Data available (ND). If an assignment of a GO term is made using the curators expert knowledge, concluding from the context of the available data, but without any “direct” evidence available, the IC evidence code is used. The ND evidence code indicates that the function is currently unknown (i.e., that no characterization of the gene is currently available). Such an annotation is made to the root of the respective ontology to indicate which functional aspect is unknown. Hence, the ND evidence code allows for a subtle difference between unannotated genes and uncharacterized genes. Note that the ND code is also different from an annotation with the NOT qualifier; the latter indicates the absence of a particular function.

1.3.4 **Inferred using a computational method**

There are eleven evidence codes associated with computational inference, ten of which include manual curation: ISS (Inferred from Sequence or Structural Similarity) with its sub-categories ISO (Inferred from Sequence Orthology), ISA (Inferred from Sequence Alignment), and ISM (Inferred from Sequence Model); IGC (Inferred from Genomic Context); IBA (Inferred from Biological aspect of Ancestor); IBD (Inferred from Biological aspect of Descendent); two evidence codes only to be used in conjunction with the NOT qualifier, IKR (Inferred from Key Residues) andIRD (Inferred from Rapid Divergence); and finally RCA (Inferred from Reviewed Computational Analysis).

The eleventh, the evidence code IEA (Inferred from Electronic Annotation) is used for all automatic inferences made without any human supervision, regardless of the method used. Consequently, IEA is by far the most abundantly used evidence code (Figure 1.4). The largest single contributor of IEA annotations is the UniProt Gene Ontology Annotation.
1.3. Where do annotations come from?

(UniProt-GOA) database [35]. The computational annotations available in the UniProt-GOA database stem from three main types of sources: 1) mappings of the available vocabularies—e.g., Enzyme Commission numbers—to the GO vocabulary, 2) methods that assign GO terms based on sequence and structure signatures, and 3) methods that propagate GO terms based on the assumed functional equivalence of homologs or groups of orthologs. Among the three types, the first one does not assign new annotations: the existing annotations are only translated to the GO vocabulary. The other two types do contribute new annotations, and are based on computational methods for functional annotation that are accepted in the community for their good coverage, albeit not always for their accuracy [110]. Sequence and structure signatures contribute the largest number of new computational annotations in UniProt-GOA via the InterPro resource [53]. InterPro uses different databases to predict the presence of domains and important sites that are used to group proteins into families; for an unannotated protein, family membership is used to infer function. Another source of new computational annotations in UniProt-GOA are the orthologous relations between proteins, based on the Ensembl Compara pipeline: we assume that one to one and apparent one to one orthologs have the same function and use these relations for the propagation of function. Similarly, orthologous microbial protein families of the HAMAP project are annotated with rules assigned by curators: membership to a HAMAP group implies function. In addition to them being used in the UniProt-GOA database, the basic principles of all methods described above are extensively used in the community, as I briefly outline below.

1.3.5 Importance of evidence and qualifiers

GO annotations should always be considered with their qualifier and evidence code in mind. A qualifier such as NOT changes the interpretation of an annotation. For example, the two evidence codes IKR and IRD are intended for use only with the NOT qualifier: the former when inferences are made from the lack of key residues in the protein necessary for the function and the latter when rapid divergence from ancestral sequence implies divergence in function. Because the biological (and in particular biomedical) literature is still heavily biased towards reporting positive results [36, 80], biological databases contain very few instances for evidence of lack of function. One recent paper described a method to infer negative annotations [142], opening the possibility to introduce computational NOT annotations.

Similarly, although the evidence code is not a direct measure of the quality of the annotation, some evidence codes are regarded as more trustworthy. For example, terms annotated with ND are typically ignored as no knowledge is available on the function of these genes. Similarly, annotations based on experimental evidence are often considered the most trustworthy. Most studies disregard all annotations assigned automatically (evidence code IEA) [96, 77, 112, 29], consequently leaving out more than 98% of the annotations in the GO (Figure 1.4). The mistrust many researchers have towards computational annotations is backed by studies suggesting that annotations from the available databases should be used with caution [110]. However, computational annotations are useful in providing the
1. General Background

first approximation of function to experimental biologists; this is one of the reasons why I explore the quality of IEA annotations in chapter 3 of my thesis.

1.4 Computational methods for function prediction

Wanting to understand the abundance of data from a new genome sequencing project—connecting the genome to the phenotype—one of the first stops for a researcher is to turn to computational methods for function prediction. In the beginning, computational methods were little more than comparing the unknown sequence to a reference database, in most cases using the BLAST algorithm [6]. Sequence alignment is still one of the first tools employed for preliminary function prediction—e.g., by searching a reference database for similar proteins—and is very often used as a baseline when evaluating more modern approaches [95].

In a typical analysis, characterized and uncharacterized genes are clustered based on sequence similarity measures and phylogenetic relationships. The function of unknown genes is then inferred from the function of characterized genes within the same cluster [e.g., 69, 113].

However, the range, the number, and the sophistication of the available computational methods for function prediction has grown to such an extent that a comprehensive overview of literature would exceed the number of pages in this thesis (and indeed it does [89]); I will instead use the GO evidence codes outlined above as a guide, and briefly describe some of the representative methods used in computational function annotation.

1.4.1 Inferring function from sequence or structure

The guiding principle of computational annotation methods based on sequence or structure similarity is that protein sequence determines its structure which in turn determines the function. Proteins with similar sequences or structures are likely to be evolutionarily related, and thus, assuming they largely kept their ancestral function, they might still have similar functional roles today. Shown to be true for many proteins, in particular those with high sequence identity (and always having important exceptions [e.g., 34, 126]), sequence/structure similarity remains arguably the easiest and most abundantly used method for computational function annotation.

Pairwise sequence alignment using tools such as BLAST or PSI-BLAST [6, 7] produces similarity matches for the query sequence; if the match has been functionally characterized, its function could give us information on the function for the query sequence [e.g., 139, 1]. Multiple sequence alignment using tools such as ClustalW [124] and MUSCLE [40] is even more powerful: identifying key features of the sequence responsible for function is more reliable when multiple sequences point to the same key sequence position [94].

The most prominent database containing multiple sequence alignments is Pfam [42]. The goal of Pfam is producing protein families that will encompass as much of the sequence space as possible. Sequences are grouped
into families: first, curators produce a high-quality seed alignment; second, the profile Hidden Markov Model (HMM \cite{39}) is built; third, the HMM is searched against the reference database derived from UniProt Knowledgebase (UniProtKB); and fourth, the application of family-specific sequence and domain gathering thresholds forms protein families (Pfam-A). With the ability to view multiple sequence alignments as well as protein domain architecture, a researcher can infer sequence homology and make reasonable hypotheses of protein function.

When the structure of a protein is known, structure motifs provide a good method to infer function: functionally important structure motifs are generally better conserved than sequence motifs. For instance, Liu et al. \cite{72} introduced a method that makes use of the similarity of protein surface pockets to infer GO terms related to the protein. However, sequence to function relationships break for some very frequent motifs. For example, TIM barrel fold is present in 27 different homologous superfamilies—groups of evolutionarily related proteins—that cover over 60 different EC classifications \cite{66}.

Alternatively, methods based on protein profiles account for the fact that sequence conservation might be very uneven across the length of two functionally related genes. This is because the function of a protein is often dictated not by the shape and structure of the whole protein, but rather by specific regions and residues, such as catalytic sites, prosthetic group attachment sites or other binding sites \cite{113}.

As Rentzsch and Orengo argue \cite{99}, one of the biggest challenges of automated function prediction is choosing the right threshold beyond which function can be propagated. Using a predefined cut-off level is not a good practice as the optimal threshold will vary depending on which genes are evaluated. Indeed, there are several instances of proteins with high sequence similarity but different functions, and conversely, of proteins with similar function but highly divergent sequences \cite{93, 9} and the user of annotations derived by computational methods should always have these intricacies in mind.

### 1.4.2 Inferring function from orthology

When doing function propagation via sequence similarity, we are implicitly propagating through homology—ancestry of sequences. There are two main subtypes of homologs: orthologs, homologs originated through speciation and paralogs, homologs originated through duplication \cite{64}. According to the standard model of genome evolution, paralogs—because they diverged through a duplication event—could obtain a new function. Conversely, orthologs diverged through a speciation event, meaning that the function in the descendant species should have been retained; orthologs are therefore expected to be more useful in functional annotation.

However, the exact quantification of the functional divergence in a pair of orthologs and a pair of paralogs is not fully resolved. It was observed that the search for homologs using the best bidirectional hit approach, without explicitly distinguishing orthologs from paralogs, produces a higher level of functional compactness via GO terms than is present in the ortholog databases Homologene and OMA \cite{2}. In addition, Studer and Robinson-
1. General Background

Rechavi list scenarios where the standard model—predicting that paralogs diverge in function more than orthologs—is invalid; for example, cases where paralogs share function, and orthologs do not [118].

A recent large-scale study further challenged the veracity of the standard model: the authors compared mouse and human ortholog and paralog pairs and surprisingly found that paralogs tend to conserve function more than orthologs [84]. This finding caused a stir in the community—demonstrating the relevance of the topic—but was subsequently challenged in two publications [4, 123].

Still, when used with caution, methods relying on sequence orthology (and paralogy) can give considerable insight in protein function [e.g., 122]. In fact, the largest database of computational functional annotation, UniProt-GOA [24] includes inferences based on orthology among the predictions: both for eukaryotes (Ensembl [44]) and prokaryotes (HAMAP [91]). In addition, many existing orthology databases can be used to generate hypotheses of function, be it by propagating function between pairs of orthologs (e.g., [86]) or among orthology group members ([e.g., 104, 69, 122]). The available literature reviews (e.g., [2]) and reports of the Quest for Orthologs consortium [31] provide a more in-depth discussion of the different methods used to infer orthology.

1.4.3 Inferring function from ancestors or descendants

IBA and IBD are among the most recently added GO evidence codes. Their addition reflects the interest in phylogenetic-based propagation of functional annotations [41].

Phylogenetic-based function inference starts with a phylogenetic tree, a gene tree. If the nodes in the gene tree are labelled with, e.g., gain or loss of function events, the function of some genes in the leaves of the tree can give clues about the function of other genes in the tree. Gaudet et al. describe a tool for just such a functional annotation [46]. In their approach, the role of the curator is crucial: he or she will check the gene tree for consistency with the available knowledge and subsequently make inferences of function for unannotated genes, based on the function of their biological ancestors or descendants, all the while taking into account the available literature evidence.

1.4.4 Inferring function from sequence models

Another approach to function prediction consists of supervised machine learning based on features derived from protein sequence [23, 68, 111, 73]. Such methods use a training set of classified sequences to learn features that can be used to infer gene functions. Although few explicit assumptions about the complex relationship between protein sequence and function are required, the results are dependent on the accuracy and completeness of the training data.
1.4.5 Inferring function from genomic context

In addition to finding two homologous proteins (or a group of homologs) and propagating function between them (or among them), we can include the information on the presence or absence of homologs within different genomes. In a now classic paper, Pellegrini et al. first described just such a method to analyse biological processes on a genome scale—phylogenetic profiling [92]. The intuition behind phylogenetic profiling is that proteins kept/lost together in different organisms might be involved in the same biochemical pathway, giving us a rationale for functional propagation between proteins with similar phylogenetic profiles. By tracking the patterns of presence or absence of *Escherichia coli* genes in different organisms, Pellegrini et al. showed evidence that proteins with similar patterns of presence or absence—similar phylogenetic profiles—tend to share functional annotations. Many extensions of the method [e.g., 119, 83, 61] indicate its usefulness. It is this approach that I use in two chapters of my thesis: first, in chapter 4 I introduce a novel method based on phylogenetic profiling, and then in chapter 5 I evaluate the influence, not of methodological improvements, but of the increase in the available data.

Various studies looked at aspects of phylogenetic profiling such as the genomes included, vocabulary used in functional annotation, methods used to find or group orthologs, and methods used to find similar phylogenetic profiles [see 63, for a review]. For example, looking at the genomes included, one study suggests that reference genomes should be selected from moderately and highly genetically distant organisms, from all three domains of life [119]. Another study noted a noticeable drop-off in performance with increased number of Eukaryotes [61]. In addition, Jothi et al. note that the over-representation of parasitic Eukaryotes and vertebrates additionally make Eukaryotes less useful in reference sets. Another study showed that phylogenetic profiles (and some other genome context methods) substantially improve in performance when a subset of phylogenetically diverse Archaeal genomes was used with Eubacteria [83]. Regarding functions that can be predicted using phylogenetic profiles, those related to translation—a very conserved pathway in the cell—should have high predictive accuracy when the phylogenetic diversity of organisms included in the profiles is highest [61]. Jothi et al. noted that performance is the worst for the translation system when they use only the superkingdom to which the test-organism belonged: e.g., when the test organism is *E. coli* and the superkingdom is Bacteria. In line with these results, in a recent benchmark study of genome-context methods based on KEGG pathways, Muley and Ranjan showed that phylogenetic profiling performs well for genetic information processing pathways (“translation,” “folding, sorting, and degradation,” and “replication and repair”) [83]. As expected, pathways involved in motility were best predicted by phylogenetic profiling: these pathways are restricted to motile organisms and show a strong co-occurrence pattern.

1.4.6 Why use the GO?

The GO vocabulary and the respective annotations are used in many different applications. For example, in a database of GO terms, one can look
1. General Background

up genes with similar functionality or location within the cell [9]. In this way a search for interacting genes in one organism or similar genes in two organisms can be narrowed down. If the terms associated with genes are too specific, more general parent terms can give a snapshot of an organism’s gene functions, thereby making a comparison easier [116]. Another standard use of the GO is to reason across the relations using an automatic logical inference tool [116]. Inferences can be made by following simple rules, for example, if A is a B and B is part of C then A is also part of C. In this way relations that are not immediately obvious, but captured in the GO, can be automatically uncovered.

The GO is frequently used to analyse the results of high-throughput experiments. One possibility is to infer the location or function of genes that are over- or under-expressed [e.g., 100]. In functional profiling the GO is used to determine which processes are different between sets of genes. This is done by using a likelihood-ratio test to determine if GO terms are represented differently between the two gene sets [e.g., 100]. Both hypothesis-generating and hypothesis-driven queries can be addressed in this way. In hypothesis-generating queries, the goal is to find which terms are significantly different between the sets, whereas in hypothesis-driven queries, it is to test if sets of terms are different. In hypothesis-generating queries a multiple-test correction needs to be applied, but because of the amount of terms in the GO, the power of the test is significantly reduced. Hypothesis-driven queries do not require any multiple-test corrections. To reduce the effect of doing a multiple-test correction the number of tests done in hypothesis-generating queries need to be minimized and this is commonly done by running the query on a GO slim. A GO slim ontology is a reduced subset of general terms [e.g., 100]. The annotations for a set of genes can then be mapped onto the GO slim. Because of the structure of the GO, an annotation may be mapped to many terms in the GO slim.

Another standard use of GO slims is to give a high-level categorization of genes based only on the terms within the GO slim. It should be added that there are two types of GO slims. The first type is a subset of the GO used to facilitate the examination of a particular taxon subdivision. The second type is a set of broad GO terms used to aggregate the GO into large bins used for the representation of annotation data.

Additionally, and most interesting in the context of my thesis, the GO is used to infer the function of unannotated genes. Genes that behave similarly to an unannotated gene are identified from the experiment and their function is evaluated to be transferred to the unannotated gene. Many of these assigned terms will be false positives, but the correct terms should appear more often than is dictated by chance or indirect effects [52].

The GO is also used in inferring protein-protein interactions (PPI) [29]. Shin et al. [112] used both PPI and GO data to show that interacting proteins are co-located within the cell. Another approach is to test the validity of inferred PPI networks by looking at the functional similarity of genes within the GO [96]. In this case, it is important that annotations inferred from previous PPI studies (IPI evidence code) are left out from the analysis.
1.5 Conclusion

Assigning functional information to biological sequences—proteins or nucleic acids—is an essential step in their analysis: knowing the function of a biological sequence is a stepping stone for further analyses, e.g., exploring functional conservation in groups of orthologs, relating biological sequences to phenotypic features such as diseases in humans or pathogenic markers in prokaryotes, and improving the yield or modifying the resulting metabolite in a biotechnological process. To describe function, the Gene Ontology (GO) remains arguably the most important vocabulary.

Based on their source, there are two main types of functional annotations in the annotation databases: those assigned based on experimental evidence, and those assigned based on computational methods, either with or without a curator’s input. For the first type, dedicated curators read and analyse many published papers that describe the function of biological sequences to fill the extant functional annotation databases. The structure imposed by the databases is then amenable to the analyses by the computation annotation community. For the second type, various algorithms that account for the sequence, structure, or evolutionary history of a biological sequence infer the function.

The number of associations in the GO has grown exponentially since its inception. There were 30,654 associations on 1 July 2000 and 7,781,954 associations on 1 July 2003 [17]. This number had grown to more than 16 million in 2007 [100] and more than 55 million in 2010. Due to the inference methods used, most of the growth has been from IEA associations. In contrast, the curated associations component has only grown linearly. The ontology itself has also been steadily growing, from less than 5,000 terms in 2000 [17] to more than 30,000 in 2010. The Reference Genome Project has been initiated to focus the annotation efforts of various groups on a number of predetermined homologous genes [46]. This will not only help in seeding the ontology, but through a concentrated effort on certain branches the overall structure of the ontology will also be improved.

