Automatic Discovery of Hidden Associations Using Vector Similarity : Application to Biological Annotation Prediction

Seyed Ziaeddin Alborzi

UNIVERSITÉ DE LORRAINE

École doctorale IAEM Lorraine

Automatic Discovery of Hidden Associations Using Vector Similarity: Application to Biological Annotation Prediction

THÈSE

présentée et soutenue publiquement le 23 Février 2018

pour l'obtention du

Doctorat de l'Université de Lorraine

(mention informatique)

par

Seyed Ziaeddin ALBORZI

Composition du jury

Rapporteurs :	Wim Vranken	PR Vrije Universiteit Brussel, Brussels
	Graham Kemp	PR Chalmers University of Technology, Gothenburg
Examinateurs :	Olivier Poch	DR CNRS, Strasbourg
	Alessandra Carbone	PR Sorbonne Université, Paris
	Anne Boyer	PR Université de Lorraine, Nancy
	Malika Smaïl-Tabbone	MC Université de Lorraine, Nancy
Encadrants :	Marie-Dominique Devignes	CR CNRS, Nancy
	David W. Ritchie	DR INRIA, Nancy

Équipe CAPSID – INRIA Nancy Grand Est

Laboratoire Lorrain de Recherche en Informatique et ses Applications (LORIA) — UMR 7503

Mis en page avec la classe thesul.

Remerciements

I would like to thank all the people who contributed in some way to the work described in this thesis. First and foremost, I thank my academic advisors, Dr. Dave Ritchie and Dr. Marie-Dominique Devignes, for accepting me into CAPSID team at Institut national de recherche en informatique et en automatique (INRIA). During my tenure, they contributed to a rewarding doctoral school experience by giving me intellectual freedom in my work, supporting my attendance at various conferences, engaging me in new ideas, and demanding a high quality of work in all my endeavors. Additionally, I would like to thank my committee members Dr. Carbone, Dr. Poch, Dr. Vranken, Dr. Kemp, Dr. Boyer, and Dr. Smaïl-Tabbone for their interest in my work. I was fortunate to have the chance to work in UniProt team at European Bioinformatics Institute (EBI). I worked (nonstop) as a predoctoral visitor in the team of Dr. Maria Martin of EBI for three months starting from October 2016. I am very grateful to Dr. Rabie Saidi and Mr. Alex Reneaux who helped me in creating rules for protein functional annotation I present in this thesis. I am grateful for the funding sources that allowed me to pursue my doctoral studies: ANR Fellowship, Region of Lorraine and INRIA. I would like to acknowledge the Doctoral School at University of Lorraine. My graduate experience benefited greatly from the courses I took. I would like to acknowledge my friends, Farhad, Saeid, Soheib, Mojdeh, Valia, Daishi, Iordan, Ehsan, Mohanna, Meysam, Younes and family members, Baba, Maman, Emad, Hesam, Carol, Baba Mapar, Maman Mapar, Mahsa, Mahna, Payam, Samira, Parinaz who supported me during my time here. Finally, I would like to thank my lovely wife, Mahta for her constant love and support.

This thesis is dedicated to
My lovely wife who always believed in me,
My wonderful parents who have raised me to be the person I am today.

Contents

List of Figures

List of Tables

Introduction en français

Objectifs de la thèse

Au cours des dernières années, les recherches dans divers domaines de la biologie ont produit d'énormes quantités de données biologiques. L' interprétation de ces grands volumes de données nécessite de mettre en place des processus de traitement et d' analyse computationnelle complexes. L'un des moyens les plus intéressants et les plus efficaces d'inférer des principes à partir d'ensembles de données biologiques est l'utilisation de l'exploration de données pour trouver une solution aux problèmes biologiques. Le volume des données biologiques est en pleine croissance, il est donc significatif que les applications d'exploration de données évoluent progressivement et se développent comme un domaine de recherche actif au sein de la bioinformatique.

L'exploration de données est un concept général qui regroupe diverses méthodes d'extraction d'informations à partir de grands ensembles de données dans le but d'apprendre des modèles. Les techniques d'exploration de données impliquent l'utilisation des méthodes d'apprentissage automatique, de systèmes de bases de données, d'intelligence artificielle, de statistiques et de visualisation [Li et al., 2013]. Les approches d'exploration de données sont exploitées dans plusieurs domaines de recherche et industries pour fournir des modèles de données : c'est ce qu'on appelle de nos jours la science des données, voire l'intelligence des données (« data science, data intelligence »). Ceci était déjà connu depuis les années 90 comme la découverte des connaissances dans les bases de données (KDD) ou l'analyse intelligente des données (IDA) [Raza, 2012].

Le processus d'exploration de données permet aux chercheurs d'améliorer leur compréhension des mécanismes biologiques afin de trouver et d'introduire des traitements modernes dans les soins de santé et de découvrir de nouvelles connaissances sur les mécanismes de la vie. Au cours des dernières années, l'analyse computationnelle, les découvertes et les prédictions fondées sur de nouveaux modèles et hypothèses biologiques ont énormément augmenté [Fogel, 2008].

Deux exemples remarquables d'exploration de données dans le domaine biologique sont la prédiction des fonctions des protéines et la prédiction des interactions protéine-protéine. Les protéines sont des macromolécules qui remplissent la plupart des fonctions biologiques dans les organismes vivants. Au niveau moléculaire, les fonctions protéiques sont souvent réalisées par des régions structurales des protéines, hautement conservées, identifiées à partir d'alignements de séquences ou de structures, qui peuvent être classées en familles de domaines. Comme de nombreux domaines protéiques se replient en structures tridimensionnelles (3D) caractéristiques, il existe souvent une relation étroite entre la structure protéique et la fonction protéique [Berg et al., 2002]. Actuellement, la base de données Pfam est l'une des classifications basées sur les séquences les plus largement utilisées pour les familles de domaines [Finn et al., 2016b]. Les bases de données CATH [Orengo et al., 1997] et SCOP [Murzin et al., 1995] sont deux exemples de classifications de domaines basées sur les structures.

