Computational prediction of gene function under the Open World Assumption

Author(s):
Škunca, Nives

Publication Date:
2014

Permanent Link:

https://doi.org/10.3929/ethz-a-010384172 →

Diss. ETH No. 22368

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

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 erste Voraussetzung die erfl̈lt sein muss damit computerberechnete Funktionsvorhersagen nützlich sind ist eine Bestimmung deren Qualität. Normalerweise würden wir dazu die vorhergesagten Funktionen mit einem Teil der verfügbaren Annotationen aus Datenbanken vergleichen. Dieses Vorgehen wird jedoch dadurch verunmöglicht, dass die Datenbanken unvollständig sind: Unsere Datenbanken geben nur einen Teil der Information aus der Literatur wieder. Zudem wird die Literatur laufend durch neue Information erweitert und verbessert. Aus diesen Gründen wird klar, dass das Nichtvorhandensein einer Funktionsannotation in einer Datenbank nicht gleichbedeutend ist mit einer falschen Vorhersage. Dieses Faktum wird in der Literatur üblicherweise als *Open World Assumption (OWA)* bezeichnet.

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.

Im zweiten Beitrag meiner Doktorarbeit vermeide ich die Probleme der Closed World Assumption (CWA) durch eine experimentelle Validierung der Vorhersagen: Als erstes entwickle ich eine Methode zur rechengestützten Vorhersage der Funktion eines Gens basierend auf phylogenetischen Profilen, die sowohl Orthologe wie auch Paraloge — Homologe, die nach einer Speziation bzw. einer Gen-Duplikation divergierten — einbezieht. Ich zeige, dass das auf phylogenetischen Profilen basierte Modell, welches sowohl Orthologe als auch Paraloge einbezieht, mehr Annotationen bei gleicher mittlerer Genauigkeit liefert als das Modell, welches ausschliesslich Orthologe verwendet. Zweitens zeige ich anhand einer experimentellen Untersuchung am Modellorganismus *Escherichia coli*, dass meine Methode zur Funktionsvorhersage eine realistische Abschätzung der Genauigkeit von auf Vorhersagen beruhenden Annotationen bietet: Ein Screening der Veränderung des Phänotyps beim Zellwachstum in Knockout-Mutanten von

E. coli ergab eine Genauigkeit von 66 Prozent—von 38 getesteten Vorhersagen wurden 25 bestätigt—was 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.

Die elektronischen Annotationen sind generell vertrauenswürdiger als ich es erwartet hätte. Wenn andere Belege zur Berechnung hinzugezogen werden als nur Experimente, welche in der Primärliteratur erwähnt sind, dann ist die Qualität dieser Annotationen sogar vergleichbar mit denen der menschlichen Kuratoren. Ich konnte aber auch signifikante Unterschiede zwischen verschiedenen Methoden der maschinellen Deduktion, verschiedenen Typen von Annotationen und verschiedenen Organismen feststellen. Diese Arbeit soll eine Anleitung für Benutzer der GO darstellen und legt die Basis für eine strukturierte Verbesserung der Algorithmen zur Deduktion von GO-Funktionen. Der Erfolg der algorithmischen Deduktion biologischer Funktion kann zurückgeführt werden auf eine Kombination von methodologischen Verbesserungen, der Erweiterung der Anzahl sequenzierter Genome, und der wachsenden Zahl verfügbarer Annotation von biologischen Funktionen. Während die meiste bisherige Literatur sich auf die methodologischen Verbesserungen beschränkt hat, wissen wir relativ wenig über den Beitrag der erhöhten Anzahl Genome und Annotationen von biologischen Funktionen.

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.

Meine Arbeit fokussiert auf numerische Methoden, aus welchen sich Genfunktionen ableiten lassen. Zusätzlich zur Präsentation einer eigenen Methode zur Vorhersage numerischer Funktionen führt meine Arbeit eine neue Metrik ein, die der Qualitätsbewertung von Vorhersagen einer anerkannten Datenbank dient. Quantifiziert wird sowohl der Beitrag der wachsenden Menge der sequenzierten Genome, als auch die neuen Annotationen. Ein dauerhafter Betrachtungsschwerpunkt wurde dabei auf das Problem der sogenannten Open World Assumption gelegt, welches postuliert, dass die Datenbanken von funktionalen Annotationen grundsätzlich unvollständig sind.