Computational functional annotation is important for generating new biological hypotheses; evaluation of its predictive accuracy is as crucial as it is challenging. Often, the only confirmation (or disproof) for a prediction can come from laborious wet-lab experiments. Still, it is imperative to benchmark methods for computational functional annotation.

One of the challenges in benchmarking is the lack of explicit negative annotations for protein function: even when we have abundant data on what a protein does, we most often may not claim that this is an exhaustive list of its functions. This is a drawback when evaluating methods for computational annotation: when a computational method provides us with a candidate annotation, and the annotation is not recorded in the extant databases, is this because the computational prediction is wrong, or is it because we do not yet know of this annotation? Because computational methods for functional annotation are at the core of my thesis, the following chapter discussed this dichotomy, referred to as the Closed vs. Open World Assumption; subsequent chapters present an evaluation of the database that is the largest single contributor of computational functional annotations, a novel method for computational functional annotation, and an evaluation of
1. General Background

the influence of new data on predictive accuracy of computational methods for functional annotation.
Chapter 2

The open world of protein function prediction

Computational prediction of protein function is among the most important unresolved problems in bioinformatics and consequently benchmarking is a crucial step when trying to understand the strengths and weaknesses of prediction methods.

Towards this end, the first “critical assessment of protein function annotation” (CAFA) experiment was recently reported, a community-wide effort to evaluate computational function prediction methods [95]. Following an open call, 23 participating teams submitted 54 algorithms for assessment that predicted gene ontology (GO) functional terms [18] for a common set of about 50,000 proteins that then lacked experimentally corroborated annotations. In the 11 months following the submission deadline, GO curators—continuing to work independently from the CAFA organisers—examined relevant literature and assigned functional annotations to 866 of those proteins. These 866 proteins became the gold standard reference set for evaluating the performance of all CAFA submissions.

CAFA is undoubtedly an important effort, but the concern remains that the primary CAFA evaluation metric fails to account for the Open World Assumption underlying GO annotations [123]: the function annotations of most proteins are incomplete and, consequently, absence of evidence of function does not amount to evidence of absence of function. This omission leads to a systematic overestimation of false-positive prediction rates, which may significantly affect the results and conclusions reported in the CAFA study.

I agree that the gold standard reference set can confirm predictions (i.e., count true positives) but, because it does not exhaustively represent all functions of the target sequences, the reference set cannot falsify predictions (i.e., count false positives). To illustrate this point, consider the first target in the reference dataset: CLC4E_MOUSE, which was assigned the molecular functions “receptor activity” and “protein binding” based on GO annotations accrued in the Swiss-Prot database [128] in the 11 months following the close of the competition (Supplementary Table 1 in [95]). Inter-
2. OWA vs. CWA

ProScan [54] also predicts a function of “carbohydrate binding”, which in CAFA would be considered a false positive because it does not appear in the gold standard set. However, this prediction is actually correct, based on experimental evidence of alpha-mannose binding [140] that is not yet recorded in Swiss-Prot GO annotations. This example is not atypical, because the Swiss-Prot database is maintained by expert curators who, owing to resource limitations, process entries according to defined priorities. Even with vastly more resources, the database would remain incomplete because most functional information has yet to be discovered through direct experiments in the first place.

![Proportion of “false-positives” that are spurious](image)

(i.e. predictions not confirmed in the reference set, but confirmed in a later release)

**Figure 2.1:** Proportion of false positives that are spurious (i.e., predictions not confirmed in the CAFA reference set but confirmed in a later release).

To quantify the extent of spurious false positives that results from disregarding the Open World assumption, I simulated the CAFA experiment by considering a different, older set of data for which we now have the benefit of hindsight: successive releases of the UniProt-GOA database [35] dating back to 2007. Analogous to predictions submitted to CAFA, I first retrieved all gene products with computational (predicted) annotation but no experimental annotation in the 2007-01-19 release. Analogous to the way the CAFA gold standard set was built, I then established which of these gene products accumulated new experimental annotations between the 19 January 2007 and 16 January 2008 releases. For these targets, and following the CAFA protocol, I counted as false positives all electronic annotations in the 19 January 2007 release that were not confirmed by an experimental annotation in the 16 January 2008 release. However, a considerable proportion of these purported false-positive predictions were in fact confirmed by experimental annotations in subsequent UniProt-GOA releases (22 January 2012, 11 January 2011, 7 February 2012, 7 January 2013), thereby contradicting the initial assessment (Figure 2.1): in this analysis, ∼14.7% of the predictions initially deemed as false positives were later confirmed to be correct. This is necessarily an underestimate of the error rate in the CAFA definition of false-positive predictions, and questions the ranking of methods reported by CAFA.

To compare prediction methods meaningfully, future CAFA sequels should consider the Open World assumption and tackle head-on the complications associated with it. For instance, explicit annotations of absence of function—identified by the keyword **NOT** in the qualifier field of GO
annotations—should be required to falsify predictions. Currently, however, only about 2,500 of the 530,000 (0.48%) experimentally confirmed molecular function and biological process annotations in UniProt-GOA are negative ones—in part because they have often been perceived as less useful than their positive counterparts. But, to improve and evaluate function prediction, negative annotations are invaluable and more of them are needed [132]. Another way of addressing the problem would be to limit the scope of function prediction to specific aspects of function that can be thoroughly assessed in experiments after the submission deadline (e.g., particular enzymatic activities) [56]. For these restricted functional aspects, the more straightforward “Closed World” assumption would apply, but the conclusions drawn might not hold in general. Neither solution constitutes a “quick fix.” Indeed, progress in assessing protein function prediction is likely to require a substantial coordinated effort and broad support from the community. In that sense, the CAFA group is already well positioned to help drive the field forward.
Chapter 3

Quality of computationally inferred Gene Ontology annotations

Gene Ontology (GO) annotations are a powerful way of capturing the functional information assigned to gene products [11, 38]. The organization of the GO in a Directed Acyclic Graph (DAG) allows for various levels of assignment specificity, while the three ontologies—Biological Process, Molecular Function, and Cellular Component—capture three aspects of the gene product’s annotation.

Some GO annotations are assigned by expert curators, either from experimental evidence in the primary literature (experimental annotations) or from other evidence such as sequence similarity, review papers and database entries (curated annotations). However, the vast majority (> 98%) of available GO annotations are assigned using computational methods, without curator oversight [38].

Uncurated—electronic—annotations are generally considered to be least reliable. Many users of GO annotations err on the safe side by assigning a lower rank/weight to electronic annotations or leave them completely out of their analyses [e.g., 37, 60, 30, 22, 14]. However, there have been very few evaluations of the quality of electronic annotations. To our knowledge, the most relevant study to date assessed the annotation quality of only 286 human proteins [25].

In this chapter, I provide the first comprehensive evaluation of electronic GO annotation quality. Based on successive releases of the UniProt Gene Ontology Annotation database (UniProt-GOA), the largest contributor of electronic annotations [13], I used experimental annotations added in newer releases to confirm or reject electronic annotations from older releases. I defined 3 measures of annotation quality for a GO term: 1) reliability measures the proportion of electronic annotations later confirmed by new experimental annotations, 2) coverage measures the power of electronic annotations to predict experimental annotations, and 3) specificity measures how informative the predicted GO terms are.
3. Quality of electronic annotations

Figure 3.1: A list of the Gene Ontology (GO) evidence and reference codes I analyzed. I group the GO evidence codes in three groups: experimental, non-experimental curated, and electronic. Gray text denotes the evidence codes that I did not include in the analysis: they are either used to indicate curation status/progress (ND), are obsolete (NR), or there is not enough data to make a reliable estimate of their quality (ISO, ISA, ISM, IGC, IBA, IBD, IKR, IRD). The subdivision of the evidence codes (green rectangles) reflects the subdivision available in the GO documentation: http://www.geneontology.org/GO.evidence.shtml.

In this chapter, I describe my new methodology and the results in detail. I first consider changes in quality in UniProt-GOA over time. I then characterize the relationship between GO term reliability and specificity. Next, I consider the possible differences in quality among the three ontologies, among computational methods used to infer the electronic annotations, and among the 12 best-annotated model organisms. Finally, I contrast electronic annotations with curated annotations that use evidence other than experiments from primary literature.

3.1 Results

To evaluate the quality of electronic annotations, I tracked changes in UniProt Gene Ontology Annotation (UniProt-GOA) database releases in overlapping three-year intervals. As a surrogate for the intuitive notion of correctness, I define reliability as the ratio of confirmed electronic annotations to confirmed and rejected/removed ones. One electronic annotation is deemed confirmed or rejected, depending on whether a new, corresponding experimental annotation supports or contradicts it. Furthermore, if an electronic annotation is removed, the annotation is deemed implicitly rejected...
3.1. Results

and thus contributes negatively to the reliability measure (Figure 3.2, panel A). As a surrogate for the intuitive notion of sensitivity, I define coverage as the proportion of newly added experimental annotations that had been correctly predicted by an electronic annotation in a previous release (Figure 3.2, panel B).

![Diagram of strategy to evaluate electronic Gene Ontology annotations]

**Figure 3.2: Outline of the strategy to evaluate electronic Gene Ontology annotations.** (A) Reliability measures the proportion of electronic annotations confirmed by future experimental annotations: an electronic annotation in an older database release is either 1) confirmed by a new experimental annotation in the later release, 2) falsified by a new, contradictory experimental annotation (corresponding GO term, but with NOT qualifier, which amounts to an explicit rejection), 3) removed from the new UniProt-GOA release (implicit rejection), or 4) unchanged, which is uninformative and does not affect the reliability measure. (B) Coverage measures the extent to which electronic annotations can predict future experimental annotations: an experimental annotation in the newer release is either 1) correctly predicted by an electronic annotation in the older release, or 2) not correctly predicted (missed). Note that the strategy is outlined for electronic annotations, but any subset of annotations can be analyzed this way, e.g. annotations assigned using a selection of evidence or reference codes.

The addition of new experimental annotations—high-quality annotations assigned by a curator—allows us to evaluate the existing electronic annotations. Unfortunately, the set of available experimental annotations is small, since obtaining them requires valuable curator time. Moreover, resource constraints require that curators focus their efforts on a selected set of model organisms [98]. Consequently, most of the available experimental annotations are distributed among the model organisms (Figure A.1); it is this set of genomes that I analyse.

### 3.1.1 Electronic annotations in subsequent UniProt-GOA releases are increasing in quality

I first sought to evaluate general trends in the overall quality of UniProt-GOA. Four summary statistics—first and third quartile, median, and mean—allow us to describe the change in quality—specificity, reliability, and coverage—of successive UniProt-GOA releases (Figure 3.3). Subsequent UniProt-GOA releases are improving with the addition of slightly
3. Quality of electronic annotations

3.1.2 GO term’s specificity is only partially indicative of the reliability of electronic annotations

Next, I investigated the association between a GO term’s specificity and reliability (Figure 3.4). Previous works based on smaller datasets have observed a negative relation between the predictive power of computational annotation and the specificity of the assigned GO term [e.g., 15, 59, 62]. My results are consistent with these results to the extent that almost all general terms are stable (Figure 3.4). Specific terms, however, span the whole range of reliability. I also observe that on average, reliability of electronic annotations hardly depends on their specificity: the variance of reliability increases with an increase in specificity, but the median stays largely constant.
3.1. Results

Figure 3.4: Reliability of electronic annotations in the 16-01-2008 UniProt-GOA release compared to the specificity of the assigned GO term—Information Content in the 16-01-2008 UniProt-GOA release. Each point represents one GO term, and its color corresponds to the ontology in the legend. Each boxplot summarizes the reliability of a selection of GO terms: those with specificity in the range denoted by the width of the boxplot. Lower, mid, and upper horizontal lines denote the first quartile, median and the third quartile, respectively. Vertical lines reach the 1.5 interquartile ranges from the respective quartiles or reach the extreme value, whichever is closer. To be visualized in these plots, a GO term needs to have assigned at least 10 electronic annotations in the 16-01-2008 UniProt-GOA release and at least 10 experimental annotations in the 11-01-2011 UniProt-GOA release.

3.1.3 The three ontologies have similar reliability, but different coverage

To assess the differences in annotation quality among the three ontologies, I analyzed the ontologies separately in terms of reliability, coverage, and specificity. On average, annotations associated with the three ontologies were similarly stable, but vary considerably in coverage (Figure 3.5). Specifically, Biological Process (BP) terms had the lowest coverage, Molecular Function (MF) terms had the highest coverage, and Cellular Component (CC) terms were in-between. This is consistent with the notion that MF terms are easiest to assign, and BP terms hardest to assign [46]. Nevertheless, this difference in difficulty translates into variable coverage but very similar reliability, suggesting that the false-positive rate of electronic annotations is controlled effectively.
3. Quality of electronic annotations

Figure 3.5: The quality of the 16-01-2008 UniProt-GOA release, evaluated by the 11-01-2011 UniProt-GOA release. A scatterplot of coverage compared to the reliability for the GO terms of the three ontologies: Biological Process, Cellular Component, and Molecular Function. The area of the disc reflects the frequency of the GO term in the 16-01-2008 UniProt-GOA release. The colored lines correspond to the mean values for the respective axes. To be visualized in this plot, a GO term needs to have assigned at least 10 electronic annotations in the 16-01-2008 UniProt-GOA release and at least 10 experimental annotations in the 11-01-2011 UniProt-GOA release. An interactive plot to explore the figure is available\footnote{...}. 

Frequency of the GO term in UniProt-GOA
3.1. Results

Figure 3.6: The quality of the 16-01-2008 UniProt-GOA release, evaluated by the 11-01-2011 UniProt-GOA release. Each reference code is evaluated separately: (A) Inferred from Enzyme Commission, (B) Inferred from UniProt Subcellular Location terms, (C) Inferred from UniProtKB keywords, (D) Inferred from Ensembl Compara, (E) Inferred from HAMAP2GO, and (F) Inferred from InterPro. The 12 model organisms included in the analysis are *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Gallus gallus*, *Danio rerio*, *Dictyostelium discoideum*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, and *Escherichia coli* K-12. The ontology is denoted by the color of the disc, while the area of the disc reflects the frequency of the GO term in the 16-01-2008 UniProt-GOA release. The coloured lines correspond to the mean values for the respective axes. To be visualized in this plot, a GO term needs to have assigned at least 10 electronic annotations in the 16-01-2008 UniProt-GOA release and at least 10 experimental annotations in the 11-01-2011 UniProt-GOA release.

3.1.4 Different sources provide annotations of different quality

To investigate differences in quality among the various sources of electronic annotations in UniProt-GOA, I repeated our analysis for each of them. The six sources can be classified in two main categories: mapping of keywords from other databases (UniProtKB keywords, UniProt Subcellular Location terms, InterPro, and Enzyme Commission) and the use of comparative genomics in functional annotation (Ensembl Compara for eukaryotes and HAMAP2GO for microbial genomes) (Figure 3.6).

\[\text{http://people.inf.ethz.ch/skuncan/SupplementaryVisualization1.htm}\]
Two sources of electronic annotations are restricted to single ontologies: the Enzyme Commission (EC) numbers map to MF GO terms, and subcellular location terms of the UniProt database map to CC GO terms (Figure 3.6, panels A and B). Both annotation sources are applied to a comparatively small number of terms, but their reliability is remarkably high: on this restricted set of GO terms, they outperform other sources of electronic annotation (Figures 3.6, A.4, A.3).

The bulk of electronic annotations are inferred from the UniProt and InterPro databases (Figure A.5). With UniProtKB keywords, GO annotations are inferred using a correspondence table between Swiss-Prot keywords associated with UniProt entries and GO terms. Note that UniProt entries consist of a small minority of manually annotated entries (Swiss-Prot entries) and a large body of entries (TrEMBL entries) automatically annotated by a rule-based system (UniRules). With InterPro, GO annotations are inferred from a correspondence table between InterPro sequence and structure signatures and GO terms. Despite similarities in the two approaches, UniProt-based annotations show considerably higher average reliability than their InterPro-based counterparts (Figure 3.6, panels C and F, horizontal lines). In terms of average coverage, the two approaches show similar performance (Figure 3.6, panels C and F, vertical lines).

Substantial manual curation is involved in obtaining electronic annotations from the two sources that rely on comparative genomics: Ensembl Compara electronic annotations transfer experimental annotations among inferred one-to-one orthologs in a subset of model organisms, and HAMAP2GO electronic annotations rely on manually created rules to propagate experimental annotations within a family of microbial proteins. Despite the intricacies involved in the annotation pipeline, these two sources have the lowest mean coverage and reliability among the six analyzed sources (Figure 3.6, panels D and E). However, note that the HAMAP rules have taxonomic restrictions on propagation that are not included in the HAMAP2GO pipeline. Hence, some aspects of HAMAP are not captured in UniProt-GOA, and therefore are not analyzed here.

This overall low reliability—a consequence of many rejected annotations—indicates that GOA strategies based on comparative genomics are currently less reliable than approaches based on sequence features (UniProtKB keywords and InterPro).

### 3.1.5 Quality of electronic annotations and the number of assigned GO terms are different among the model organisms

To investigate the difference in electronic annotation quality among the model organisms, I repeated our analysis for each model organism separately. Overall, repeating the analysis confirmed our general findings above. However, I observed variations among organisms, both in the number of available annotations and their quality (Figures 3.7, A.8, A.2, A.7).