En plus des classifications basées sur la séquence et sur la structure, les protéines peuvent également être classées en fonction de leur fonction. Par exemple, l'Ontologie des Gènes ou "Gene Ontology" (GO) [Ashburner et al., 2000] consiste en un vocabulaire contrôlé de termes qui décrivent la fonction des produits des gènes dans une cellule. La Commission des Enzymes (EC) a proposé un autre schéma de classification particulier pour les enzymes [Webb et al., 1992]. A priori, les systèmes de classification des fonctions sont conçus et utilisés pour décrire les fonctions des protéines entières. Au niveau des domaines protéiques, un pourcentage très limité de domaines bénéficie d'une annotation GO manuelle. Récemment, un travail intéressant publié sous le nom de dcGO [Fang and Gough, 2013] a tenté de dériver, à partir des annotations de protéines entières, des annotations fonctionnelles (telles que GO) pour la plupart des domaines protéiques. Néanmoins, nous avons constaté qu'il existe plusieurs associations GO-Pfam organisées par InterPro [Finn et al., 2016a], qui ne sont pas présentes dans dcGO. Selon l'analyse [Alborzi et al., 2017b], on estime que les associations dcGO ne peuvent annoter que 43

Plus généralement, il y a des millions de séquences de protéines dans UniProtKB/TrEMBL [Apweiler et al., 2017] qui manquent actuellement d'annotations GO. Or, il existe seulement un nombre relativement limité de familles distinctes de domaines protéiques, qui sont réutilisés et combinés de différentes manières dans différentes protéines. En effet, comparées au grand nombre de séquences différentes qui existent, les classifications de domaines actuelles contiennent de l'ordre de seulement 15 000 familles de domaines protéiques distincts. Par conséquent, il est naturel de supposer que si des annotations de structures et de séquences protéiques connues pouvaient être attribuées à des termes GO (ou EC) au niveau du domaine, beaucoup de ces annotations pourraient être transférées à un très grand nombre de protéines non annotées. Cependant, associer des termes GO aux domaines protéiques est un problème non trivial car, à l'exception des protéines à domaine unique où la cartographie est évidente, de nombreuses relations peuvent se produire entre les domaines et les fonctions. Ce manque d'annotations et la complexité du problème nous intéressent pour cibler le problème de l'annotation des domaines protéiques.

En effet, dans quelle mesure les domaines protéiques annotés peuvent-ils être utilisés pour annoter fonctionnellement des protéines entières ? L'annotation fonctionnelle de protéines entières est d'une importance cruciale pour une meilleure compréhension des processus biologiques au niveau moléculaire, et a des implications considérables dans la recherche biomédicale et pharmaceutique. Cependant, la caractérisation expérimentale des protéines ne peut pas facilement être réalisée à grande échelle parce que c'est un processus difficile et coûteux [Liolios et al., 2009]. En outre, la vérification de l'annotation des séquences protéiques existantes par des conservateurs experts est presque aussi coûteuse et longue. Ainsi, l'annotation automatique de la fonction des protéines est devenue un problème computationnel critique en bioinformatique [Radivojac et al., 2013]. Au cours de la dernière décennie, plusieurs approches de prédiction de la fonction protéique ont été décrites [Bork et al., 1998, Rost et al., 2003, Watson et al., 2005, Friedberg, 2006, Sharan et al., 2007, Lee et al., 2007, Punta and Ofran, 2008, Rentzsch and Orengo, 2009, Xin and Radivojac, 2011]. La plupart des approches utilisent BLAST [Altschul et al., 1997] pour comparer les séquences de nouvelles protéines avec des protéines dont la fonction a déjà été déterminée expérimentalement, tandis que d'autres appliquent des principes similaires au niveau du domaine.

Ces dernières années, des techniques d'acquisition de données expérimentales à haut débit pour l'analyse génomique, transcriptomique, protéomique et interactomique chez de nombreuses espèces ont ouvert de nouvelles possibilités pour la prédiction automatique de la fonction des protéines. Par exemple, des méthodes utilisant des réseaux d'interaction protéine-protéine peuvent assigner des classes fonctionnelles à des protéines à partir de leurs réseaux d'interactions physiques [Vazquez et al., 2003]. D'autres approches exploitent l'information à partir de combinaisons de domaines protéiques et d'interactions de domaines [Peng et al., 2014]. Les données d'expression génique et d'interaction moléculaire peuvent également être

utilisées pour créer un réseau de gènes fonctionnellement connectés à partir desquels des annotations fonctionnelles peuvent être propagées à travers le réseau [Massjouni et al., 2006], et des informations taxonomiques peuvent être utilisées pour filtrer les fausses prévisions [Zhu et al., 2007]. L'application de l'apprentissage automatique aux relations évolutives entre les produits géniques et les contextes génomiques est un autre moyen d'inférer les annotations fonctionnelles des protéines [Enault et al., 2005, Li et al., 2007]. Des techniques d'apprentissage automatique sont également utilisées pour identifier et extraire des caractéristiques fonctionnelles à partir de protéines représentatives et pour propager des fonctions à des protéines inconnues. Ces méthodes utilisent généralement des techniques probabilistes pour extraire des fonctions des réseaux d'interactions protéiques [Nariai et al., 2007] ou des informations phylogénétiques [Engelhardt et al., 2005]. Une autre approche utilise des techniques d'exploration de règles d'association pour construire des modèles prédictifs basés sur des règles [Boudellioua et al., 2016].

Les informations structurelles sur les protéines peuvent également être utilisées pour faciliter l'annotation des fonctions. Par exemple, dans [Roy et al., 2012], des protéines modèles ayant des repliements et des sites fonctionnels similaires sont créées, et une protéine cible est ensuite comparée à la matrice homologue la plus proche. Parce que les structures tridimensionnelles des protéines sont souvent plus conservées au cours de l'évolution que leurs séquences, l'utilisation de modèles structurels est un moyen précis de trouver des fonctions similaires dans différentes séquences protéiques [Whisstock and Lesk, 2003]. Cependant, les algorithmes basés sur un modèle échoueront si aucun modèle homologue n'est disponible. Les méthodes hybrides peuvent prédire les fonctions protéiques basées sur l'apprentissage et trouver des scores consensuels calculés à partir d'une combinaison de sources de protéines différentes [Hooper et al., 2014] ou d'un mélange de méthodes différentes pour retourner une liste classée d'annotations [You et al., 2017].

Plusieurs méthodes d'annotation fonctionnelle utilisent les familles de domaines protéiques comme unité de base de la similarité protéique [Peng et al., 2014, Forslund and Sonnhammer, 2008]. Néanmoins, malgré la grande variété de techniques d'annotation de fonctions existantes, la prédiction de la fonction des protéines reste un problème ouvert, car il n'existe aucune méthode universelle qui fournisse clairement les meilleures annotations fonctionnelles. En réponse à ce besoin, l'expérience CAFA (Critical Assessment of Protein Function Annotation) [Radivojac et al., 2013] a été lancée pour évaluer l'état actuel de l'art dans l'annotation des fonctions protéiques et encourager les développements dans ce domaine. Cela nous a également motivé à concevoir une approche pour annoter les protéines de manière fonctionnelle.