Organisms with the largest number of changes—confirmations and rejections—tend to have the highest quality of annotation: the three unicellular organisms and the three mammals (Figures 3.7, top and bottom.
3.1. Results

Figure 3.7: Quality of the 16-01-2008 UniProt-GOA release, evaluated by the 11-01-2011 UniProt-GOA release; each model organism is evaluated separately. Common background shading denotes a depiction of the same set of GO terms (full data is presented in Figure A.6). The ontology is denoted by the color of the disc, while the area of the disc reflects the frequency of the GO term in the 16-01-2008 UniProt-GOA release. To be visualized in this plot, a GO term needs to have assigned at least 10 electronic annotations in the 16-01-2008 UniProt-GOA release and at least 10 experimental annotations in the 11-01-2011 UniProt-GOA release for each model organism. The coloured lines correspond to the mean values for the respective axes.

Experimenting, describing and interpreting results on unicellular organisms is arguably more straightforward than on multicellular organisms; it might explain the relatively high quality of electronic annotations.
3. Quality of electronic annotations

for the three unicellular model organisms (Figure 3.7, bottom row). The average quality measures for the three mammals—*Homo sapiens*, *Mus musculus*, and *Rattus norvegicus*—are comparably high (Figure 3.7, top row), but many specific low-quality annotations somewhat reduce the means of reliability and coverage.

Our observation that general GO terms tend to have higher reliability holds for each model organism. Nevertheless, assigning mainly general GO terms guarantees neither high reliability nor high coverage. I observe the worst electronic annotation quality on *Gallus gallus*, *Danio rerio* and *Dictostelium discoideum* gene products, despite a mean specificity of 1.79, versus 4.47 for mammals.

3.1.6 The reliability of electronic annotations rivals that of non-experimental curated annotations

![Figure 3.8: Quality of electronic and curated annotations on a common set of GO terms. Quality of the 16-01-2008 UniProt-GOA release is evaluated by the 11-01-2011 UniProt-GOA release; coverage is on the x-axis and reliability is on the y-axis. The ontology is denoted by the color of the disc, while the area of the disc reflects the frequency of the GO term in the 16-01-2008 UniProt-GOA release. The coloured lines correspond to the mean values for the respective axes. To be visualized in the plot, a GO term needs to have assigned at least 10 electronic/curated annotations in the 16-01-2008 UniProt-GOA release, and at least 10 experimental annotations in the 11-01-2011 UniProt-GOA release.](image)

To put the quality of electronic annotations in perspective, I contrasted them to curated annotations (evidence codes RCA, ISS, TAS, NAS, and IC), i.e., annotations inferred by curators without direct experimental evidence (Figure 3.8). Curated annotations contain annotations assigned using evidence codes perceived as of particularly high quality: for instance, del Pozo et al. [30] consider the TAS evidence code to “offer the highest confidence [along with the IDA evidence code]”. Buza et al. [22] rank TAS and
IC evidence code second only to the group of annotation codes that rely on direct experimental evidence. In Benabderrahmane et al. [14], TAS is the only evidence code to receive the weight of 1.0.

Compared to electronic annotations, it is not surprising that curated annotations have a considerably lower average coverage (Figure 3.8, vertical lines). Indeed, the main appeal of electronic annotations is precisely that they scale efficiently to large quantities of data. But in terms of reliability, and contrary to current beliefs, curated annotations that use evidence other than experiments from primary literature do not fare better than electronic annotations (Figure 3.8, horizontal lines, Figure A.9). In fact, I observed a higher reliability for electronic annotations than for curated annotations (0.52 vs. 0.33).

A more detailed analysis revealed that the lower mean reliability of curated annotations in the 16-01-2008 UniProt-GOA release is mainly due to removed annotations with evidence code Reviewed Computational Analysis (RCA) (Fig. S10 in Text S1). The low reliability of RCA annotations is caused by the removal of many RCA annotations assigned to the M. musculus gene products (Figure A.8, yellow bar in the panel denoted M. musculus); these were removed as there were concerns about the veracity of results from some papers that had been annotated (Emily Dimmer, personal correspondence).

When I exclude annotations assigned using the RCA evidence code, the reliability of non-experimental curated annotations rises to 0.58. But even then, the reliability of electronic annotations (0.52) remains competitive with that of curated annotations (Figure A.10).

### 3.2 Discussion

Electronic annotations constitute the bulk of GO annotations, yet their correctness has not been systematically assessed until now. Direct, experimental verification by means of new experiments would be prohibitively expensive even for a small subset of the annotations. Instead, I sought to exploit existing, but newly available experimental data to evaluate electronic annotations. Specifically, I defined and used a measure I call reliability as an indicator of correctness: a GO term has high reliability if, in a subsequent release, many associated electronic annotations are confirmed experimentally while few associated annotations are removed or explicitly negated. This approach at verifying electronic annotations is both efficient (as it reuses existing experiments) and powerful (as it potentially applies to any term). At the same time, the measure is only as accurate and representative as the newly recorded experimental annotations. For instance, there are far more “positive” function annotations than “negative” ones (annotations with a NOT qualifier, which indicates lack of function), which could result in inflated reliability estimates. On the other hand, I attempt to compensate for this bias by considering all removed electronic annotations as negative ones. While it might be argued that the removal of an electronic annotation does not necessarily imply that it is wrong, from a user standpoint, the removal of an annotation hardly suggests that it can be relied upon.
3. Quality of electronic annotations

Despite analyzing 193 027 gene products, my approach leaves out a number of uninformative electronic annotations, which are neither confirmed nor rejected in a given time interval. Due to the incomplete nature of GO (referred to as the Open World Assumption), absence of an annotation does not imply absence of the corresponding function. This is reflected by the fact that most gene products in GOA have been updated at least once—with the period between updates lasting as long as 12 years (Figure A.11).

Electronic annotations have often been perceived as unreliable, but my study provides a more differentiated picture. First, I observed that the reliability and, to a lesser extent, the specificity of electronic GO annotation has steadily improved in recent years. This is a remarkable achievement, given that the number of electronic annotations has been growing exponentially during the same time period [38].

Second, despite these overall encouraging results, there are significant variations in performance among the different types of electronic annotations. The two most reliable sources also happen to be the most specialized ones: annotations derived from UniProt Sub cellular Location terms and EC numbers. This suggests that specialization can be advantageous.

Also highly reliable are annotations obtained from mapping Swiss-Prot keywords associated with UniProtKB entries to GO terms. In particular, the high mean reliability of predictions of Biological Process GO terms stands out, on what is arguably the most difficult ontology to assign [46]. There are nevertheless a handful of general UniProtKB keywords derived GO terms that have low reliability (Figure 3.6); in particular, Molecular Function terms related to metal ion binding have proven to be unreliable throughout all three analyzed UniProt-GOA releases due to a number of removed annotations (GO terms denoted in Figure 3.6 C, Dataset S1; an interactive plot is available at http://people.inf.ethz.ch/skuncan/SupplementaryVisualization2.html). In addition, a few annotations related to ion transport were explicitly rejected with the NOT qualifier, e.g., UniProtID Q6R3K9 now has a NOT annotation for “iron ion transport”, UniProtID Q3YL57 now has a NOT annotation for “sodium ion transport”, and UniProtID Q9UN42 now has a NOT annotation for “monovalent inorganic cation transport”.

Since the UniProt database includes manually annotated entries (“Swiss-Prot entries”) in addition to electronically annotated (“TrEMBL entries”), this could introduce some circularity in our analysis. However, the proportion of manually annotated entries in UniProt is very small (3.06% in the September 2011 UniProt release), so any bias so incurred cannot affect our conclusions. The importance of the automated component of the UniProt pipeline is also reflected in the large number of electronic annotations derived from it—almost a quarter of all electronic annotations (Figure A.5).

Besides UniProtKB keywords, InterPro sequence and structure signatures constitute the other large source of electronic annotations (42%; Figure A.5). Their average reliability is however not as good as UniProtKB keywords-derived terms. Consider for instance the Cellular Component term “integral to membrane” and its parent term “intrinsic to membrane” (Figure 3.6, panel F). The reliability of annotations associated with these
3.2. Discussion

These observations are consistent with a recent article reporting “promiscuous hits limited to solely [signal peptide or transmembrane helix] part among clearly unrelated proteins” [137]. Moreover, I observed more InterPro annotations rejected with the NOT qualifier than UniProtKB-based annotations (Dataset S1). For example, UniProtIDs Q8IZE3, Q96R7, and Q8BKG3 now have a NOT annotation for “kinase activity”; UniProtID Q2L385 now has a NOT annotation for “channel activity”; UniProtIDs Q9LQ10, Q8GYY0, and Q06429 now have a NOT annotation for “1-aminocyclopropane-1-carboxylate synthase activity.”

As for strategies based on comparative genomics, namely HAMAP2GO and Ensembl Compara, they yielded the least reliable annotations of those I analysed. But because they have been introduced in the UniProt-GOA releases relatively recently, I could only assess their performance on one or two overlapping time intervals (Figure A.12). If transient, the low reliability of an annotation source could be the result of a large change in the annotation pipeline that ultimately results in more reliable resource. For instance, when looking for the cause of low reliability for the annotations Inferred from HAMAP2GO (3.6, panel E), I found the HAMAP2GO file—mapping HAMAP annotations to GO terms—is currently being substantially revised (Alan Bridge and Emily Dimmer, personal correspondence). A recent change in policy towards more conservative predictions resulted in the large number of removed annotations I observed. Because of the lagging nature of our quality measures, I will only be able to assess the new pipeline in a few releases’ time.

Despite these considerable variations among sources of annotations, all electronic annotations are currently labelled with the same evidence code (“IEA”)—with the source information relegated to the more obscure which/from attribute. As many users and tools tend to ignore the latter database column, I recommend making these differences more explicit by introducing multiple evidence codes for electronic annotations; the new evidence codes might take into account the subdivisions available in the ECO ontology [135].

The third and arguably most unexpected finding of this study is that the reliability of electronic annotations rivals that of annotations assigned by an expert curator using sources other than direct experimental evidence (Figure 3.8, horizontal lines). At the same time, the coverage of electronic annotations—which measures the ability to predict future experimental annotations—is far superior (Figure 3.8, vertical lines). For example, the mean reliability of the BP ontology is slightly lower when inferred from electronic annotations than when the annotations are based on sequence similarity and approved by the curator (evidence code ISS). Still, the mean reliabilities for the CC and MF ontologies are slightly higher for electronic annotations, and the mean coverage of electronic annotations for all three ontologies is visibly higher (Figure A.13).

This challenges the widespread notion that annotations inferred by algorithms are less reliable than annotations inferred by curators using evidence codes.
other than direct experimental evidence found in primary literature—a notion that might have had validity when automated annotations consisted of relatively crude approaches, such as global sequence similarity with ready-made thresholds. Although occasionally still in use, such annotation strategies have been largely superseded by the approaches highlighted here and described elsewhere in more detail [13, 50, 24].

3.3 Conclusion

To narrow the gap between the number of sequenced gene products and those with functional annotation, computational methods are indispensable [18], [19], even more so for the non-model organisms (Fig. S4 in Text S1). I introduced three measures to evaluate the quality of electronic annotations: one accounts for the specificity of the assigned GO term, and two reliability and coverage assess the performance of electronic annotation sources by tracking changes in subsequent releases of annotation files.

Although the performance of electronic annotations varies among inference methods (sources), the overall quality of electronic annotations rivals the quality of curated non-experimental annotations.

This is not to say that the curators have made themselves redundant. On the contrary, as I highlight above, most electronic annotations heavily rely on manually curated UniProtKB keywords and InterPro entries. Moreover, given the essential role of curators in embedding experimental results into ontologies, so does the present study.

3.4 Materials and Methods

3.4.1 Data

I used the January 2011 release of the OBO-XML file to obtain the GO terms, definitions and the ontology structure needed in the analysis. The file was downloaded from the GO FTP site.

The annotations (mappings of gene products to GO terms) were downloaded from the European Institute for Bioinformatics (EBI) FTP site. Each file, created as part of the UniProt Gene Ontology Annotation (UniProt-GOA) project, is a many-to-many mapping of UniProtKB IDs to GO terms. All dates mentioned in this study refer to the release date of these annotation files, not the date attribute of individual annotations.

I analyzed 193,027 UniProtKB IDs; GO terms can be assigned to these sequences using any of the evidence or reference codes. The distribution of annotations among the 12 Gene Ontology Reference genomes is shown in Figure A.2. This set of model organisms has by far the largest number of high-quality experimental annotations, allowing us to make the most reliable estimate of the annotation quality (Figure A.1).

4 ftp://ftp.ebi.ac.uk/pub/databases/GO/goa/UNIPROT/
3.4. Materials and Methods

The structure of the GO vocabulary is changing as a response to consistency checks, new biological insights, and intricacies involved in annotating various model organisms [82, 141, 67]. To account for these changes, for each pair of GO releases analysed I only consider terms that are present in both releases.

3.4.2 Gene Ontology meta-information

The source of an annotation is recorded in the evidence code. I group GO evidence codes into 3 broad categories: 1) codes reflecting annotations assigned by curators using direct experimental evidence from the literature (experimental evidence codes EXP, IMP, IGI, IPI, IEP, IDA), 2) codes reflecting annotations inferred by curators using other types of evidence (curated evidence codes ISS, RCA, IC, NAS, TAS) and 3) electronic evidence code (IEA), denoting annotations which are inferred computationally (Figure 3.1). Several evidence codes were not included in the analysis: they are either used to indicate curation status/progress (ND), are obsolete (NR), or there is not enough data to make a reliable estimate of their quality (ISO, ISA, ISM, IGC, IBA, IBID, IKR, IRD).

A reference code captures the source of an electronic annotation. I analyse six reference codes available in UniProt-GOA: three are based on cross-referencing keywords from other databases: UniProtKB keywords, UniProt Subcellular Location terms, and Enzyme Commission [12, 127]; two are based on the propagation of annotations within a family of proteins: InterPro and HAMAP2GO [53, 70]; one reference code uses comparative genomics in projecting experimental annotations to unannotated inferred one-to-one orthologs—Ensembl Compara [131].

When a NOT qualifier accompanies an annotation, it explicitly states that the gene product is not associated with the respective GO term. A subtle use of the NOT qualifier comes into play because the isoform distinctions are not reflected in the annotation files at this time; a gene product can be mapped to the GO term in a given spatial/temporal context, but the mapping is not valid in another context (Judith Blake and Pascale Gaudet, personal correspondence). Such gene products will be mapped to one GO term twice—one accompanied by a NOT qualifier and one without it. For consistency, I ignore all such occurrences. The 11-01-2011 UniProt-GOA release contains 493 gene products with such annotations.

3.4.3 Qualitative evaluation of Gene Ontology annotations using successive releases of the UniProt-GOA file

All analyses are performed on overlapping 3-year periods between 2006 and 2011. Unless stated otherwise, I show the results associated with the most recent period analysed in this work (2008-2011).

The three measures of quality I introduced are specificity, reliability, and coverage. For clarity, the definitions are given and described for electronic annotations. Nevertheless, any subset of annotations can be analysed this

\[^5\text{http://www.geneontology.org/GO.evidence.shtml}\]
3. Quality of electronic annotations

way, e.g., annotations assigned using one or a subset of evidence or reference codes.

I measure the specificity (opposite of generality) of a GO term GO_i with respect to its information content [98, 78, 5]:

\[ \text{Specificity}(GO_i) = -\log_2(\text{freq}(GO_i)) \]

where \( \text{freq}(GO_i) \) is the frequency of GO_i among all annotations considered.

To calculate the reliability for a GO term, I count all the confirmed and rejected electronic annotations associated with this term (Figure 3.2, panel A). An electronic annotation is confirmed if it is corroborated by a new experimental annotation, i.e., an experimental annotation added during the time interval. An electronic annotation is rejected if it is falsified by a new experimental annotation that comes with a NOT qualifier, or if this electronic annotation has been removed in the later UniProt-GOA release. More formally,

\[ \text{Reliability}(GO_i) = \frac{C_{GO_i}}{C_{GO_i} + R_{GO_i}} \]

where \( C_{GO_i} \) is the set of confirmed annotations associated with term GO_i and \( R_{GO_i} \) is the set of rejected and removed annotations associated with term GO_i.

To calculate the coverage for a GO term in a UniProt-GOA release, I count all the new experimental annotations in the later UniProt-GOA release correctly predicted by an electronic annotation in the earlier release, and those not correctly predicted (missed) by electronic annotations in the earlier release (Figure 3.2, panel B). More formally,

\[ \text{Coverage}(GO_i) = \frac{P_{GO_i}}{P_{GO_i} + M_{GO_i}} \]

where \( P_{GO_i} \) is the set of correctly predicted new experimental annotations associated with term GO_i and \( M_{GO_i} \) is the set of ”missed” new experimental annotations associated with term GO_i.

To calculate any of the measures of quality, I take into account the GO Direct Acyclic Graph (DAG) structure. To calculate the frequency of a GO term, I account for all annotations derived by inheritance. Consequently, the specificity of any child term is necessarily greater than or equal to the specificity of its parents. When calculating reliability, an annotation that is replaced by a more specific annotation (a descendent) is not considered rejected, as the descendent still implies it. Similarly, an annotation is confirmed by the arrival of an experimentally ascertained descendent, as the more specific term implies the more general term. Conversely, if an annotation is followed by the arrival of a less specific experimental annotation, only the subset of its ancestral terms implied by the less specific experimental annotation is deemed as confirmed; the rest is uninformative (neither confirmed, rejected, nor removed).

All the results of the described analysis are available as Dataset S2 of the accompanying publication [132].

3.4.4 Visualization

The analysis was done using a combination of in-house Java classes, SQL queries to the custom database, and R scripts. Summaries were done using the plyr package of the R language [136]; all plots were created using the gg-
plot2 package of the R language [135], and the interactive plots were created using the googleVis package of the R language; the respective R packages are available from the CRAN repository. REVIGO web server [120] was used to summarize the lists of GO terms and select those highlighted in the Results section.
Chapter 4

Phylogenetic Profiling with Cliques of Orthologs Is Enhanced by Signatures of Paralogy Relationships

This chapter is derived from the peer-reviewed publication: Škunca N, Bošnjak M, Kriško A, Panov P, Džeroski S, Šmuc T, Supek F. Phyletic profiling with cliques of orthologs is enhanced by signatures of paralogy relationships. PLoS Comput Biol. 2013

While both the number and the diversity of sequenced prokaryotic genomes grow rapidly, the number of specific assignments of gene functions in the databases remains low and skewed toward the model prokaryote *Escherichia coli*. In closing the gap between the number of sequenced genes and the number of functionally annotated genes, computational methods for functional annotation are invaluable.

Many computational methods for functional annotation of genes are based on a search for sequences with common evolutionary descent—homologs. One possible encoding of homology is the use of phylogenetic profiles: each row in the phylogenetic profile represents one gene, and the columns represent the presence or absence of homologs in sequenced genomes [92, 63].

There are two main ways in which phylogenetic profiles can be used for annotation of gene function. Both of them involve propagating the annotation label. First, one could create phylogenetic profiles and propagate the annotation label within the profile—from genes with known function to their homologs included in the profile. This is homology-based annotation, and many schemes for doing so are possible [74]. Second, one could propagate labels between the profiles by finding similar profiles: assuming that genes that are inherited together tend to work together, one transfers annotation from a better-studied group of homologs to a profile that is similar but contains genes that are not as well studied. Again, this can be done in many ways. For example, phylogenetic profiles can be grouped by similarity using a variety of distance measures [e.g., 92, 97], possibly involving a machine learning framework [e.g., 97, 130, 94]. Rows in the phylogenetic profile can stand for genes or groups of genes [e.g., 92, 107, 138]; functional
annotation can be assigned using a range of vocabularies, e.g., UniProt controlled vocabulary of keywords [128], Enzyme Commission numbers [125], or arguably the most widespread vocabulary, the Gene Ontology [48]. In addition, one could employ some hybrid between the first two approaches, e.g., when the evidence in favour of within-profile label propagation is used to improve the confidence of between-profile propagation and vice versa.

Refinements of homology-based annotation include making a distinction between two types of homologous relationships: orthologs—sequences derived from the same gene in the last common ancestor, and paralogs—sequences derived from a duplication event [43]. Because orthologous pairs are expected to keep the same function [27, 122, 64] and paralogous pairs are expected to diverge in function [57], the canonical approach to functional annotation relies on transfer of function between orthologs. However, the latest evidence suggests that, relative to pairs of paralogs, the conservation of function between pairs of orthologs is not as strong as the standard model would imply [4].

In this chapter, my goal was to create a functional annotation model that learns to associate gene function with specific patterns in phylogenetic profiles—the presence and absence of different types of homologs in different organisms. To create the phylogenetic profiles, I combined ortholog cliques—fully interconnected groups of orthologs—with both additional orthologs and additional paralogs. I found that, instead of reducing the predictive accuracy, paralogs provide valuable information: compared to the model that includes only orthologs, the model that includes both orthologs and paralogs gave more predictions at the same average correctness.

In addition, experimental assays in the model organism *Escherichia coli* showed that the annotation model provides realistic assessments of confidence for the predicted annotations: a growth phenotype screen on *E. coli* knockout mutants indicated an overall Precision of 66%—out of 38 tested genes, the assay confirmed predictions for 25 genes—agreeing with the expected Precision of 60%.

With the work detailed in this chapter, I predict Gene Ontology annotations at various levels of specificity for about 1.3 million poorly annotated genes in 998 prokaryotes at a stringent threshold of 90% Precision: about 19 000 of those are highly specific functions. In addition to these, the function annotation model I introduce in this chapter provides many more predictions at less stringent cut-offs in a Web resource GORBI (http://gorbi.irb.hr/).

### 4.1 Results

I created the functional annotation models in three steps: 1) constructing the phylogenetic profiles, 2) functionally annotating them where possible, and 3) using a decision tree-based classifier to find groups of profiles that are similar or dissimilar (Figure 4.1). I detail these steps below.

The first step is constructing the phylogenetic profiles; in fact, this step is what differentiates between models proposed in this work. To choose among these models for functional annotation, I constructed four kinds of phylogenetic profiles (Figure 4.1). First, phylogenetic profiles of OMA cliques of or-
4.1. Results

Figure 4.1: Constructing phylogenetic profiles with the relations inferred by the OMA algorithm. A) One OMA group and the possible relations used in constructing phylogenetic profiles: members of an OMA group are all connected by orthologous relations and they form a clique (red); some orthologous proteins were left out in the process of forming cliques because they lack an orthologous connection to at least one group member (blue); a witness to non-orthology infers paralogs (green) [35] B) Constructing phylogenetic profiles: presence of the corresponding homolog is shown with the colours and their combinations. For example, when constructing the phylogenetic profile that accounts for OMA clique members (red) and all left out orthologs (blue), the cell in the 1st column and 1st row will have 1: Species 1 has an OMA 1 clique member (red) and at least one more protein in an orthologous relationship with at least one protein from OMA 1 (blue); the cell in the 998th column and 2nd row will have 0: Species 998 only has protein(s) in a paralogous relationship to OMA 2 members. In the Function column, the Gene Ontology annotations are assigned when at least half of the OMA clique members have the respective annotation.
4. Phylogenetic profiling with orthologs and paralogs

orthologs. Third, I added presence patterns for all paralogs inferred by the OMA algorithm; these are in fact inferred between-species paralogs—broken pairs in the OMA algorithm. The within-species paralogs are accounted for implicitly: if an OMA clique member is connected to a within-species paralog, the binary phylogenetic profile does not change. Fourth, I made a separate set of phylogenetic profiles that only include clique members and paralogs, but not the orthologs outside of the clique.

The second step is annotating the phylogenetic profiles with a GO term if at least half of clique members had the respective GO term assigned to them. I determined this threshold empirically (see Materials and Methods) in order to maximize functional consistency of known annotations within OMA cliques. The additional orthologs and paralogs were not considered in GO term annotation, even when their presence/absence was used in creating the profile. In other words, the difference between the functional annotation models is in the pattern of presence/absence of different types of homologs, and not in the functional annotations assigned to the phylogenetic profile.

The final step is measuring the (dis)similarity between profiles. I presented both the annotated and the poorly annotated phylogenetic profiles to a machine learning algorithm based on decision trees. In the decision tree algorithm, the groups of phylogenetic profiles are recursively divided into subsets based on their presence/absence patterns. In fact, the similarity measure is not defined a priori, but is instead inferred from the data: those homologs whose presence/absence best discriminates between GO terms are used to determine which profiles are more similar. In the final step of the decision tree algorithm, the most similar phylogenetic profiles are placed in leaves: this allows us to propagate the GO term annotation across profiles within these leaves.

Here, I used an algorithm based on decision trees, Clus-HMC-Ens [129]. Clus-HMC-Ens is based on combining multiple decision trees in a Random Forest-like setting [20], and can handle multiple labels—here, GO terms—for each phylogenetic profile. Furthermore, Clus-HMC-Ens is aware of the hierarchical relationships between the multiple labels and uses this information to improve predictive accuracy [129].

I report three performance measures: Precision, Recall, and Area Under the Precision-Recall Curve (AUPRC). Precision stands for the fraction of predictions that are known to be true, Recall stands for the fraction of known annotations that were successfully predicted, and AUPRC summarizes both Precision and Recall at all possible stringency thresholds of the annotation model. Formal definitions of these measures and the machine-learning algorithm’s train/test procedure used to obtain them are detailed in the Materials and methods section.

4.1.1 Both orthologs and paralogs contribute to predictive accuracy of phylogenetic profiles

In one OMA clique, inferred orthologous relations connect each protein to every other protein, so it is not surprising that they group proteins with mostly the same function (see Materials and methods). However, OMA cliques leave out many of the existing orthologous relations. Consequently,
4.1. Results

Figure 4.2: Predictive performance of the four analysed models for the three Gene Ontologies. A) Biological Process, B) Cellular Component, and C) Molecular Function. The x axis represents the models: phylogenetic profiles are based on (a) OMA cliques of orthologs; (b) OMA cliques of orthologs and OMA inferred orthologs; (c) OMA cliques of orthologs and OMA inferred paralogs; and (d) OMA cliques of orthologs, OMA inferred orthologs, and OMA inferred paralogs. The y-axis represents the Area Under the Precision-Recall Curve (AUPRC). Each disc represents one GO term; its colour represents the ontology, while the area of the disc is proportional to the generality of the GO term: the frequency of the GO term among all annotations available in 07-02-2012 UniProt-GOA release. Each boxplot summarizes AUPRC for the dataset indicated on the x-axis. Lower, mid, and upper horizontal lines denote the first quartile, median and the third quartile, respectively; vertical lines reach 1.5 interquartile range from the respective quartile or the extreme value, whichever is closer.

Phylogenetic profiles of OMA cliques are incomplete, leading to poor performance in our classification model: many of the true orthologous relations are missing, and the model can successfully annotate using only the most general GO terms (Figure 4.2, model a; Figure B.1, panels A, B, and C). If I compensate for the missing orthologous relations in OMA cliques by adding all inferred one-to-one, one-to-many, many-to-one, and many-to-many orthologs left out when constructing the cliques, the model improves: the mean AUPRC is 0.8 (Figure 4.2, model b; Figure B.1, panels D, E, and F). I also tested whether adding paralogs to phylogenetic profiles of OMA
cliques improves the mean AUPRC: it does, showing that the functional information I obtain from paralogs is far from useless (Figure 4.2, model c; Figure B.1, panels G, H, and I). Still, the mean AUPRC is 0.65—lower than if I enrich phylogenetic profiles with orthologous relationships. However, it is the combined information from orthologs and paralogs that provides us with the best model for functional annotation (Figure 4.2, model d; Figure B.1, panels J, K, and L): the mean AUPRC increases to 0.85.

In the above experiments, adding only orthologs improved AUPRC more than adding only paralogs (Figure 4.2, models b and c, respectively). To test whether accounting for the ortholog/paralog distinction would further increase AUPRC, I encoded the phylogenetic profiles with three levels: presence of an OMA clique member or another ortholog (2), presence of a paralog (1), or absence of any of these (0). I found a small gain in AUPRC resulting from the ortholog/paralog distinction (Figure B.2, panel B), but I also found that increasing the number of levels in the dataset from the original two to the above-described three decreases the AUPRC (Figure B.2, panel A). Taken together, accounting for the ortholog/paralog distinction did not yield an overall gain in AUPRC in the current machine learning setup (Figure B.2, panel C), so I chose the binary model as the principal result.

4.1.2 Consistent gains in accuracy across GO terms

In this binary model that includes both orthologs and paralogs, most of the general GO terms have high AUPRC. More specific GO terms span a wide range of AUPRC (Figure B.3). Nevertheless, both specific and general GO terms benefit from the inclusion of orthologs and paralogs. Specific GO terms such as “lysine biosynthetic process via diaminopimelate,” “organic acid\[ratio\]sodium symporter activity,” or “bacterial-type flagellum basal body” are used in less than 0.1% of annotations in the 07-02-2012 UniProt-GOA release (their Information Content is higher than 10): the mean AUPRC of this subset of specific GO terms rises from 0.78 in the model that includes orthologs (Figure 4.2, model b) to 0.83 in the model that includes both orthologs and paralogs (Figure 4.2, model d). For the general GO terms such as “protein transport,” “kinase activity,” or “plasma membrane,” each used in more than 3% of annotations in the 07-02-2012 UniProt-GOA release (their Information Content is lower than 5), the corresponding change in AUPRC is from 0.80 to 0.88.

Intuitively, phylogenetic profiling should perform best for the Biological Process (BP) GO terms: proteins with similar profiles are expected to be involved in the same BP but not necessarily to have the same Molecular Function (MF). For example, one kinase and one glucosidase may be involved in the same process of sporulation despite having different MF. As a result, one would expect phylogenetic profiling to be more appropriate for assigning BP GO terms than MF GO terms.

Here, I report high predictive accuracy for all three ontologies (Figure 4.2, model d). In fact, among the best performing and most specific predictions are those for Molecular Function (MF) GO terms “acyl-CoA dehydrogenase activity,” “transposase activity,” “organic acid\[ratio\]sodium symporter activity” and its parent term “solute\[ratio\]sodium symporter ac-
4.1. Results

tivity,” “penicillin binding” and its parent term “drug binding” (Figure B.3).

4.1.3 Model that includes paralogs provides more predictions with the same correctness

![Figure 4.3: The relationship between Precision and Recall for GO terms, at various model stringency cut-offs.](image)

Predictions for each GO term are evaluated at one of three cut-offs: (A), (B), and (C) show results at cut-offs 0.1, 0.3, and 0.7 respectively, for the model including OMA cliques of orthologs and OMA inferred orthologs; (D), (E), and (F) show results at model cut-offs 0.1, 0.3, and 0.7 respectively, for the model including OMA cliques of orthologs, OMA inferred orthologs, and OMA inferred paralogs. Each disc represents one GO term; the colour denotes the ontology, and the area of the disc reflects the frequency of the GO term in the 07-02-2012 UniProt-GOA release. The coloured lines correspond to the mean values for the respective axes, for the respective ontology. The model made at least 50 predictions for each visualized GO term.

The AUPRC provides us with a view on predictive accuracy that values both the comprehensiveness of predicted annotations for a given GO term (Recall) and their correctness (Precision) across the entire range of model stringency cut-offs. To further explore the relationship between Precision and Recall at specific levels of model stringency, I chose three cut-offs—0.1 (permissive cut-off), 0.3 (medium cut-off), and 0.7 (stringent cut-off), for the two best models—the model including orthologs (corresponding to AUPRC values in Figure 4.2, model b) and the model including both orthologs and paralogs (corresponding to AUPRC values in Figure 4.2, model d). The combination of data and cut-offs resulted in six plots (Figure 4.3).

For any of the cut-offs, the mean Precision for GO terms between the two models is similar (Figure 4.3, horizontal lines between A and D; B and E; C and F). However, there is a difference for Recall, in particular for the more stringent cut-offs (Figure 4.3, vertical lines between B and E; C and F). It is this increase in Recall that increases AUPRC, as I observed...
4. Phylogenetic profiling with orthologs and paralogs

before (Figure 4.2). For example, at the most stringent cut-off the model including only orthologs predicts annotations with 414 GO terms for at least 50 poorly characterized genes in the 998 genomes, while the model including both orthologs and paralogs predicts annotations with 573 GO terms for at least 50 genes.

To each unnannotated OMA clique, the model assigned a cut-off that indicates the probability of being annotated with a GO term. To have an interpretable measure of confidence for each prediction, I transformed this cut-off to the corresponding Precision (see the Materials and methods section). I then propagated the function of each OMA clique to the member genes and obtained the functional annotations, along with the estimates of Precision for each annotation.

As a consequence of the increased Recall, the model that includes both orthologs and paralogs provides more annotations at the same Precision (Figure 4.4, panel A). The increased Recall allows us to assign specific annotations at a very stringent threshold of 90% Precision. For example, I predict new annotations for *E. coli*, both using the most general, as well as many specific GO terms (Figure 4.4, panel B, Figures B.4 and B.5).

4.1.4 Experimental validation of the model’s accuracy estimates

In the comparisons above, I obtained the best predictive performance for the model based on cliques of orthologs enhanced by both inferred orthologs and paralogs. I evaluated the ability of each model to generalize to novel data, the poorly characterized genes, with an out-of-bag method for testing predictive performance: I measured accuracy on a random subset of the annotated phylogenetic profiles left out when inferring the functional annotation model. This method was shown to give unbiased estimates of predictive performance.

To validate how realistic are these out-of-bag performance estimates, I chose annotations for 38 genes in *Escherichia coli* K-12 having at least 60% expected Precision, for three GO terms that were straightforward to investigate experimentally using readily available antibiotics: “DNA damage response,” “translation,” and “peptidoglycan-based cell wall biogenesis.” The 38 *E. coli* strains, each with the deletion of one among the 38 selected genes, were grown in the presence of antibiotics that target the above Biological Processes: nalidixic acid (causes severe DNA damage, including double-strand breaks), kasugamycine (inhibitor of translation initiation), and ampicillin (inhibitor of cell wall synthesis) (Figure 4.5).