Par ailleurs, il convient de noter que les protéines exercent rarement leurs fonctions seules. elles coopèrent généralement avec d'autres protéines en construisant un large réseau d'interactions protéine-protéine [Gavin et al., 2002]. Les interactions protéine-protéine sont responsables de la majorité des fonctions cellulaires et l'identification de ces interactions est un moyen de mieux comprendre les divers processus cellulaires et les mécanismes moléculaires des cellules. Grâce aux approches de génomique à haut débit, la quantité de séquences protéiques augmente considérablement tandis que les méthodes expérimentales pour découvrir leurs interactions sont loin derrière. De nombreuses méthodes de calcul ont été proposées pour combler l'écart entre les connaissances sur les séquences de protéines connues et celles sur leurs interactions. Étudier les interactions moléculaires au niveau de la protéine fournit une compréhension intuitive précieuse de la façon dont une molécule joue ses rôles à l'intérieur d'une cellule particulière. Cependant, des compléments d'information essentiels peuvent être apportés par l'analyse des interactions au niveau des domaines protéiques. Il existe un petit nombre de protéines à domaine unique impliquées dans des interactions protéine-protéine, la plupart ont plus d'un domaine [Apic et al., 2001]. Les interactions dans ces protéines multi-domaines impliquent souvent la coopération entre deux ou plusieurs domaines [Bhaskara and Srinivasan, 2011]. Par conséquent, l'identification des interactions protéiques au niveau du domaine est indispensable pour comprendre les détails atomiques

précis dans les interactions protéiques et pour apprendre à prédire de nouvelles interactions.

Au cours des dernières années, les chercheurs se sont concentrés sur l'énumération et la description informatisées des interactions protéiques au niveau des domaines. Une façon de découvrir les interactions domaine-domaine consiste à utiliser des structures tridimensionnelles de protéines. KBDOCK [Ghoorah et al., 2013b], 3did [Stein et al., 2010], iPfam [Finn et al., 2013] et INstruct [Meyer et al., 2013] sont quatre bases de données contenant des informations structurelles sur les interactions domaine-domaine observées, principalement déduites des données de la PDB. La qualité de ces interactions observées est très élevée, mais leur nombre reste limité par la disponibilité d'informations structurelles sur les complexes protéiques. Même si ces méthodes ont fourni des milliers d'interactions de domaine connues, le nombre d'interactions protéine-protéine inférées en utilisant ces interactions domaine-domaine est beaucoup moindre que le nombre réel d'interactions protéine- protéines. Les interactions de domaine déduites des données structurelles en 2010 ne peuvent couvrir qu'environ 5

Contributions

Les contributions de cette thèse concernent plusieurs thèmes de recherche à la fois : découverte des connaissances à partir de données biologiques, annotation fonctionnelle des protéines et interactions protéiques. Dans chacun de ces thèmes, nous avons proposé de nouvelles méthodes et applications.

Dans un premier temps, nous avons proposé une approche de fouille de données appelée CODAC (COmputational Discovery of Direct Associations using Common neighbours) pour découvrir des associations directes entre les fonctions protéiques et les domaines protéiques [Alborzi et al., 2018, Alborzi et al., 2017b, Alborzi et al., 2017c].

Il nous est alors apparu que notre méthode CODAC pour l'annotation fonctionnelle des domaines protéiques pouvait servir de point de départ à l'annotation fonctionnelle automatique de l'ensemble des séquences protéiques. Ceci nous a conduit à développer une extension de CODAC que nous appelons CARDM (Combinatorial Association Rules Domain Miner). CARDM combine l'étape d'apprentissage CODAC, dans laquelle les annotations fonctionnelles sont associées aux domaines protéiques, avec une génération combinatoire de règles et une procédure de filtrage à partir desquelles des modèles prédictifs spécifiques aux taxons sont construits et utilisés pour annoter automatiquement les séquences et structures protéiques.

Nous avons finalement introduit une nouvelle façon de résoudre le problème de la découverte des interactions entre les domaines protéiques. Notre méthode appelée PPIDM est dérivée de notre méthode CODAC précédemment développée et est à notre connaissance la première méthode qui prédit les interactions entre des ensembles de domaines protéiques.

Les méthodes proposées dans cette thèse ne produisent pas de résultats validés comme le ferait la vérification manuelle, mais elles contiennent une phase d'apprentissage à partir de données vérifiées et une combinaison de différentes techniques et bases de données qui les rendent extrêmement puissantes. Les résultats produits peuvent être utilisés par l'expérimentateur pour réduire l'espace de recherche pour trouver des candidats pour certaines associations ou interactions. Des collaborations avec des biologistes sont en cours pour valider les résultats de nos annotations de domaine.

Vue d'ensemble de la thèse

La thèse est organisé comme suit:

Chapitre 1: Comprendre la science des données et le contexte biologique est indispensable. Ainsi, ce chapitre couvre l'essentiel de la science des données, comme la découverte de connaissances et l'exploration

de données, l'extraction de règles d'association, le modèle d'espace vectoriel, les graphes k-partis, les tests d'hypothèses statistiques et le filtrage d'informations. De plus, une introduction générale aux structures et séquences protéiques, aux domaines protéiques, aux fonctions et annotations protéiques, et aux interactions protéine-protéine est donnée dans ce chapitre. Ce chapitre présente également les ressources utilisées dans la thèse.

Chapitre 2: Ce chapitre commence par notre premier problème: attribuer des numéros EC aux domaines protéiques. Notre logiciel pour prédire les associations de domaine EC s'appelle ECDomainMiner.

Chapitre 3: Ce chapitre décrit une approche générale de la découverte computationnelle d'associations entre différents ensembles d'annotations en formalisant le problème sous la forme d'un problème d'enrichissement de graphe biparti dans le cadre d'un graphe triparti.

Chapitre 4: Dans ce chapitre, nous décrivons un nouveau système de prévision de la fonction des protéines (CARDM), qui est utilisé pour l'annotation fonctionnelle des séquences de protéines dans UniProtKB/TrEMBL. En utilisant nos modèles de prédiction générés, notre équipe CAPSID a participé à un défi appelé CAFA dont les résultats sont également expliqués en détail.

Chapitre 5: : Ce chapitre décrit l'approche PPIDM (abréviation de « Protein-Protein Interaction Domain Miner ») pour découvrir par calcul les interactions entre un seul ou des sous-ensembles de domaines protéiques Pfam. PPIDM est dérivé de la méthode CODAC décrite précédemment pour la découverte informatique des associations directes en utilisant des voisins communs.

Chapitre 6: Ce chapitre résume les contributions de cette thèse et présente plusieurs orientations futures à court terme et à long terme.

Annexe A: Cette annexe décrit les serveurs Web ECDomainMiner et GODomainMiner, qui fournissent un accès public aux ressources EC-Pfam et GO-Pfam / CATH / SCOP.

Annexe B: Cette annexe décrit l'intégration des annotations fonctionnelles dans le serveur Web KBDOCK2.

Annexe C: Cette annexe contient des copies des articles publiés et des affiches présentées dans les conférences.

Introduction

Thesis Aims and Objectives

Over recent years, researches in diverse fields of biology have extensively produced huge amount of biological data. Concluding with an interpretation from such big data is in need of complex computational analysis. One of the most interesting and efficacious ways of inferring principles out of biological datasets is usage of data mining to find a solution for biological problems. Biological data are immensely growing, thus, it is significant that the data mining applications progressively evolve in order to maintain it as an active research area within bioinformatics.

Data mining is a general concept that groups various methods of extracting information from large datasets for the purpose of learning patterns and models. Data mining techniques involve usage of machine learning techniques, database systems, artificial intelligence, statistics, and visualisation [Li et al., 2013]. Data mining approaches is exploited in several various research fields and industries to provide data patterns (data intelligence). This is often known as Knowledge Discovery in Databases (KDD) or Intelligent Data Analysis (IDA) [Raza, 2012].

Data mining process allows researchers to enhance their understanding of biological mechanisms in order to find and introduce modern treatments in healthcare and discover new knowledge of life. In the last few years, computational analysis, discoveries and predictions such as new biological patterns and hypothesis, have enormously increased [Fogel, 2008].

Two remarkable examples of data mining in the biological domain are protein function prediction and protein-protein interaction prediction. Proteins are macromolecules which carry out many biological functions in living organisms. At the molecular level, protein functions are often performed by highly conserved structural regions identified from sequence or structure alignments, which may be classified into families of domains. Because many protein domains fold into characteristic three-dimensional (3D) structures, there is often a close relationship between protein structure and protein function [Berg et al., 2002]. Currently, the Pfam database is one of the most widely used sequence-based classifications of protein domains and domain families [Finn et al., 2016b]. The CATH [Orengo et al., 1997] and SCOP [Murzin et al., 1995] databases are two examples of structural domain classifications.

As well as sequence-based and structure-based classifications, proteins may also be classified according to their function. For example, the Gene Ontology (GO) [Ashburner et al., 2000] consists of a controlled vocabulary of GO terms which describe the function of gene products in a cell. The Enzyme Commission number (EC number) is another classification scheme particularly for enzymes [Webb et al., 1992]. However, function classification systems annotate the entire proteins. One interesting exception is the dcGO database [Fang and Gough, 2013] which provides multiple ontological annotations (such as GO) for protein domains. Nonetheless, we found that there are several manually curated GO-Pfam associations from InterPro [Finn et al., 2016a] which are not present in dcGO. According to the analysis

[Alborzi et al., 2017b], it is estimated that dcGO associations can only annotate 43% of the un-annotated structures in the Protein Data bank (PDB) [Gutmanas et al., 2014].

More generally, there are many millions of protein sequences in UniProtKB/TrEMBL [Apweiler et al., 2017] that currently lack GO annotations. On the other hand, only a relatively small number of distinct protein domain families exist, and which are re-used and combined in different ways in different proteins. Indeed, compared to the vast number of different sequences that exist, current domain classifications contain of the order of only 15,000 distinct protein domain families. Therefore, it is natural to suppose that if known protein structure and sequence annotations could be assigned GO terms (or EC numbers) at the domain level, many of these annotations could be transferred to a potentially very large number of unannotated proteins. However, associating GO terms with protein domains is a non-trivial problem because, except for single-domain proteins where the mapping is obvious, many to many relationships can occur between the domains and functions. These lack of annotations and the complexity of the problem interest us to target the problem of annotating protein domains.

Annotating protein domains can be extended to functionally annotate entire proteins. The functional annotation of entire proteins is crucially important for a better understanding of biological processes at the molecular level, and has considerable implications in biomedical and pharmaceutical research. However, the experimental characterization of proteins cannot easily be scaled up because this is a difficult and costly process [Liolios et al., 2009]. Furthermore, the curation and annotation of existing protein sequences by expert curators is almost equally expensive and time-consuming. Thus, the automatic annotation of protein function has become a critical computational problem in bioinformatics [Radivojac et al., 2013]. During the past decade, several protein function prediction approaches have been described [Bork et al., 1998, Rost et al., 2003, Watson et al., 2005, Friedberg, 2006, Sharan et al., 2007, Lee et al., 2007, Punta and Ofran, 2008, Rentzsch and Orengo, 2009, Xin and Radivojac, 2011]. Most approaches use BLAST [Altschul et al., 1997] to compare the sequences of new proteins with proteins whose function have previously been determined experimentally, while some others apply similar principles at the domain level.

In recent years, high-throughput experimental data acquisition techniques for genomic, transcriptomic, proteomic, interactomic analysis in many species has opened new possibilities for automatic protein function prediction. For instance, methods using protein-protein interaction networks may assign functional classes to proteins from their physical interaction networks [Vazquez et al., 2003]. Other approaches exploit information from combinations of protein domains and domain interactions [Peng et al., 2014]. Gene expression and molecular interaction data may also be used to create a network of functionally connected genes from which functional annotation may be propagated across the network [Massjouni et al., 2006], and taxonomy information may be used to filter false predictions [Zhu et al., 2007]. Applying machine learning to evolutionary relationships between gene products and genomic contexts is another way to infer protein function annotations [Enault et al., 2005, Li et al., 2007]. Machine learning techniques are also used to identify and extract functional features from representative proteins, and to propagate functions to unknown proteins. Such methods typically use probabilistic techniques to extract functions from protein interaction networks [Nariai et al., 2007] or phylogenetic information [Engelhardt et al., 2005]. Other approach uses association rule mining techniques to construct rule-based predictive models [Boudellioua et al., 2016].