To each of the 38 genes the model assigned a Precision, as explained in the Materials and methods section. For example, Precision associated with the *E. coli* gene *yfgI* for “DNA damage response” was 62%; for “translation” and “peptidoglycan-based cell wall biogenesis” it was lower than 1%. I would therefore predict this gene to be involved in “DNA damage response” with a probability of being a false positive of 38% (100 − 62). For the other two GO terms the probability of being a false positive would be over 99%: the annotation model inferred that these are unlikely functions for this gene.

To experimentally evaluate a predicted annotation, the *E. coli* mutant
4.1. Results

A

Number of specific Gene Ontology annotations

<table>
<thead>
<tr>
<th></th>
<th>Experimental and curated annotations available in 2012-07-02 UniProt-GOA release</th>
<th>Model including orthologs (Precision threshold 90%)</th>
<th>Model including orthologs and paralogs (Precision threshold 90%)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

B

Biological Process annotations for *Escherichia coli* K-12

- cellular process (910)
- metabolic process (1521)

Selected examples of specific annotations:
- cilary or flagellar motility (3)
- cellular biogenesis or biogenic amine biosynthetic process (1)
- pentose-phosphate shunt (1)
- RNA aminoacylation for protein translation (4)
- NADP regeneration (1)
- regulation of cell shape (3)
- peptidoglycan-based cell wall biogenesis (1)
- protein secretion (4)
deleted in the gene whose function I predicted was compared to the E. coli wild type when grown in the presence of the antibiotic that inhibits the predicted function. If the gene is indeed involved in the predicted function, the survival of the mutant is expected to be lower than the survival of the wild type. For example, I predicted “DNA damage response” for the E. coli yfgI gene, so the corresponding mutant and the wild type were grown in the presence of DNA-damaging nalidixic acid; we expect the mutant to have lower survival than the wild type because its DNA repair capabilities are diminished.

We might predict a particular function, such as “DNA damage response,” for an important gene that is indirectly involved in many biological processes. Deleting such a gene may lower survival non-specifically and thus appear to validate our prediction. To control for this, each mutant was grown in the presence of the two additional antibiotics. For the above example of the yfgI gene, if my prediction is correct, the survival of the mutant should not be different from the survival of the wild type when grown on kasugamycin or ampicillin.

Therefore, I considered a prediction confirmed only if both of the following criteria were satisfied: 1) the survival of a mutant was lower than 25% of the wild type when grown with the addition of the antibiotic inhibiting the process predicted by our model, and 2) the survival of the mutant was higher than 50% of the wild type when grown on the other two antibiotics.

For example, I predicted “DNA damage response” for the E. coli yfgI gene: when grown on DNA-damaging nalidixic acid, the yfgI mutant had 7% survival of the wild type, but when grown on kasugamycin or ampicillin, the survival was much higher: 98% and 74% of the wild type, respectively (please see Table S1 in the accompanying publication). I therefore consider the prediction for the involvement of the yfgI gene in DNA repair processes confirmed: the yfgI mutant is sensitive to a DNA-damaging agent, while exhibiting wild type-like resistance to other tested stresses.

With these criteria, 25 out of 38 genes had confirmed predictions, which is equivalent to the experimental Precision of 66% (Figure 4.5). Since the selected genes had an expected Precision of 60%, the experiments show that

**Figure 4.4 (preceding page): Existing and predicted annotations for the representative prokaryotes.** A) The number of our model’s predictions at Precision 90% compared to the available curated Gene Ontology (GO) annotations. Each bar summarizes the data for one prokaryote: *Escherichia coli* K-12, *Listeria monocytogenes* serotype 4b str. F2365, *Mycobacterium tuberculosis* H37Ra, *Pseudomonas aeruginosa* UCBPP-PA14, *Staphylococcus aureus* subsp. *aureus* NCTC 8325, and *Streptococcus pneumoniae* R6. For both our predictions and the available GO annotations I show GO term annotations that have Information Content higher than 3. The colour of the bar denotes the source of the annotations: yellow for the model that includes all inferred one-to-one, one-to-many, many-to-one, and many-to-many orthologs, blue for the model that includes both these orthologs and the paralogs, and green for the curated annotations available in the 07-02-2012 UniProt-GOA release. B) Biological Process (BP) annotations that the model including both orthologs and paralogs assigned to *E. coli* genes at Precision 90%. Apart from the most general terms in the BP ontology, I highlight some more specific annotations.
4.1. Results

the estimates of accuracy provided by the model are realistic. In fact, 14 of the 38 tested genes have Precision \( \geq 85\% \). For these genes, the experiments have shown 11 out of 14 (79\%) to be correct, approximately matching the expected precision of 85\%.

Consequently, these estimates can be used to guide decisions when prioritizing genes for an in-depth experimental investigation of function in the wet lab.

4.1.5 Examples of novel functions for *Escherichia coli* genes supported by literature evidence

In addition to the systematic experimental verification of novel annotations in three GO categories as described above, here I highlight individual predictions for which I found supporting evidence in the publicly available literature.
4. Phylogenetic profiling with orthologs and paralogs

This information was not available to the classifiers at the time when the models were constructed. The following examples are for *E. coli* K12, as this is by far the best-studied model prokaryote [58].

Genes *hypC* and *hybG* are predicted to have “nickel cation binding.” These genes had no GO terms assigned in the 07-02-2012 UniProt-GOA release ([http://www.uniprot.org/uniprot/P0AAM3](http://www.uniprot.org/uniprot/P0AAM3) and [http://www.uniprot.org/uniprot/P0AAM7](http://www.uniprot.org/uniprot/P0AAM7)), and we therefore considered them unannotated. In the meantime, *hypC* was annotated with “metal ion binding” using experimental evidence: this is a parent GO term of our prediction. Moreover, when examining the literature, I found evidence that these two genes are involved in the biosynthesis of the [NiFe] cluster [21].

Another prediction is for *gltL*: it is predicted to be annotated with “ATP-binding cassette (ABC) transporter complex.” In line with our predictions, PortEco, a portal that includes information from 14 different *E. coli* data resources [81], labels the gene as “ATP-binding component of ABC superfamily.” Note that more general electronic GO annotations were available for this gene, e.g., “ATP binding,” “ATPase activity,” and “ATP catabolic process” ([http://www.uniprot.org/uniprot/P0AAG3](http://www.uniprot.org/uniprot/P0AAG3)).

A similar prediction of a more detailed function is for *ybgI*, where I predict GO terms from both BP and MF ontologies. This gene is known to be a conserved metal-binding protein [65], having an electronic GO annotation “metal ion binding”; I predict it is annotated with the BP GO term “Mo-molybdopterin cofactor metabolic process.” Based on the structure of the protein, Ladner *et al.* hypothesize this protein is a hydrolase-oxidase enzyme [65]; we predict this protein is annotated with the MF GO term “hydrolase activity, acting on acid anhydrides, in phosphorus-containing anhydrides.”

### 4.1.6 Predictions available to browse or download from the GORBI website

Because I showed my functional annotation model gives realistic estimates of predictive accuracy, I made our predictions freely available in a Web site [GORBI](http://gorbi.irb.hr/). The predictions can be queried either using GO accession number, NCBI taxonomy ID, or gene/protein ID (Figure 4.6). For example, one can focus on more general or more specific GO terms, depending on their position in the ”Gene Ontology DAG” (Figure 4.6, insert A). In addition, an experimenter can tune the search parameters to get a small number of high-confidence candidates, or a larger number of candidates that potentially have more false predictions, depending on the availability of annotations for the desired function and the available resources for experimental validation.

### 4.2 Discussion

#### 4.2.1 Phylogenetic profiles in functional annotation

The intuition of phylogenetic profiling is that corresponding genes gained and lost together in different genomes are likely to share function: they
4.2. Discussion

Figure 4.6: A screenshot of the GORBI web site. Predictions can be browsed in three ways, and any of their combination: 1) query using a Gene Ontology (GO) accession number (e.g., 6418 for "tRNA aminoacylation for protein translation"); alternatively, one can browse the "Gene Ontology DAG" (insert A) to find interesting GO terms, 2) query using an organism’s NCBI taxonomy ID (e.g., 288681 for Bacillus cereus E33L); alternatively, one can browse the "Taxonomy tree" (insert B) to find the NCBI taxonomy IDs of interesting organisms, 3) query using protein identifiers: NCBI Gene ID, UniProtKB protein ID, RefSeq ID, UniRef ID, UniParc, or EMBL ID. The results (insert C) list NCBI taxonomy ID and the name of the organism, Entrez Gene ID, UniProt ID that links to the corresponding protein’s page in the UniProt Knowledge Base, GO accession number, the name and the expected precision for the prediction.

could be involved in the same metabolic pathway, which is therefore incomplete without all the members in a genome [92]. Additionally, even if the two genes are parts of separate pathways and don’t strictly require each other for function, they could both share a role beneficial for survival in a particular environment [10].

The standard way of finding corresponding genes in different genomes is via sequence homology: in addition to inferring function via homology, a phylogenetic profile allows to infer function based on the presence or absence of the corresponding genes in a range of organisms.

4.2.2 Orthologs and paralogs in functional annotation

In functional annotation, one often differentiates between two subtypes of homologs, orthologs and paralogs [e.g., 121]. According to the standard model of genome evolution, paralogs—pairs of genes diverged through a
4. Phylogenetic profiling with orthologs and paralogs

duplication event—could obtain a new function \[133\]. Conversely, orthologs are pairs of genes diverged through a speciation event and should be more likely to retain function; they are therefore expected to be more informative in functional annotation \[64\].

However, the exact quantification of the functional divergence in a pair of orthologs and a pair of paralogs is not fully resolved. It was observed that the search for homologs using the best bidirectional hit approach, without explicitly distinguishing orthologs from paralogs, produces a higher level of functional compactness via Gene Ontology (GO) terms than is present in the ortholog databases Homologene \[106\] and OMA \[3\]. In addition, Studer and Robinson-Rechavi list scenarios where the standard model—predicting that paralogs diverge in function more than orthologs—is invalid; for example, cases where paralogs share function, and orthologs do not \[118\].

A recent large-scale study further challenged the veracity of the standard model: the authors compared mouse and human ortholog and paralog pairs and surprisingly found that paralogs tend to conserve function more than orthologs \[84\]. This finding caused a stir in the community—demonstrating the relevance of the topic—but was subsequently challenged in two publications \[4, 123\].

Nevertheless, a recent systematic survey showed that the divergence in function between paralogs is not as strong as the standard model would imply \[4\]. In addition, we know that homologs—orthologs and paralogs combined—are useful in functional annotation, especially when their sequence similarity is above the ”twilight zone” \[103\]. Further, orthologs and paralogs share a common ancestor: paralogs, as well as orthologs, could carry functional information useful for annotation.

4.2.3 Paralogs enrich phylogenetic profiles

In line with the recent research \[4\], our results show that the standard model, when viewed in the functional annotation context, tends to draw too strong of a line between orthologs and paralogs. When I enriched clique-only annotation models with additional orthologs and additional paralogs, I obtained a model that outperformed both the model that was enriched only with orthologs and the model that was enriched with refined homologs at different evolutionary distances (Figure B.6). The improvement is most notable in the number of new annotations I was able to assign: while keeping the Precision at the same high level, our best model increases Recall (Figure 4.3), and consequently gives us more predictions at the same level of correctness (Figure 4.4, panel A).

Even so, our results do not contradict the standard model in two major points: 1) cliques of orthologs/groups where all genes are connected with orthologous relations are indeed functionally very similar (Figure 4.7), and 2) our results support the current widespread annotation efforts that use homology: even when we disregard the orthology/paralogy relationships to enhance cliques, we obtain high predictive accuracy (Figure B.6).
4.2.4 Orthologs-only model outperforms paralogs-only model

The OMA algorithm infers paralogs (i.e., non-orthologs) among genes linked as the best bidirectional hits in the respective genomes: a witness to non-orthology breaks the link between two genes [104]. Because only one witness is enough to break the orthologous relationship, the OMA algorithm produces orthologous groups with high specificity [2]. As a trade-off, the set of inferred paralogs might contain pairs whose orthologous link was erroneously broken; the probability for this to happen increases with the addition of new genomes (A. Altenhoff, personal communication). Therefore, our set of paralogs might contain orthologs that were misclassified as paralogs.

Even so, when I enriched clique-only annotation models with the inferred paralogs, predictive accuracy increased less than when I enriched clique-only annotation models with the missing orthologs (Figure 4.2, models c and b, respectively). I obtained these results despite enriching with a larger number of paralogous pairs than orthologous pairs: it is not the number of added pairs that improves the predictive accuracy, but the genome they are located in.

4.2.5 Experimental validation shows that the model’s performance estimate is realistic

An important output of any computational annotation model is an estimate of confidence for the annotations: it can subsequently be used to guide decisions about experimental validation. In fact, one project that provides a framework for the exchange of information between the computational and experimental communities is COMBREX [102, 101]. To meaningfully contribute to growing resources such as COMBREX, I wanted to evaluate whether our annotation model provides realistic estimates of confidence for the individual annotations.

Probing growth profiles of knockout E. coli mutants with sub-lethal concentrations of antibiotics is an established method of functional annotation [85, 33], allowing an experimental validation of whether a gene is involved in the predicted Biological Process by growing the respective knockout E. coli mutant in a medium containing the antibiotic that targets the Biological Process I predicted. The experimental results support the estimates of Precision obtained from a cross-validation procedure, serving as a proof of principle that our phylogenetic profiling-based model is useful when searching for novel functions of poorly annotated genes in a microbiology lab.

Our annotation model assigns GO terms from across the GO hierarchy, for both general and specific terms. Overall, more general terms tend to have a higher cross-validation Area Under the Precision-Recall Curve (AUPRC) (Figure B.3) and consequently the annotations assigned with these terms are more likely to be correct.

The AUPRC scores such as the one I use serve as a test of the internal consistency of the model. On the one hand, the model captures the similarities of the phylogenetic profiles of the OMA cliques (and the enriched OMA cliques) of orthologs; on the other hand, the model captures the GO
terms assigned to the OMA cliques of orthologs. For a given GO term, the AUPRC scores will be low if the phylogenetic profiles’ features cannot be used to transfer the function between the profiles. Thus, I make no prior assumptions whether a GO term at a certain level of specificity can be transferred across profiles, but rather infer this from the data itself in a systematic manner.

An experimenter can focus on more general or more specific GO terms depending on the trade-off of reported Precision and the cost/time required of experiments; when using the GORBI Web site, GO terms can be selected depending on their position in the Gene Ontology hierarchy (Figure 4.6, insert A).

To facilitate the use of the generated computational annotations, I provide them in a Web site GORBI ([http://gorbi.irb.hr/](http://gorbi.irb.hr/)) where each prediction is accompanied by the annotation model’s estimate of confidence.

### 4.3 Conclusions

I contribute a solution for computational annotation of genes that utilizes a distinction between two types of homologs—orthologs and paralogs—to yield an innovative annotation model: phylogenetic profiles derived from cliques of orthologs enriched with both orthologs and paralogs have shown the best predictive accuracy. Our results are in line with related recent research: while it is generally accepted that pairs of orthologs have a lower rate of functional divergence, the divergence in paralogous pairs is not as strong as the standard model would imply [4].

In addition, validation experiments in knockout mutants of *E. coli*, show that my annotation model reports realistic measures of predictive performance. The agreement with the experimental results implies that my functional annotations—and the corresponding confidence estimates—can be used to narrow the search space for potential function candidates and thereby help to bridge the widening gap between the sequenced and characterized proteins.

For successful annotation of newly sequenced proteins, we need contributions from both the computational community—a large number of credible annotations—and the experimental community—validating the most interesting computational annotations. In turn, the validated findings from the wet-lab can be fed into the computational annotation pipelines, helping to propel a virtuous circle that increases the number of experimentally annotated genes.

My research aims to contribute to the understanding of the deluge of data we face, whether from complete microbial genomes for which we provide annotation predictions ([http://gorbi.irb.hr/](http://gorbi.irb.hr/)), or from the metagenomics projects, in particular the emerging human microbiomes, to which I can apply my annotation model.
4.4 Materials and Methods

4.4.1 Annotation data

I downloaded all the annotation data from the FTP site of the UniProt-GOA database [13]. I used the Gene Ontology (GO) vocabulary for functional annotation [11]. I included all annotations assigned by a curator (evidence codes EXP, IMP, IGI, IPI, IEP, IDA, ISS, RCA, IC, NAS, TAS), and from the non-curated annotations (evidence code IEA), I included those inferred from UniProtKB keywords, UniProt Subcellular Location terms, Enzyme Commission numbers, and InterPro (reference codes GO_REF:0000004, GO_REF:0000023, GO_REF:0000003, and GO_REF:0000002, respectively). Despite not being curated, my work described in chapter 3.2 showed these electronic annotations are of high quality, in particular for the only analysed Prokaryote, *E. coli* (Figure B.7).

I express the specificity (opposite of generality) of a GO term $GO_i$ with respect to its Information Content:

$$\text{InformationContent}(GO_i) = -\log_2(freq(GO_i)),$$

where $freq(GO_i)$ is the frequency of $GO_i$ among all annotations for the twelve Reference genomes [98].

4.4.2 The OMA algorithm and the OMA database

The OMA algorithm is a graph-based method of orthology inference [104]. Roth et al. provide full details of the algorithm, and I summarize the main points relevant to my work. The algorithm starts with an all-against-all sequence alignment: proteins from two species are connected if they are best bidirectional hits, within a confidence interval, in the compared species. The connections between a pair of proteins are broken when one of them is the best bidirectional hit with one of the proteins in a connected pair in some third species, and the other is the best bidirectional hit with the second protein in the same pair; the broken pairs are inferred paralogs. The remaining connections are inferred orthologs. Finally, OMA cliques of orthologs are sub-graphs where all proteins are connected by orthologous relationships (Figure 4.1).