Protein structural information can also be used to aid function annotation. For example, in [Roy et al., 2012] template proteins having similar folds and functional sites are created, and a target protein is then compared to the closest homologous template. Because the three-dimensional structures of proteins are often more evolutionary conserved than their sequences, using structural templates is an accurate way to find similar functions in different protein sequences [Whisstock and Lesk, 2003]. However, template-based

algorithms will fail if no homologous template is available. Hybrid methods can predict protein functions based on learning and finding consensus scores computed from a combination of different protein sources [Hooper et al., 2014] or from a mixture of different methods in order to return a ranked list of annotations [You et al., 2017].

Several functional annotation methods use protein domain families as the basic unit of protein similarity [Peng et al., 2014, Forslund and Sonnhammer, 2008]. Nonetheless, despite the wide variety of existing function annotation techniques, protein function prediction is still an open problem because no universal method exists which clearly provides the best functional annotations. In response to this need, the CAFA (Critical Assessment of protein Function Annotation) experiment [Radivojac et al., 2013] was launched to assess the current state of the art in protein function annotation and to encourage developments in the field. This also motivated us to devise an approach to functionally annotate proteins.

Nonetheless, it should be noted that proteins rarely carry out their functions alone. They generally cooperate with other proteins by constructing a large network of protein-protein interactions [Gavin et al., 2002]. Protein-protein interactions are responsible for the majority of cellular functions and identification of such interactions is a way toward a better understanding of diverse cellular processes and molecular machineries of cells. Thanks to the high-throughput genomics approaches, the amount of protein sequences are dramatically increasing while experimental methods to discover their interactions are far behind. Many computational methods have been proposed to bridge the gap between known protein sequences and their interaction information. Studying molecular interactions at the protein level provides valuable intuitive understanding of how a molecule plays its roles inside a particular cell. However, for deeper insights into the interaction properties we found that predicting interactions at the protein domain level are very interesting and useful. There are a small number of single domain proteins that interact with their biological associates through their domains, a larger number of proteins have more than one domain [Apic et al., 2001]. Interactions in these multi-domain proteins, often, involve cooperating between two or more domains [Bhaskara and Srinivasan, 2011]. Therefore, identification of protein interactions at the domain level is logically useful to understand accurate atomic details in protein interactions and predict new interactions.

During the past few years, researchers have concentrated on computationally unearthing protein interactions at the domain level. One way to discover domain-domain interactions is using three-dimensional structures of proteins. KBDOCK [Ghoorah et al., 2013b], 3did [Stein et al., 2010], iPfam [Finn et al., 2013], and INstruct [Meyer et al., 2013] are four databases containing structural information about observed domain-domain interactions principally inferred from PDB chains. The quality of such observed interactions is very high, but the number of known domain-domain interactions is bounded by the availability of structural information about protein complexes. Even though these methods provided thousands of known domain interactions, the number of inferred protein interactions using these domain-domain interactions is far fewer than the actual number of protein interactions. Domain interactions inferred from structural data in 2010, can only cover around 5% of protein interactions in Saccharomyces cerevisiae and 19% of protein interactions in Homo sapiens [Yellaboina et al., 2010]. This encouraged us to introduce new methods to uncover all possible domain interactions.

Contributions

In this thesis we contributed to domains of knowledge discovery, protein function annotation, and protein interactions by proposing novel methods and applications. First, we proposed a data mining approach called CODAC for discovering direct associations between protein functions and protein domains

[Alborzi et al., 2018, Alborzi et al., 2017b, Alborzi et al., 2017c].

It then became apparent to us that our CODAC method for functional annotation of protein at the domain level could also be applied to the automatic functional annotation of the entire protein sequences. This led us to develop an extension of CODAC which we call CARDM (Combinatorial Association Rules Domain Miner). CARDM combines the CODAC learning step, in which function annotations are associated with protein domains, with a combinatorial rule generation and filtering procedure from which aggregated taxon-specific predictive models are constructed and used to annotate protein sequences and structures automatically.

We finally introduced a novel way to tackle the problem of discovering interactions between protein domains. Our method called PPIDM is derived from our previously developed CODAC method [Alborzi et al., 2018] and is to our knowledge the first method that predicts interactions between sets of protein domains.

It is worth mentioning that like any automatic mining approach, the methods proposed in this thesis do not produce validated results as manual curation would do, but they contain learning phase from manually curated data and combination of different techniques and databases that may make these methods more and more reliable. The produced results can be used by experimentalist to reduce the search space for finding candidates for certain association or interaction. Collaboration with biologist is ongoing to validate the results of our domain annotations.

Overview of Thesis

The rest of this thesis is organized as follows:

Chapter 1: Understanding data science and biological context is indispensable. Thus, this chapter covers essential data science context, such as knowledge discovery and data mining, association rule mining, vector space model, k-partite graphs, statistical hypothesis testing, and information filtering. Moreover, a general introduction to proteins structures and sequences, protein domains, protein functions and annotations, and protein-protein interactions is given in this chapter. This chapter also introduces resources which are used in the thesis.

Chapter 2: This chapter begins with our first problem : to assign EC numbers to protein domains. Our software to predict EC-domain associations is called ECDomaniMiner.

Chapter 3: This chapter describes a general approach for the computational discovery of associations between different sets of annotations by formalizing the problem as a bipartite graph enrichment problem in the setting of a tripartite graph.

Chapter 4: In this chapter, we describe a novel and comprehensive protein function prediction system (CARDM), which is used for the functional annotation of protein sequences in UniProtKB/TrEMBL. Using our generated prediction models, our CAPSID team participated in a challenge called CAFA that the results are also explained in detail.

Chapter 5: This chapter describes "PPIDM" (stands for Protein-Protein Interaction Domain Miner) to computationally discover interactions between single or subsets of Pfam protein domains. PPIDM is derived from the previously described CODAC method for computational discovering of direct associations using common neighbors.

Chapter 6: This chapter summarizes the contributions of this thesis, and it presents several short-term and long-term future directions.

Appendix A: This appendix describes the ECDomainMiner and GoDomainMiner web servers, which provides public access to the EC-Pfam and GO-Pfam/CATH/SCOP resources.

Appendix B: It depicts the integration of the functional annotations into the KBDOCK2 webserver.
Appendix C: It contains copies of the published articles and presented posters.

Chapter 1

Background

Contents

1.1 Data Science Context - Data Preparation, Mining, and Interpretation

1.1.1 Knowledge Discovery from Data and Data Mining

Many people treat data mining as a synonym of the knowledge discovery from data (KDD), however, we agree that they have different definitions. The process of scrutinizing large amounts of data for discovering patterns (considered as knowledge about the data) is described as Knowledge discovery [Frawley et al., 1992]. But data mining is an essential step in the process of knowledge discovery and refers to extracting or mining knowledge from large amounts of data stored in databases, data warehouses, or other data repositories [Fayyad et al., 1996]. Knowledge discovery as a process consists of an iterative sequence of following steps [Han et al., 2011], depicted in Figure 1.1:

- Data cleaning: To remove noise and inconsistent data,
- Data integration: To combine multiple data sources,
- Data selection: To retrieve data from database, relevant to the analysis task,

- Data transformation: To tranform or consolidate data into proper forms for mining such as summary or aggregation,
- Data mining: To apply intelligent approaches in order to extract data patterns and trends,
- Pattern evaluation: To identify the really interesting patterns representing knowledge based on some measures,
- Knowledge presentation: To visualize and represent mined knowledge to the users.

Step 1 to 4 are different forms of preprocessing of data where the data are prepared for mining. The data mining is an essential step that could include an interaction with user or a knowledge base. The interesting patterns are presented to the user and may be stored as new knowledge in the knowledge base. The architecture of a typical data mining system may have the main components illustrated in Figure 1.2 [Han et al., 2011]. In Figure 1.2, data are gathered from one or a set of data repositories. Data cleaning and integration techniques may be carried out on the data. A server fetches the data based on the request in the database server module. Data mining engine as the most essential part of the system, perform functional modules such as characterization, associations, classification, prediction, cluster, and other analysis. Interesting patterns are searched using particular measures in pattern evaluation module. User interface modules allows user interact with data mining system to specify a query or a task. The knowledge base is the domain knowledge that is used to guide the process or assess the interestingness of the patterns.

Data mining involves an integration of methods and techniques from multiple disciplines such as database or big data technologies, high-performance computing, deep learning, statistics, machine learning, data visualization, information retrieval, and etc. Note that data mining systems have to handle large amount of data, unless they should appropriately be called machine learning systems, statistical analysis tools, or experimental system prototypes.

1.1.2 Machine Learning and Data Mining

Data Mining can be defined as the subprocess of knowledge discovery process that starts from apparently unstructured data tries to extract knowledge and/or unknown interesting patterns. There are different ways to discover patterns out of datasets such as visualization techniques, topological data analysis, or machine learning. Machine Learning is a sub-field of data science that focuses on design and development of the algorithms with which machines gain the capability to learn without being explicitly programmed [Samuel, 2000]. It is obvious here that machine learning can be used for data mining. Nonetheless, data mining can exploit other techniques in addition to or on top of machine learning. Machine learning contains three types of learning; Supervised Learning, Unsupervised Learning, and Semi-Supervised Learning [Han et al., 2011, Witten et al., 2016]. Supervised learning refers to problems where the input variables and outputs are defined, and there is a need of an algorithm to learn the mapping function from the input to the output. The objective of the algorithm is to approximate the mapping function in a way that with a new input record, we can predict the correct output. This learning is called supervised because the process of learning from the training dataset can be assumed as a teacher supervising the process of learning. The correct answers are known, while the algorithm iteratively predicts based on the training data and in each iteration the answer is corrected by the teacher. Learning process ends when the algorithm reaches an acceptable level of performance.

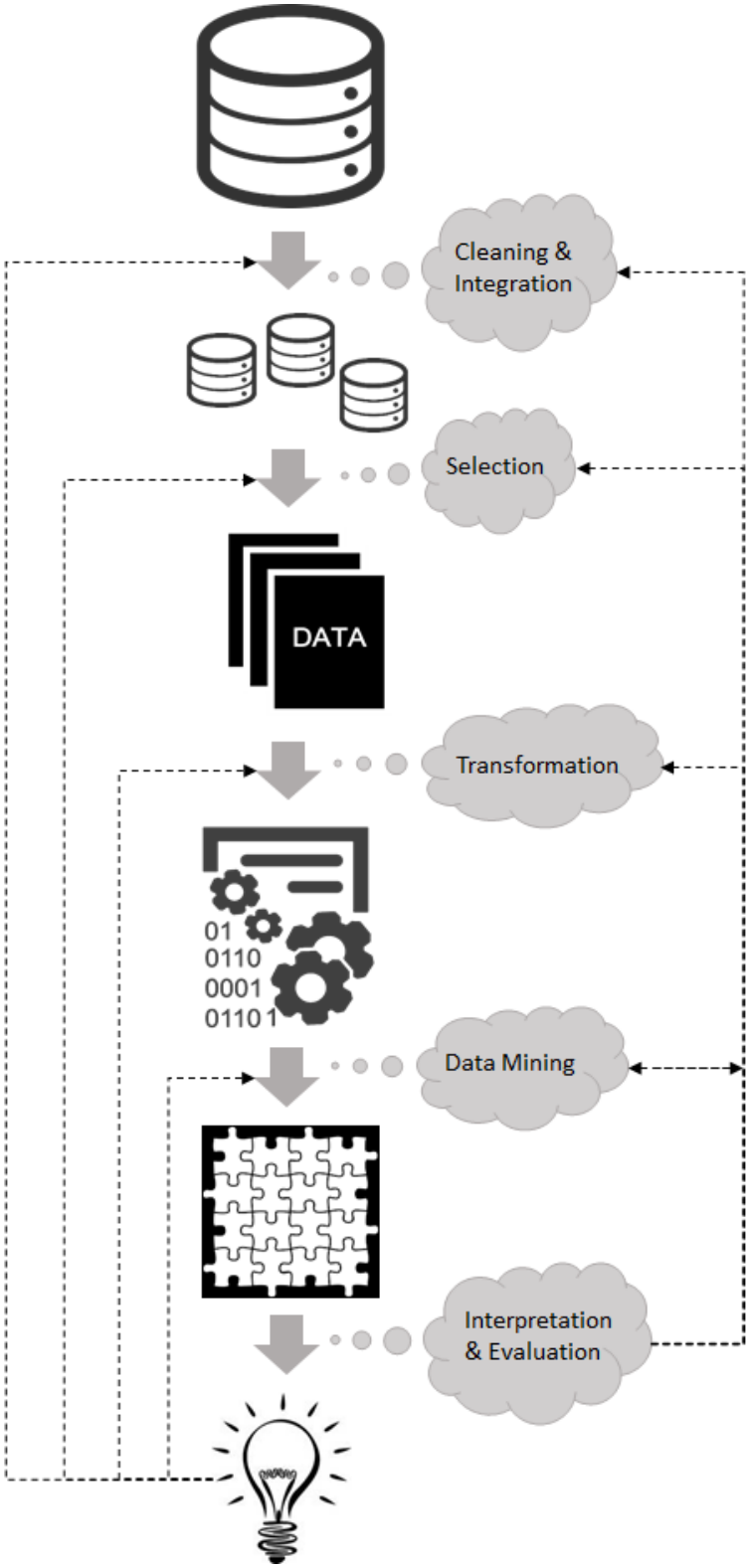


Figure 1.1: The process of knowledge discovery that considers data mining as a step in the process. Data cleaning, integration, selection, and transformation may be considered as data preparation for the data mining step. Discovered knowledge could be refined by returning to the previous steps.