In this work, I only use OMA cliques that group at least 10 members.

The OMA algorithm is available as a stand-alone version; the results can also be browsed on the OMA web site [108].

4.4.3 Annotating OMA cliques of orthologs

Because one essential component of our work is annotating OMA cliques of orthologs based on the proteins they contain, I first checked whether OMA cliques contain proteins with the same function. First, unannotated OMA members were labelled with the GO terms of annotated OMA members at four thresholds: if 30, 50, 70, or 90% of OMA members have the respective function. To assign these labels, I used only annotations available in the 16-01-2008 UniProt-GOA release.
Next, I checked the annotations in the more recent 17-10-2011 UniProt-GOA release. For each unannotated protein, I consider the labelled function to be confirmed if the protein holds the respective annotation in the more recent release; I consider the labelled function to be rejected if the protein holds the same annotation alongside a ”NOT” qualifier (explicit rejection) or a new annotation that is not the propagated one (implicit rejection). To make a more robust measure, I summarize the confirmed and rejected annotations for each GO term. I named this measure ”Coherence of a GO term.” More formally,

\[
\text{Coherence}(\text{GO}_i) = \frac{C_{\text{GO}_i}}{C_{\text{GO}_i} + R_{\text{GO}_i}}
\]

where \( C_{\text{GO}_i} \) is the set of confirmed annotations associated with term \( \text{GO}_i \) and \( R_{\text{GO}_i} \) is the set of rejected annotations associated with term \( \text{GO}_i \). I account for the definition of the GO: the assignment of any GO annotation assumes the assignment of all the GO parent terms.

This is a conservative estimate of Coherence: I consider as rejected an annotation that might not have been added to the database yet. Annotations are continuously being added to UniProt-GOA database, and the annotation update interval for a gene can be as long as 12 years \[132\]. To compensate for this bias, I evaluated coherence on a three-year interval, as most genes in \textit{E. coli} are updated within that time frame.

![Figure 4.7](image)

**Figure 4.7: Functional coherence of GO annotations** Each panel presents the results of evaluation for the annotations inferred from A) 30%, B) 50%, C) 70%, and D) 90% of OMA members. I propagated the annotations available for the OMA group members in the 2008-01-16 UniProt-GOA release to unannotated group members, and evaluated the predictions with the newly arrived annotations in the more recent 17-10-2011 UniProt-GOA release.

For each GO term, the functional coherence depends on the imposed an-
notation threshold (Figure 4.7): when a larger fraction of OMA members in 2008 supported the GO annotation, I found more newly annotated proteins that support this propagated GO annotation in 2011. The drawback of the increasing threshold was a smaller number of GO terms that can be used in annotation and consequently a smaller number of annotated OMA groups used in training the annotation model. I chose the threshold of 50% as a compromise: for most GO terms, the newly annotated proteins are in accordance with the propagated functions—fraction of correctly predicted newly arrived annotations is greater than 0.9—and we are left with enough specific GO terms for functional annotation (Figure 4.7, panel C): 422 GO terms from the Biological Process ontology, 48 GO terms from the Cellular Component ontology, and 264 GO terms from the Molecular Function ontology. I use the 50% threshold throughout this work.

4.4.4 Phylogenetic profiles

The phylogenetic profile of an OMA clique of orthologs is encoded as a vector of binary values. The vector’s length is 998 items—the number of prokaryotic genomes included in our work. Each position in the vector indicates the presence or absence of an OMA clique member in the respective genome. There are 64052 annotated and unannotated OMA phylogenetic profiles in our dataset (Figure 4.1).

I enriched the phylogenetic profiles, first by connecting the missing orthologs to OMA clique members (Figure 4.1, full lines), and second by connecting the paralogs (Figure 4.1, dashed lines) to OMA clique members. Orthologs include one-to-one orthologs, one-to-many orthologs, many-to-one orthologs, and many-to-many orthologs.

4.4.5 Machine learning algorithms

The Clus-HMC algorithm [107] builds decision trees for hierarchical multi-label classification (HMC). In contrast to ordinary classification trees [17], which can be used for single-label annotation, Clus-HMC is able to deal with multiple, hierarchically organized class labels, such as terms from the Gene Ontology. It builds decision trees for HMC by extending the standard decision tree learning algorithm: It splits the training data into subsets based on attribute values, in order to minimize the weighted sum of variances for all class labels within the subsets resulting after the split [107].

In this weighted sum, a parameter \( w_0 \) can be used to place more weight on either the more specific, or the more general GO terms. The default value of this parameter is 0.75, which places more weight on more general terms. Changing the default value of the \( w_0 \) parameter to place more weight on the specific terms will favour them, possibly trading off the accuracy of the more general terms for a gain in accuracy of the more specific terms. To test for possible gain, I experimented with different values of the \( w_0 \) parameter to place higher weight on either the more general GO terms (default value, \( w_0 = 0.75; w_0 = 0.5 \)) or on the more specific GO terms (\( w_0 = 1/0.75 = 1.33; w_0 = 1.75; w_0 = 2.0; w_0 = 3.0 \)). Clus-HMC-Ens proved to be robust to the value of the \( w_0 \) parameter: I did not record a significant change in the AUPRC values (p-value was not lower than 0.28 in the five tested values.
of the $w_0$ parameter, Wilcoxon signed-rank test), and I therefore used the default value in all our computational experiments.

In addition, the hierarchy of class labels introduces dependencies between the classes: Clus-HMC is aware of the hierarchical relationships between the multiple labels and uses this information to improve predictive performance.

The Clus-HMC algorithm was extended to an ensemble setting (Clus-HMC-Ens) [129], where a forest of decision trees for HMC is learned: The predictions of the individual trees are combined to obtain the overall prediction of the ensemble. Clus-HMC-Ens implements, among other methods, the Random Forest (RF) ensemble approach [20], where the individual trees are constructed by using a randomized version of Clus-HMC. Each tree is constructed from a different sample of the training dataset: The bagging (Bootstrap aggregating) methodology of resampling the dataset [18] is used to construct the different samples. One bootstrap sample consists of the same number of examples as the original dataset, but they are randomly drawn with replacement; consequently a bootstrap sample contains about two thirds of unique examples. A model—Clus-HMC decision tree—is produced from each of the bootstrap samples.

When estimating the classification error, out-of-bag estimates are calculated. The examples that were omitted from the bootstrap sample—one third of the original dataset—are used in calculating Precision, Recall, and Area Under the Precision-Recall Curve (AUPRC). The estimates are based on the random sample, and the measures are therefore unbiased. To check whether adding paralogs improves the functional annotation model regardless of the machine learning algorithm used, I inferred functional annotation models with the standard approach used in phylogenetic profiling: transfer of function via pairwise distance measures between phylogenetic profiles, as implemented in a kNN classifier (Figure B.7). The conclusions presented above do not change: the model that includes both orthologs and paralogs outperforms the model that includes only orthologs. Because Clus-HMC-Ens outperforms kNN in computational efficiency and predictive accuracy, I used Clus-HMC-Ens throughout this work.

### 4.4.6 Evaluating the functional annotation models

I compare models of functional annotation using Precision-Recall curves: in the Precision-Recall space, Recall is on the x-axis, and Precision is on the y-axis. Traditionally, Precision and Recall are defined for binary classification: an instance either has or does not have the label; in our case, each OMA clique either has or does not have a GO annotation. Precision and Recall are defined for each GO term:

$$\text{Precision}(GO_i) = \frac{TP_{GO_i}}{TP_{GO_i} + FP_{GO_i}}$$

$$\text{Recall}(GO_i) = \frac{TP_{GO_i}}{TP_{GO_i} + FN_{GO_i}}$$

where $TP_{GO_i}$ is the number of correctly predicted true annotations ("True Positives"), $FP_{GO_i}$ is the number of incorrectly predicted true annotations ("False Positives"), and $FN_{GO_i}$ is the number of missed true annotations ("False Negatives").

Precision stands for the fraction of correctly predicted examples out of all the predictions, and Recall stands for the fraction of correctly predicted
examples out of all known to be true.

Here, we are dealing with a multi-class problem: each OMA clique can be annotated with multiple GO terms. The classifier I am using is adapted for such a problem and assigns a probability that each OMA group is assigned each of the GO terms. By varying a cut-off for the probability from 1.0 to 0.0, I am relaxing the stringency of the predictions: an increasing number of OMA groups are assigned an increasing number of GO terms. Fixing this cut-off at the three values and calculating Precision and Recall for each GO term created visualizations in Figure 4.3.

The probabilities allow us to have a ranking of GO annotation predictions for OMA cliques and proteins therein. In addition to the ranking, I wanted to have an intuition for the number of candidates one needs to experimentally examine in order to get confirmed annotations. Therefore, I translated the probabilities to Precision for each GO term. Similarly as above, I varied the cut-off for the probability, and calculated the corresponding Precision for each GO term at each probability cut-off: out of all the OMA clique annotations that pass the threshold, I counted the number of true positives, and the number of false positives.

To compare models in Figure 4.2, I used a single measure of performance that combines Precision and Recall: Area Under the Precision-Recall Curve (AUPRC). To calculate AUPRC, I first varied the probability cut-off from 1.0 to 0.0 and obtained the Precision-Recall curve. I then calculated the area that is enclosed between the Recall axis and the curve. The closer AUPRC is to 1.0, the better the model.

4.4.7 Bacterial strains, growth conditions, and antibiotic treatments

All deletion mutants used herein were derived from wild-type sequenced *Escherichia coli* MG1655 by P1 transduction. P1 phage was grown on a series of Keio collection deletion mutants listed in Table S1 of the original publication [115]. Successfully transduced mutants were selected on LB plates supplemented with kanamycine.

Bacteria were grown in LB broth at 37°C, to the exponential phase (OD600 = 0.20.3). Viable cell counts were estimated by plating serial dilutions on LB plates, as well as LB plates supplemented with 400 ug/mL kasugamycine (inhibitor of translation initiation), 4 ug/mL nalidixic acid (causes severe DNA damage, including double strand breaks), and 3 ug/mL ampicillin (inhibitor of cell wall synthesis). Plates were incubated overnight at 37°C. The concentrations of antibiotics used in this study were selected as the concentrations that lead to ∼10% survival of the wild type *E.coli*.

4.4.8 Sources of data and software

1. The orthology and paralogy data from the OMA database, May 2011 version was kindly provided by A. Altenhoff.

3. GO annotations were downloaded from the UniProt-GOA FTP site. I used the 2008-01-16 and the 2011-10-17 UniProt-GOA releases to evaluate the consistency of OMA group annotations, 2011-10-17 UniProt-GOA release to create all the annotation models, and 07-02-2012 UniProt-GOA release to estimate the frequency of occurrence of a GO term in the UniProt-GOA database [http://www.ebi.ac.uk/GOA/](http://www.ebi.ac.uk/GOA/).


5. Final dataset files in ARFF format, as given to the Clus-HMC-Ens algorithm (Datasets S1 and S2 of the original publication [115]).

Chapter 5

Phylogenetic profiling: how much input data is enough?

Phylogenetic profiling is a technique to infer gene function from patterns of presence and absence across species [92], using the principle of “guilt-by-association”: if homologs are inherited or lost co-dependently, they are likely to be functionally or physically interacting. Phylogenetic profiling has been successfully used to discover protein-protein interactions and to annotate new genomes, particularly in prokaryotes; reviewed in [63]. These successes can be attributed to a combination of methodological refinements [28, 138, 26, 115], increase in the number of sequenced genomes, and increase in the number of functional annotations—Gene Ontology (GO) annotations in particular [38]. However, whereas most of the previous literature on phylogenetic profiling has focused on methodological refinements, comparatively little is known about the relative contribution of more genomes and more functional annotations. As there are now about 17,500 bacterial genomes sequenced [88] including all of them is not practical. Wanting to use phylogenetic profiling to functionally annotate a new bacterial genome, a researcher might thus ask: how much input data is enough to achieve good functional predictions?

Here, I sought to answer this question by decomposing the effect of additional genomes and additional functional annotations on phylogenetic profiling performance. In our tests, phylogenetic profiling generally improved with more input data, but the marginal benefit of additional genomes rapidly diminished beyond ∼100 genomes and practically vanished beyond ∼400 genomes.

In general, the availability of new annotations is greatly beneficial for phylogenetic profiling: the annotations added in the new releases of the UniProt-GOA database has drastically increased the predictive accuracy of phylogenetic profiling. Conversely, decreasing the number of annotations lead to a clear detrimental effect on the overall performance. For a minority of relatively general GO terms, there are already sufficient annotations in
the database to train accurate predictors of functional annotation, but for most GO terms, annotations are still lacking.

Previous work has purported the view that genomes need to be carefully selected, typically by maximising the phylogenetic diversity among the genomes [61, 114, 83]. Our analyses confirmed that this is the case and that, conversely, considering only particular subclades decreased predictive accuracy. I also observed that while maximising phylogenetic diversity lead to the best results, the simple strategy of randomly selecting subsets from all available genomes was nearly as effective.

Finally, I show that my results hold even when I account for the lack of comprehensive annotations in biological databases. Because we can never be sure whether particular gene products are comprehensively annotated—biological databases are subject to the Open World Assumption [32]—I introduce a new testing framework based on subsets of well-annotated proteins and sparsely-annotated ones, and ascertain that my conclusions hold under the different conditions.

5.1 Results and Discussion

First, I show that increasing the amounts of data in biological databases has led to better predictive accuracy of phylogenetic profiling. Then, I separately consider the influence of the amount of annotations and the number of sequenced genomes. I also discuss the problem of selecting optimal subsets of genomes. Finally, I provide evidence that our conclusions hold in light of the Open World Assumption—the notion that functional annotation databases are incomplete [123, 32].

5.1.1 Phylogenetic profiling strongly benefits from more data

To evaluate the influence of new annotations and newly sequenced genomes on phylogenetic profiling, I first retrieved a set of 1093 GO terms for which I can reliably assign annotations using the 2013 UniProt-GOA database (see Methods). I then retrospectively applied my phylogenetic profiling method on successive past versions of the UniProt-GOA database, comparing the predictive accuracy of the method on these sparser datasets.

For these GO terms, the combination of the newly sequenced genomes and the new annotations had a favourable effect: on average, the predictive performance of phylogenetic profiling steadily improved each year (Figure 5.1, panel A). The difference in predictive accuracy between the annotation model available in 2005 and 2013 is extreme.

For some GO terms, the earlier releases of the database do not have enough data to infer annotations: for example, the 2005 annotation database has insufficient information for 846 of the above 1093 GO terms; there, AUPRC is zero resulting in the heavy base of the first series in Figure 5.1 panel A. In the second set of experiments, I focused on the GO terms that had sufficient annotation information throughout the analysed releases, thereby evaluating the influence of new annotations and newly sequenced genomes on this smaller—and more general—set of 243 GO terms.
5.1. Results and Discussion

![Figure 5.1](image)

**Figure 5.1: Performance of phylogenetic profiling, measured in Area Under the Precision-Recall Curve (AUPRC), when I change the amount of data available for training the model for functional annotation.** For each year from 2005 to 2013 denoted on the x-axis, the corresponding dataset includes those genomes that were available both in the OMA database, as well as the NCBI taxonomy database in the respective year; the phylogenetic profiles are annotated using the UniProt-GOA file available in January of the respective year. Each violin plot summarizes the distribution of GO terms according to the AUPRC value: the area of the plot corresponds to the probability density of GO terms at different values of AUPRC; the black dot denotes the mean value of AUPRC. A) I consider 1093 GO terms in total—those that had sufficient annotation information in the most recent database releases. If the model does not have enough data to infer annotations for one of the 1093 GO terms, as will be the case for, e.g., 846 of these GO terms using the data from 2005, its AUPRC score is zero. B) I consider only the GO terms that had sufficient annotation information throughout the analysed releases. C) I consider all the GO terms from the prokaryotic GO set.

Just like in the first set of experiments, in the second set of experiments the average predictive accuracy increases with more data (Figure 5.1, panel B). However, here I see very little increase in predictive accuracy after 2010: it is tempting to speculate that for this set of GO terms we have approached the maximum predictive accuracy that our implementation of phylogenetic profiling can achieve, regardless of the number of newly sequenced genomes or new annotations.

In our third set of experiments, I supplemented my set of 1093 GO terms with all the GO terms that can be assigned to prokaryotic proteins, thereby obtaining 7666 GO terms. Even using the most recent release of the database, for 6573 (7666 − 1093) of these AUPRC will be zero, as I do not have enough data to make a prediction (Figure 5.1, panel C). Considering the data from this angle shows that the functional annotation of genomes remains largely incomplete: despite assigning many GO terms with high AUPRC values and seeing a general trend of increasing predictive accuracy, there are many more GO terms for which I cannot make predictions because of the lack of annotations.

In all three sets of experiments, the observed increase in predictive accuracy is a consequence of both the newly sequenced genomes, which increased the number of genomes included in the phylogenetic profiles, as well as the
new functional annotations. In what follows, I separately evaluate the influence of each of these two factors, focusing on the most recent database releases available and the set of 1093 GO terms I could therewith reliably predict (see Methods).