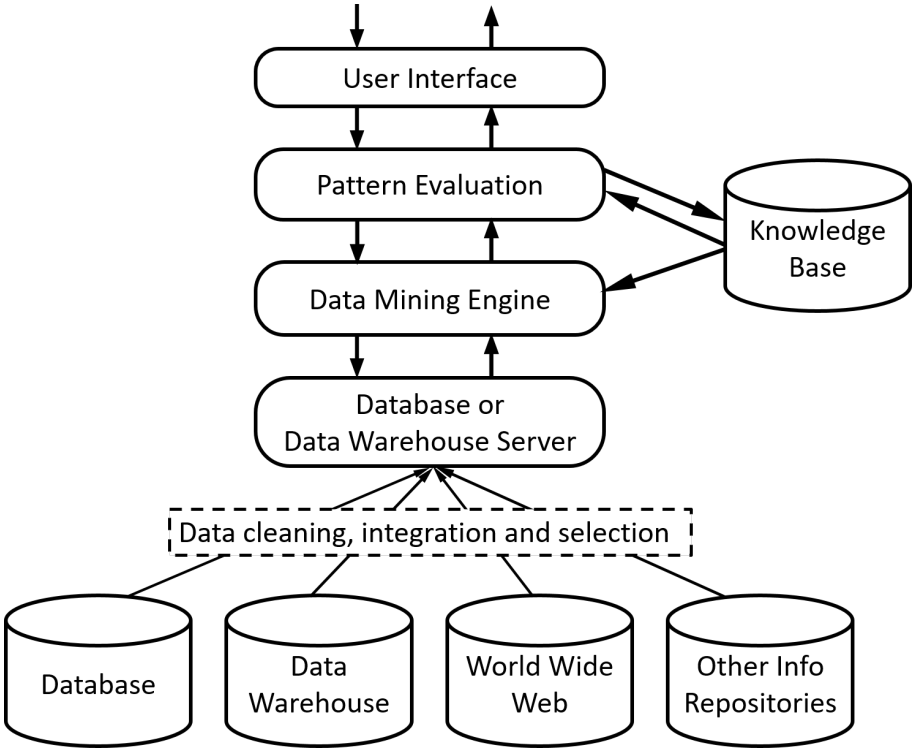


Figure 1.2: Architecture of a typical data mining system.

Unsupervised learning refers to problems where only input data is present and the corresponding output is unknown. The objective of the unsupervised algorithm is to model the fundamental organization or distribution in the data. Since there is no correct answers and there is no teacher, this type of learning is called unsupervised. Algorithms are expected to devise a system to discover and present the interesting structure in the data.

Semi-supervised learning is where a large amount of input data is unlabeled while only small number of output are labeled. These types of problems are in between the other two learnings. Many machine learning problems in the real world should be solved by semi-supervised algorithms. This is due to the fact that labeling data by domain experts is time-consuming and expensive whereas unlabeled data are easily collected in a much cheaper way. In semi-supervised learning, unsupervised learning techniques could be used to discover the structure in the input data. This can lead for instance to detect high-density regions in which labeled data can be used together with unlabeled data by supervised learning techniques [Chapelle et al., 2003].

Inferred data patterns in data mining are generally used to solve grouping similar data (Clustering) [Berkhin et al., 2006], classifying new data into known classes (Classification) [Phyu, 2009], finding unusual data (Anomaly Detection) [Chandola et al., 2009], finding a model of data (Regression) [Kotsiantis and Pintelas, 2009], representing data in a compact manner (Summarization) [Afantenos et al., 2005], constructing a new set of features from the original feature set [Wang et al., 2001], and finding dependencies between variables (Association Rule Mining) [Hipp et al., 2000]. Figure 1.3 illustrates the differences between the general problems in data mining, machine learning algorithms to tackle the problems, learning types, and the examples of usages.

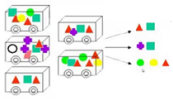
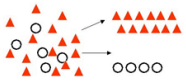
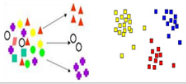
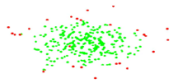
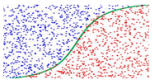
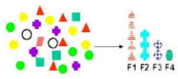
Problem	Supervision	Illustration	Algorithm	Applicability
Association Rules	Mainly Unsupervised		Apriori FP-Growth	Link analysis Market cart analysis
Classification	Supervised		Decision Trees Naïve Bayes Support Vector Machine	Text mining Wide and narrow data Patterns analysis
Clustering	Unsupervised		Hierarchical K-means Hierarchical O-cluster	Text mining Product grouping
Anomaly Detection	Supervised Unsupervised		One class SVM	Lack of samples
Regression	Supervised		Support Vector Machine Generalized Linear Model	Text mining Wide and narrow data
Feature Extraction	Supervised Unsupervised		Multilayer perceptron Non-negative Matrix Factorization	Text analysis Feature reduction

Figure 1.3: Learning algorithms to solve the data mining problems (images from gerardnico.com).

Frequent Patterns and Association Rules

Frequent patterns and association rules will be detailed here as an example of data mining technique and because they will be used later in the thesis. Frequent patterns are patterns that frequently appear in a dataset. For instance, a set of items that appears frequently in a transaction list is a frequent itemset [Han et al., 2011]. Discovering such frequently appearing patterns plays an significant role in mining associations among data. Frequent itemset mining is to discover associations and correlations among items in a large transactional dataset. Market basket analysis is a typical example of frequent itemset mining. This process analyzes customer purchasing behaviors by finding associations between various items that customers buy. Finding such associations assists sellers to develop their marketing strategies by understanding which items are frequently sold together. For example, if customers are buying milk, how likely do they also buy bread at the same time?