5.1.2 How many annotations are enough?

![Graph showing the performance of phylogenetic profiling](image)

**Figure 5.2:** Performance of phylogenetic profiling, measured in AUPRC, when I reduce the number of annotations used for phylogenetic profiling. For each of the experiments denoted on the x-axis, I only used a fraction of the available annotations in the most recent dataset. Dashed and full lines connect the dots of the mean AUPRC scores for two sets of experiments: random sub-selection of genomes (full lines) and sub-selection to keep maximum diversity among the selected genomes (dashed lines). Colour denotes the number of genomes used in the phylogenetic profiles.

In theory, an increase in the number of functional annotations should always be beneficial for phylogenetic profiling: the more training data I have, the better our functional annotation. I was interested in evaluating what effect changing the amount of annotations has on predictive accuracy: I systematically reduced the number of annotations available to annotate our phylogenetic profiles and then evaluated predictive accuracy (Figure 5.2).

For the overall predictive accuracy of the 1093 GO terms included in this analysis, removing annotations drastically reduced predictive accuracy.
This held across different number of genomes considered in the phylogenetic profiles, and for different genome selection criteria (factors which I discuss in detail below). In particular, upon discarding annotations down to 20% of the current levels, phylogenetic profiling became completely ineffectual regardless of the number of species considered (Figure 5.2). To put this result in perspective, one should consider the viewpoint of future database releases: the amount of currently known annotations is what I simulate by discarding annotations (left side of the curves in Figure 5.2).

At the right end of the curves in Figure 5.2 the improvement in performance of phylogenetic profiling tapers off beyond 80% of added annotations. However, because I only consider terms that are associated with at least 50 phylogenetic profiles (see Methods), the 1093 GO terms included in this analysis are inherently skewed toward well-annotated terms. Consequently, at the point when 80% of annotations are considered, there often remains enough annotations for terms to be reliably assigned to phylogenetic profiles. Indeed, if I consider an even better annotated subset of 777 GO terms associated with at least 100 profiles each, the tapering beyond 80% becomes more pronounced (Figure C.1).

5.1.3 How many genomes are enough?

As previous studies have highlighted, including more genomes in phylogenetic profiles is not necessarily beneficial in terms of predictive accuracy [61, 114, 83]. Additional genomes in the profiles can influence predictive accuracy in two ways: 1) they may provide useful information for the annotation models, thereby increasing predictive accuracy; 2) they may provide redundant information, which would be at best neutral, or disruptive to methods that do not properly account for the correlation arising from evolutionary relationships between species [26].

To investigate how the number of genomes influences predictive accuracy, I selected genomes to include in phylogenetic profiles either randomly or to obtain maximum phylogenetic diversity (Figure 5.3). For each experiment in this and the next section, regardless of the number of genomes used, I kept the annotations of phylogenetic profiles the same: the annotations were those assigned to the phylogenetic profile when I used all the available data, so any difference in predictive accuracy is a consequence of the different genomes used.

The usefulness of adding more genomes is visible in our experiments: the more genomes get included in phylogenetic profiles, the better the predictive accuracy. In fact, contrary to previous reports [114], I obtained best predictive accuracy when I use all the available bacteria in our phylogenetic profiles.

It is the effect of adding redundant genomes (i.e., phylogenetically close genomes) that is discussed in many reports in the literature, arguing that for good predictive accuracy a well chosen subset of genomes is imperative [119]. In our experiments, redundant information in phylogenetic profiles did not lead to performance degradation because our algorithm properly deals with potential correlation among profiles (see Methods). As a result, additional genomes are never detrimental in terms of predictive accuracy.
However, additional genomes in the phylogenetic profiles result in a heavier computational burden. This raises the question of how many genomes are required for phylogenetic pattern methods to be accurate. Our analyses suggest that independent of the amount of annotations considered, phylogenetic profiling only marginally improves beyond \( \sim 100 \) genomes, with no practical difference beyond \( \sim 400 \) genomes (Figure 5.3).
5.1.4 Influence of genome selection on phylogenetic profiling: maximal phylogenetic diversity performs best and can nearly be achieved through random sampling

A reasonable strategy for the selection of genome subsets is to keep the largest phylogenetic diversity within the subset, either manually selecting the subset [61] or using a computational method [114]. In fact, preserving the maximum diversity within a set of genomes used for phylogenetic profiling was the best strategy in our experiments (Figure 5.3).

To get a better insight into the influence of larger phylogenetic diversity on the performance of phylogenetic profiling, I focused on two possible extremes when selecting my subsets: 1) I included phylogenetically close genomes that form a clade; 2) I used two phylogenetically close previously published sets of genomes shown to perform well in phylogenetic profiling.

First, I focused on five major phylogenetic groups of sequenced bacteria: Proteobacteria (538 genomes), Firmicutes (218 genomes), Actinobacteria (106 genomes), Bacteroidetes (35 genomes), and Cyanobacteria (39 genomes). I carried out three groups of experiments. First, I created phylogenetic profiles using only genomes in the respective groups. Second, I created phylogenetic profiles that include a random selection of bacteria such that the number of randomly selected bacteria is the same as the number of bacteria in each major group. I performed ten random selections of genomes, and the results reported in this work—the AUPRC values—are the mean values of these experiments. Third, I created phylogenetic profiles that include a selection of bacteria that maximizes phylogenetic diversity within the set such that the number of selected bacteria is the same as the number of bacteria in each major group.

In my experiments, phylogenetic profiling based only on clades is a poor strategy: the maximally diverse set always outperforms the randomly chosen set (Figure 5.3); even the dataset based on our largest clade of bacteria, Proteobacteria (538 genomes), is outperformed by the dataset containing the same number of maximally diverse bacteria (Figures 5.3 and C.2).

Second, I investigated how well my phylogenetic profiling method works with published subsets of genomes: I used two subsets available in the literature: a manually selected subset by Jothi et al. [61] and a computationally selected subset by Simonsen et al. [114].

Jothi et al. [61] evaluated a number of manually curated sets of genomes and found that a non-redundant set containing members from all three kingdoms dubbed BAE3a had on average the best accuracy when predicting functional linkages. In our experiments, the performance using the BAE3a set is inferior to the performance when using the same number of genomes, but selected from our set of 1078 bacteria, be it a random selection, or a selection to obtain a maximally diverse set of genomes (Figures 5.3 and C.2). Our selection was made from a much more diverse set of genomes than the set available in 2007 when Jothi et al. made their selection, and the difference in predictive accuracy might indicate the advantage of having a larger pool of sequenced genomes to chose from.

More recently, Simonsen et al. [114] presented methods to automate the
selection of genome subsets used in phylogenetic profiling; they focused on the inference of protein-protein interactions. Similar to our results when predicting function, they showed that genome selection with the largest phylogenetic diversity improves predictive accuracy. By removing closely related genomes from the set, they created a reference genome set that improved the predictive accuracy compared to both their full set of 980 genomes and the BAE3a set of Jothi et al. [61]. I reconstructed this set and compared its performance in my framework (Figure 5.3). In our experiments, the Simonsen et al. [114] dataset has a similar predictive accuracy as the BAE3a dataset and shows slightly lower, but similar predictive accuracy as our datasets obtained by either random sampling or a sampling that preserves the highest diversity within the dataset (Figures 5.3 and C.3).

5.1.5 Our main conclusions hold under the Open World Assumption

One challenge in assessing function prediction is to account for the inherent incompleteness of biological databases. Indeed, because gene products can have multiple functions, absence of functional annotation is not evidence of absence of function—a notion referred to as the Open World assumption (OWA) [123, 32]. In this section, I provide corroborating evidence that my findings hold in light of the OWA.

First, I ask: how are my analyses affected by the OWA? The predictive accuracy metric I used in this study (Area Under the Precision-Recall Curve, AUPRC) is computed from the out-of bag error: for each iteration of the experiment, a predictor is built using \( \sim 63\% \) of the data selected randomly. To each sample in the remaining \( \sim 37\% \)—samples not used to train the predictor—the predictor assigns the probability that a given OMA group of gene products is associated with each GO term. Thus, comparing these predictions with the actual GO term(s) assigned to each gene product yields True Positives, False Positives, True Negative, and False Negative counts, from which Precision and Recall can be computed.

However, counting predictions that are absent from the reference set as False Positives yields biased results under the OWA: absence from the reference set could be due to mere incompleteness of the database.

To control for the effect of this bias on our conclusions, I performed two sets of control experiments: 1) I repeated my analyses on a dataset that is likely to suffer less bias from violating the OWA: gene products with at least five annotations in the UniProt-GOA database. All else being equal, by virtue of having more annotations, these gene products can be expected to be more comprehensively annotated. 2) I repeated my analyses on a dataset that is likely to suffer more bias from violating the OWA than our normal dataset: I randomly left out 60\% of all the annotations in the UniProt-GOA database.

Both alternate datasets yielded highly consistent results (Figures 5.3 and 5.4). When I included in my phylogenetic profiles only the well-annotated proteins, I replicated the shape of the curve in Figure 5.3: selection to obtain maximally diverse sets of genomes outperforms the random selection, but marginally so; the effect of adding more than \( \sim 100 \) genomes diminishes.
5.2 Conclusion

When phylogenetic profiling was first described fifteen years ago [92], even with the few sequenced genomes that were available back then, the method provided new insights into protein function of *Escherichia coli*. Since then, there have been numerous refinements of the method, many focusing on the optimal selection of genomes for increasing the predictive accuracy of functional annotation.

Here, I have shown that an increase in available input data alone has
More data in phylogenetic profiling drastically improved the effectiveness of phylogenetic profiling. However, adding more genomes beyond \( \sim 100 \) only has a marginal effect on predictive accuracy. Furthermore, I have confirmed that maximising phylogenetic diversity yields best results, and that a straightforward random selection procedure from a broad range of species nearly achieves the same result.

Should the future releases of the annotation databases provide more of the more specific functional annotations, phylogenetic profiling will benefit and propagate these annotations to non-model organisms. Moreover, with the increase in the number of sequenced genomes, we will have a larger set of genomes to choose from and thereby further improve the phylogenetic profiles.

Phylogenetic profiling was a good idea fifteen years ago; owing to the additional functional and genomic data we now have available, it is an even better one today.

5.3 Methods

5.3.1 Phylogenetic profiling

To explore the extent to which newly sequenced genomes and the new information in the annotation databases influence phylogenetic profiling, I focused on a recently published method for phylogenetic profiling [115]. In my implementation, I tracked the patterns of presence or absence of homologs inferred by the OMA algorithm [104]; I included both orthologs and paralogs because our previous research showed that paralogs, when used in addition to orthologs in phylogenetic profiling, improve predictive accuracy [132]. Therefore, each row in the binary phylogenetic profile represents one OMA group, and a 1 or a 0 represents the presence or absence of an OMA group member and its orthologs or paralogs in the included genomes.

I assigned Gene Ontology (GO) terms [11, 38] to OMA phylogenetic profiles: when at least half of the OMA group members—those homologs that are all mutually orthologous—are annotated with the respective GO term, I assign the GO term to the OMA group itself, and consequently also to the OMA phylogenetic profile. For the functional annotations, I used the releases of the UniProt-GOA database available from the UniProt-GOA FTP site; for each denoted year, I used the first yearly release. If not otherwise noted, I only included GO terms that are assigned to at least 50 OMA phylogenetic profiles.

To annotate the phylogenetic profiles, I included GO terms assigned using evidence codes EXP, IMP, IGI, IPI, IEP, and IDA, as well as selected GO annotations assigned using the evidence code IEA: those inferred from UniProtKB keywords, UniProt Subcellular Location terms, Enzyme Commission numbers, and InterPro. In a recent report, I showed these electronic annotations are highly reliable, particularly in the prokaryote Escherichia coli [132]. To obtain functional annotations, I used various releases of the UniProt-GOA database, all of them freely available at the UniProt-GOA FTP site.

I presented this phylogenetic profile to a machine learning algorithm based on decision trees [129], thereby obtaining a model for functional an-
notation, as well as the estimation of predictive accuracy for each GO term. More details on the algorithm and our implementation of phylogenetic profiling are available in the original publication [115].

5.3.2 Genome data

For the patterns of presence or absence, I used the data available in the December 2012 release of the OMA database. This release contained 1078 Bacteria, 107 Archaea, and 135 Eukaryotes. Each of the subsets of genomes in our work is derived from this set of genomes.

In this work, I focused on Bacteria because this is by far the most numerous kingdom in the OMA database. Moreover, eukaryotic phylogenetic profiles showed limited success, both in the literature [61, 117, 105, 109], as well as in our own experiments (Figure C.4). Interestingly, [61] showed best average performance for the dataset containing a few selected eukaryotes, while the performance deteriorated when their datasets included all the eukaryotes available at the time: they note that the over-representation of parasitic eukaryotes and vertebrates made eukaryotes less useful for phylogenetic profiling.

5.3.3 Species selection

I considered four strategies to select genome subsets: random selection, selection to preserve the maximum diversity among genomes, selecting genomes that belong to a particular clade, and previous strategies from the literature.

When selecting for maximum diversity, I iterated through the distance matrix used to construct the species phylogenetic tree in the OMA database: I searched for the two closest species in the matrix and excluded the one with fewer GO annotations, repeating this process until I was left with the desired number of genomes. For these experiments that evaluate the influence of the number and diversity of genomes on phylogenetic profiling, I only changed the number of genomes in phylogenetic profiles, keeping as the annotations of phylogenetic profiles those assigned when I used all the available data.

I evaluated the predictive accuracy of our implementation of phylogenetic profiling on two published subsets of genomes. First, a manually curated subset of genomes by Jothi et al. [61]: I created a subset of strains that closely matches that named BAE3a, for which they report best overall performance. The December 2012 OMA database release does not contain three strains of *Chlamydomphila pneumoniae*: AR39, CWL029, and J138. They were replaced by the strain available in the OMA database: LPCoLN. In addition, the OMA database does not contain three of the strains listed in the BAE3a set; I replaced those with closely related strains. The final dataset had 58 genomes: their selection very closely resembled the BAE3a set.

The second published set was derived from Simonsen et al. [114]: Martin Simonsen kindly provided the list of 100 genomes that showed best predictive accuracy for *E. coli* phylogenetic profiling. This set had been automatically obtained using their method of Tree-Based Search. Out of
these 100, we did not have 4 strains in the OMA database: *Xenorhabdus nematophila* (strain ATCC 19061 / DSM 3370 / LMG 1036 / NCIB 9965 / AN6), *Kangiella koreensis* (strain DSM 16069 / KCTC 12182 / SW-125), *Escherichia fergusonii* (strain ATCC 35469 / DSM 13698 / CDC 0568-73), and *Tolumonas auensis* (strain DSM 9187 / TA4). I omitted these, thus obtaining a dataset with 96 genomes.

### 5.3.4 Evaluating predictive accuracy

To infer our functional annotation models, I used the implementation of the Random Forest algorithm [20] in the CLUS software [107]. This software is a Predictive Clustering System that builds decision trees able to deal with multiple hierarchically organized class labels, such as terms from the Gene Ontology, for Hierarchical Multi-label Classification (HMC). The Random Forest Ensemble approach used—combining multiple decision trees—obtains better predictive performance than could be obtained from individual decision trees [129].

The first important feature of the Random Forest Ensemble approach is that individual decision trees are constructed by searching over a random subset of the available decisions—in our case presence/absence patterns of OMA members in genomes—when splitting a node. These individual trees are combined to produce the final result. Because of this combination, Random Forest reduces the influence of redundant decision points, patterns of presence or absence, that would influence methods relying on distance metrics [83]. Consequently, when testing whether adding new genomes is beneficial for phylogenetic profiling, the addition of possibly redundant decision points (e.g., phylogenetically close genomes) will not have a detrimental effect, while the addition of possibly useful decision points (e.g., phylogenetically distant genomes) will increase predictive accuracy (Figure 5.5).

The second important feature of the Random Forest Ensemble approach is that each decision tree is constructed from a different sample of the training dataset: the bagging (Bootstrap aggregating) methodology samples with replacement [19]. In this way, about 63% of unique samples are selected for training the decision tree and produce the functional annotation model—the decision tree that can infer GO terms assigned to an OMA group based on the presence/absence of OMA members in the genomes.

The procedure above allowed us to calculate the True Positives, False Positives, and False Negatives, thereby obtaining the corresponding AUPRC score. For each experiment, I repeated the procedure 500 times, averaging the probabilities of this ensemble of decision trees, thereby obtaining the final estimates of Precision, Recall, and the AUPRC.

### 5.3.5 Analysed data

The file that contains the AUPRC values obtained for each GO term in each of the experiments performed for this work can be downloaded from http://lab.dessimoz.org/14.phylprof. On the same web page, I also provide a list of all the 1093 GO terms that formed the backbone of our analysis. The CLUS algorithm I used in my work is freely available from
5.3. Methods

Figure 5.5: Performance of phylogenetic profiling was not affected when I used many strains of the same organism. I used 31 strains of *Escherichia coli* and I added to this set: A) 31 random organisms, B) 62 random organisms, C) 93 random organisms, and D) 124 random organisms. Each plot in a panel corresponds either to the combination of the 31 *E. coli* strains and the randomly selected organisms (left) or just the randomly selected organisms (right). Each boxplot summarizes AUPRC scores for GO terms in the dataset indicated on the x-axis. Lower, mid, and upper horizontal lines denote the first quartile, median and the third quartile, respectively; vertical lines reach 1.5 interquartile range from the respective quartile or the extreme value, whichever is closer. Each plot summarizes the results for ten independent random organism selections.

Chapter 6
Conclusions and Outlook

Evaluating the quality of different function prediction methods is hindered by the incompleteness of the annotation databases, a notion referred to as the Open World Assumption (OWA). In my thesis, I first discussed OWA’s effect on the standard measures that evaluate function prediction methods. Second, I exploited the changes in successive releases of the UniProt Gene Ontology Annotation database, and thereby assessed the quality of computational function predictions in terms of their Specificity, Reliability, and Coverage. Third, I presented a novel method for computational function prediction based on phylogenetic profiling, therein demonstrating the usefulness of both orthologs and paralogs. Additionally, experiments showed that my method provides realistic estimates of predictive accuracy. Finally, I assessed the influence of more data, as opposed to more methodological improvements, on predictive accuracy of an established method for function prediction, phylogenetic profiling.

Despite notable advances in evaluating the quality of function prediction [95, 134], OWA remains an important obstacle when trying to fairly evaluate function prediction algorithms [32, 49]. Indeed, all efforts to evaluate function on a larger scale, i.e., evaluating multiple functions in multiple organisms, rely on the observed data that can be found, or will be found in a few months’ time, in the annotation databases. Unfortunately, the desired solution of performing large(r)-scale experiments to obtain comprehensively annotated proteins across a range of model organisms is infeasible due to practical considerations.

New approaches to resolve the Open World of protein function prediction could include assigning negative annotations—providing evidence that a protein does not perform a function [142]; in the ideal case where we have comprehensive negative annotations, there is no practical difference between the Open and the Closed world. A wisely chosen set of trusted negative annotations might provide an additional angle on the evaluations of quality of function prediction.

It is only if the advances in the development of new algorithms are coupled with increasingly robust evaluations of these algorithms that computational methods for function prediction will continue to contribute to generating useful biological hypotheses. Not only will such work contribute to closing the gap between the Open and the Closed world, but it will increase our understanding of the increasing pool of new sequences.
Appendix A

Supplementary Information for the analysis of quality of computationally inferred Gene Ontology annotations

Figure A.1: Distribution of experimental, curated, and electronic annotations among the 12 model and 241,860 non-model organisms in the 11-01-2011 UniProt-GOA release. The numbers account for annotations available in the UniProt-GOA release, i.e., we do not report the annotations available through the definition of the ontology. The 12 model organisms are *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Gallus gallus*, *Danio rerio*, *Dictyostelium discoideum*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, and *Escherichia coli* K-12. Out of non-model organisms, 838 have at least one experimental annotation, while the rest have no experimental annotations.
A. Quality of electronic Gene Ontology annotations

Inferred from Sequence or Structural Similarity
Inferred from Reviewed Computational Analysis
Inferred from Traceable Author Statement
Inferred from Non-traceable Author Statement
Inferred by Curator
Experimental

Inferred from UniProt Subcellular Location terms
Inferred from UniProtKB Keywords
Inferred from Ensembl Compara (H. sapiens, R. norvegicus, M. musculus, G. gallus)
Inferred from HAMAP2GO (E. coli)
Inferred from InterPro
Inferred from Enzyme Commission

B
Figure A.2 (preceding page): Curated (A) and electronic (B) annotations we analyse for the 12 model organisms. Each panel summarizes the data for the model organism denoted in the header; the color denotes the evidence or the reference code, and the y axis denotes the number of annotations analysed for the period denoted on the x axis.
Figure A.3: The quality of the 16-01-2008 UniProt-GOA release, evaluated using the 11-01-2011 UniProt-GOA release; comparison on the subset of the cellular component GO terms used in electronic annotations inferred from UniProt subcellular mappings. The 12 model organisms are *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Gallus gallus*, *Danio rerio*, *Dictyostelium discoideum*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, and *Escherichia coli* K-12. The area of the disc reflects the frequency of the GO term in the 16-01-2008 UniProt-GOA release. To be visualized in this plot, a GO term needs to have assigned at least 10 electronic annotations in the 16-01-2008 UniProt-GOA release and at least 10 experimental annotations in the 11-01-2011 UniProt-GOA release. The green lines correspond to the mean values for the respective axes.
Figure A.4: The quality of the 16-01-2008 UniProt-GOA release, evaluated using the 11-01-2011 UniProt-GOA release; comparison on the subset of molecular function GO terms used in electronic annotations inferred from Enzyme Commission mappings. The 12 model organisms are Homo sapiens, Mus musculus, Rattus norvegicus, Caenorhabditis elegans, Drosophila melanogaster, Arabidopsis thaliana, Gallus gallus, Danio rerio, Dictyostelium discoideum, Saccharomyces cerevisiae, Schizosaccharomyces pombe, and Escherichia coli K-12. The area of the disc reflects the frequency of the GO term in the 16-01-2008 UniProt-GOA release. To be visualized in this plot, a GO term needs to have assigned at least 10 electronic annotations in the 16-01-2008 UniProt-GOA release and at least 10 experimental annotations in the 11-01-2011 UniProt-GOA release. The orange lines correspond to the mean values for the respective axes.

Figure A.5: Distribution of experimental and electronic annotations among the 12 model and 241 860 non-model organisms in the 11 01 2011 GOA release. The numbers account for annotations available in the GOA release, i.e. we do not report the annotations available through the definition of the ontology. The 12 model organisms are Homo sapiens, Mus musculus, Rattus norvegicus, Caenorhabditis elegans, Drosophila melanogaster, Arabidopsis thaliana, Gallus gallus, Danio rerio, Dictyostelium discoideum, Saccharomyces cerevisiae, Schizosaccharomyces pombe, and Escherichia coli K-12.
Figure A.6: The quality of the 16-01-2008 UniProt-GOA release, evaluated by the 11-01-2011 UniProt-GOA release; each model organism is evaluated separately. The ontology is denoted by the color of the disc, while the area of the disc reflects the frequency of the GO term in the 16-01-2008 UniProt-GOA release. To be visualized in this plot, a GO term needs to have assigned at least 10 electronic annotations in the 16-01-2008 UniProt-GOA release and at least 10 experimental annotations in the 11-01-2011 GOA release. The coloured lines correspond to the mean values for the respective axes.
Figure A.7: Distribution of annotations from the three sources—experimental, curated and electronic—among the 12 reference genomes. Numbers denote the annotations available for the respective bar; these include annotations available in the 11-01-2011 UniProt-GOA release, as well as annotations available through the definition of the ontology (i.e., all parent GO terms).
Figure A.8: Counts of confirmed, rejected and newly available experimental annotations we analyze for the 12 model organisms. Each panel summarizes the data for the model organism denoted in the header; the color denotes the group of evidence codes and the y axis denotes the number of annotations counted in the period denoted on the x axis.
Figure A.9: Counts of confirmed and rejected annotations for the analyzed evidence or reference codes. The y axis denotes the number of confirmed (green) or rejected/removed (red) annotations counted in the period denoted on the x axis.
A. Quality of electronic Gene Ontology annotations

Figure A.10: Counts of confirmed and rejected annotations for the analyzed evidence or reference codes. The y axis denotes the number of confirmed (green) or rejected/removed (red) annotations counted in the period denoted on the x axis.

Figure A.11: The span of curated annotation updates in the UniProt-GOA releases. Each panel summarizes the data for the model organism denoted in the header. The y-axis denotes the number of gene products with the longest annotation update interval denoted on the x-axis. The longest update interval is the time between the first and the last annotation assigned to one gene product.
Figure A.12: Summary statistics of reliability; each reference code is evaluated separately. The 12 model organisms are *Homo sapiens*, *Mus musculus*, *Rattus norvegicus*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Arabidopsis thaliana*, *Gallus gallus*, *Danio rerio*, *Dictyostelium discoideum*, *Saccharomyces cerevisiae*, *Schizosaccharomyces pombe*, and *Escherichia coli* K-12. Each boxplot summarizes the evaluation period indicated on the x axis: reliability of the 13-01-2006 GOA release, evaluated by the 21-01-2009 UniProt-GOA release; reliability of the 19-01-2007 UniProt-GOA release, evaluated by the 22-01-2010 UniProt-GOA release; reliability of the 16-01-2008 GOA release, evaluated by the 11-01-2011 UniProt-GOA release. Lower, mid, and upper horizontal lines denote the first quartile, median and the third quartile, respectively, while the black dot denotes the mean value. Values further than 1.5 interquartile range from the respective quartile are depicted by black points. An asterisk (*) below the boxplot denotes a significant difference of the median with respect to the previous interval, at a confidence level of 0.05 (Mann-Whitney U test, two-tailed).
**A. Quality of electronic Gene Ontology annotations**

Figure A.13: Quality of electronic annotations and annotations Inferred by Sequence Similarity (ISS) on a common set of GO terms. Electronic sources are inferred from InterPro, UniProt-GOA, UniProt-GOA (subcellular), Enzyme Commission, Ensembl Compara, and HAMAP. Quality of the 16-01-2008 GOA release is evaluated by the 11-01-2011 GOA release; coverage is on the x-axis and reliability is on the y-axis. The ontology is denoted by the color of the disc, while the area of the disc reflects the frequency of the GO term in the 16-01-2008 GOA release. To be visualized in the plot, a GO term needs to have assigned at least 10 electronic/curated annotations in the 16-01-2008 GOA release, and at least 10 experimental annotations in the 11-01-2011 GOA release. The coloured lines correspond to the mean values for the respective axes.
Appendix B

Supplementary Information for the analysis of the influence of orthologs and paralogs on phylogenetic profiling
Figure B.1: Examples of Precision-Recall curves for three Gene Ontology (GO) terms for the four annotation models. Each panel shows the Precision-Recall curve for the GO term denoted in the top title panel, obtained in the annotation model denoted in the right title panel. The x-axis represents Recall, and the y-axis represents Precision. The plots were created by increasing the cut-off for the model’s annotation probability: each point in the plot represents the cumulative Precision and Recall at each step of varying the cut-off.
Figure B.2: A comparison of Area Under the Precision-Recall Curve (AUPRC) and Information Content (IC) for the functional annotation model that includes OMA cliques of orthologs, OMA inferred orthologs, and OMA inferred paralogs. Each point represents one Gene Ontology term and its color denotes the ontology. The IC is calculated as the negative logarithm of the frequency of the GO term in the 07-02-2012 UniProtGOA release.

Figure B.3: Molecular Function (MF) annotations that the model including both orthologs and paralogs assigned to *E. coli* genes at Precision 90%. Apart from the most general terms in the MF ontology, we highlight some more specific annotations.
B. Phylogenetic profiling with orthologs and paralogs

Figure B.4: Cellular Component (CC) annotations that the model including both orthologs and paralogs assigned to *E. coli* genes at Precision 90%. Apart from the most general terms in the CC ontology, we highlight some more specific annotations.

Figure B.4: Cellular Component (CC) annotations that the model including both orthologs and paralogs assigned to *E. coli* genes at Precision 90%. Apart from the most general terms in the CC ontology, we highlight some more specific annotations.
B. Phylogenetic profiling with orthologs and paralogs

Figure B.5 *(preceding page)*: Predictive performance of the annotation models that account only for the evolutionary distance between refined homologs inferred by the OMA algorithm, for the three Gene Ontologies: A) Biological Process, B) Cellular Component, and C) Molecular Function. The x-axis represents the models: each model includes refined homologs that are closer than the denoted PAM distance. These homologs and their corresponding pairwise evolutionary distances were inferred by the OMA algorithm [104] all-against-all local sequence alignments were refined using two empirical criteria: first, only homologs with an E-value of roughly 10-14 were considered significant and second, only alignments where the shorter sequence is at least 61% length of the longer sequence are considered. The y-axis represents the Area Under the Precision Recall Curve (AUPRC). Each disc represents one GO term; its colour represents the ontology, while the area of the disc is proportional to the frequency of the GO term among all annotations available in 07-02-2012 UniProt-GOA release. Each boxplot summarizes AUPRC for the dataset indicated on the x-axis. Lower, mid, and upper horizontal lines denote the first quartile, median and the third quartile, respectively; vertical lines reach 1.5 interquartile range from the respective quartile or the extreme value, whichever is closer.
Figure B.6: The quality of the 21-01-2009 UniProt-GOA release, evaluated by the 07-02-2012 UniProt-GOA release for electronic annotations assigned to *Escherichia coli*. A scatterplot of coverage compared to the reliability for the GO terms of the three ontologies: Biological Process, Cellular Component, and Molecular Function. The area of the disc reflects the frequency of the GO term in the 16 01 2008 UniProt-GOA release. The coloured lines correspond to the mean values for the respective axes. To be visualized in this plot, a GO term needs to have assigned at least 10 electronic annotations in the 21-01-2009 UniProt-GOA release and at least 10 experimental annotations in the 07-02-2012 UniProt-GOA release. The methodology used to obtain the data presented in this figure is described in [132]. Note that this figure shows the analysis with more recent data than what is published by Škunca *et al.*
B. Phylogenetic profiling with orthologs and paralogs

A. Area Under the Receiver Operating Characteristic Curve (10-fold cross-validation error estimate)

B. Area Under the Receiver Operating Characteristic Curve (Out Of Bag error estimate)

Generality of a GO term
Figure B.7 (preceding page): Predictive performance of the annotation models compared with the Area Under the Receiver Operating Characteristic Curve (AUC) for the Cellular Component ontology. A) Annotation model inferred using the kNN classifier: each panel shows the data for the k value denoted in the panel header. B) Annotation model built using the Clus-HMC classifier. The x axis represents the data used to infer the models: phylogenetic profiles are based on (C) OMA cliques of orthologs and OMA inferred orthologs and (D) OMA cliques of orthologs, OMA inferred orthologs, and OMA inferred paralogs. The y-axis represents the AUC. Each disc represents one GO term; its colour represents the ontology, while the area of the disc is proportional to the generality of the GO term: the frequency of the GO term among all annotations available in 07-02-2012 UniProt-GOA release. Each box-plot summarizes AUC for the dataset indicated on the x-axis. Lower, mid, and upper horizontal lines denote the first quartile, median and the third quartile, respectively; vertical lines reach 1.5 interquartile range from the respective quartile or the extreme value, whichever is closer.
Appendix C

Supplementary Information for the analysis of the influence of more data on phylogenetic profiling
More data in phylogenetic profiling

Figure C.1: Performance of phylogenetic profiling, measured in AUPRC, when we reduce the number of annotations used for phylogenetic profiling. For each of the experiments denoted on the x-axis, we only used a fraction of the available annotations in the most recent dataset. Dashed and full lines connect the dots of the mean AUPRC scores for two sets of experiments: random sub-selection of genomes (full lines) and sub-selection to keep maximum diversity among the selected genomes (dashed lines). Colour denotes the number of genomes used in the phylogenetic profiles. For these experiments, we evaluated predictive accuracy for the 777 GO terms that were assigned to at least 100 phylogenetic profiles.
Proteobacteria, 538
Firmicutes, 218
Cyanobacteria, 39
Actinobacteria, 106
Bacteroidetes, 35
Figure C.2 (preceding page): Performance of phylogenetic profiling, measured in AUPRC, on five major groups of sequenced bacteria, available in the OMA database Dec 2012 release. A) Proteobacteria (538 organisms), B) Firmicutes (218 organisms), C) Actinobacteria (106 organisms), D) Bacteroidetes (35 organisms), and E) Cyanobacteria (39 organisms). Each plot in a panel corresponds either to the group of bacteria (left), a subset of organisms with the maximum phylogenetic diversity, having the same number as the number of organisms in the group of bacteria (middle), or a random subset of organisms, having the same number as the number of organisms in the group of bacteria (right). The area of each violin plot summarizes the distribution of GO terms according to the AUPRC value: the area of the plot corresponds to the probability density of GO terms at different values of AUPRC. The black dot denotes the mean value of AUPRC for the respective dataset.

Figure C.3: Performance of phylogenetic profiling, measured in AUPRC, for two triplets of datasets with the same number of organisms, but of different composition. A) Manually assembled dataset shown to have the best average score among the datasets tested by Jothi et al., 2007. For the left panel, the manually selected 58 organisms, predominantly bacteria; for the middle panel, randomly selected 58 bacteria from our pool of 1078 bacteria; for the right panel, 58 bacteria selected to have the highest phylogenetic diversity in the set. B) Automatically selected set, shown to be best among those examined by Simonsen et al., 2012. For the left panel, the 96 automatically selected genomes; for the middle panel, randomly selected 96 bacteria from our pool of 1078 bacteria; for the right panel, 96 bacteria selected to have the highest phylogenetic diversity in the set.
Figure C.4: Performance of phylogenetic profiling on three different kingdoms: eukaryotes, bacteria, and archaea. Each boxplot summarizes Area Under the Precision-Recall Curve (AUPRC) scores for the dataset indicated on the x-axis. Lower, mid, and upper horizontal lines denote the first quartile, median and the third quartile, respectively; vertical lines reach 1.5 interquartile range from the respective quartile or the extreme value, whichever is closer.
Bibliography


[79] Elaine R Mardis. The 1,000 genome, the 100,000 analysis? Genome Med., 2(11):84, 26 November 2010.