At a store with set of items, each item has a Boolean variable representing the absence or presence of that item on the store shelf. Then, each customer can be a Boolean vector of values assigned to these variables. Customers, Boolean vectors, can be analyzed for purchasing patterns. These patterns reflect items that are frequently purchased, associated, together, and be presented in the form of association rules.

In the other words, association rules are if-then statements aiming to uncover dependencies of implicitly related data in a data repository [Hipp et al., 2000]. Such information can be applied as the foundation of decision making processes in variety of areas such as bioinformatics [Alborzi et al., 2012]. It should be noted that association rules generally do not take the order of items into account.

A typical association rule has two parts of "if" and "then" namely called an antecedent (if) and a consequent (then).

Selecting interesting rules from a set of rules is based on the restriction on diverse ways to calculate the significance and interestingness of rules. Three main interesting measures of rules are "Support", "Confidence", and "Lift". Support indicates how frequently items are present in the data source (equation

1.1 and 1.2), confidence indicates that the number of the antecedent-consequent statements found to be true in the data source (equation 1.3), and lift shows if antecedent and consequent are independent (equation 1.4). According to the definition by [Agrawal et al., 1993], let $I = \{i_1, i_2, \ldots, i_n\}$ be a set of n items, and $T = \{t_1, t_2, \ldots, t_m\}$ be a set of transactions called the database, where each transaction with unique transaction identifier in T contains a subset of items from I. Therefore, an association rule is defined as $I_l \Rightarrow I_r$, where both I_l and I_r are itemsets composed of the items in I and $I_l \cap I_r = \emptyset$. The support of each itemset, I_x, with respect to the T, is introduced as the ratio of transactions, t, in the T which contains items in I_x to the whole size of T,

$$Support(I_x) = \frac{|t \in T; I_x \subseteq t|}{|T|} \tag{1.1}$$

the support of a rule, $I_l \Rightarrow I_r$, with respect to the T, is introduced as the ratio of transactions, t, in the T which contains $I_l \cup I_r$ to the whole size of T,

$$Support(I_l \Rightarrow I_r) = \frac{|t \in T; I_l \cup I_r \subseteq t|}{|T|} \tag{1.2}$$

the confidence of a rule, $I_l \Rightarrow I_r$, with respect to the T, is the ratio of the transactions containing both I_l and I_r to transactions having I_l,

$$Confidence(I_l \Rightarrow I_r) = \frac{Support(I_l \cup I_r)}{Support(I_l)} \tag{1.3}$$

and the lift of a rule, $I_l \Rightarrow I_r$, with respect to the T, is the ratio of the transactions, containing both I_l and I_r to transactions having I_l multiplied by transactions having I_r,

$$Lift(I_l \Rightarrow I_r) = \frac{Support(I_l \cup I_r)}{Support(I_l) \times Support(I_r)} \tag{1.4}$$

It is worth highlighting that an association rule with a confidence value close to 1 has usually high quality, while a rule with lift close to 1 implies that its antecedent and consequent are independent from each other. When two itemsets are independent, no association rule can be established involving those two itemsets. Association rules built by frequent itemsets across all transactions might have high confidence values. However, it is also possible that the lift values of the rules are very close to 1, and the relations between their itemsets could easily be a fluke. Thus, exploiting both confidence and lift values of a rule enhances our understanding of its reliability. There are several techniques, with their individual features, to generate association rules such as Apriori [Agrawal et al., 1994], Eclat [Zaki, 2000], and FP-growth [Han et al., 2000].

Figure 1.4 shows an example of how to calculate measures for candidate association rules.

1.1.3 Information Filtering and Recommendation Systems

The amount of disseminated information and data are abundantly increasing [Wurman, 1989]. An information filtering (IF) system uses automatic or semi-automatic methods to eliminate information which are undesired to users from flows of information. The main purpose of information systems is to form filters in order to deal with overloads of data due to the information explosion. Recommendation system (recommender system) is a subclass of information filtering systems that actively attempts to predict items that users are interested in [Ricci et al., 2011, Robillard et al., 2014]. Recommendation systems build a profile for each user and then compare it to multiple reference attributes. These attributes are typically stemmed from the characteristics of items (content-based filtering approaches)

Rule	Support	Confidence	Lift
$A \rightarrow D$	2/5	2/3	10/9
$C \rightarrow A$	2/5	2/4	5/6
$A \rightarrow C$	2/5	2/3	5/6
$B\&C \rightarrow D$	1/5	1/3	5/9

Figure 1.4: An example of association rule mining. Transactions are presented as combinations of A, B, C, D and E items in the baskets. Item supports are Supp(A) = 3/5, Supp(B) = 3/5, Supp(C) = 4/5, Supp(D) = 3/5, and Supp(E) = 2/5. Support, confidence and lift measures of the four sample rules are expressed. For the first rule, there are two out of five transactions containing items A and D together (Support), item D exists in two transactions out of the three transactions that item A exists (Confidence). The lift of the first rule equals to $2/5 \div (3/5 \times 3/5)$.

[Balabanović and Shoham, 1997, Lops et al., 2011] or prior interests and behavior of users (collaborative filtering approaches) [Sarwar et al., 2001, Koren and Bell, 2015] or a hybrid of the collaborative filtering and content-based filtering approaches [Burke, 2002]. Collaborative filtering approaches are divided in two groups. User-based approaches which recommend items by finding similar users, and item-based approaches which calculate similarity between items and then make recommendations. Figure 1.5 displays the differences between collaborative and content-bases filtering in recommendation systems.

Usage and popularity of recommender systems are more and more increasing and now it is used as a solution in diverse areas such as online movie recommendation [Davidson et al., 2010, Gomez-Uribe and Hunt, 2016], financial services [Felfernig et al., 2007], and collaborative research [Chen et al., 2011].

1.1.4 Data Structure and Representation

Vector Space Model

Vector space model is the representation of a set of objects (particularly text documents) as vectors and their attributes as dimensions with identifiers in a vector space. It is a fundamental topic for data representation in information retrieval, information filtering, indexing and relevancy rankings [Raghavan and Wong, 1986].

Similarity between vectors (document and query vectors) in vector space models is calculated using associative coefficients such as Cosine, Jaccard and Dice coefficients. These are measures based on the normalized scalar product of two vectors where shared attribute indicates similarity. The most commonly used similarity measure for real-valued vectors in high-dimensional positive spaces is the cosine coefficient.

Cosine similarity is a measure of similarity between two non-zero vectors of an inner product space that gauges the cosine of the angle between them. Cosine similarity is exclusively used in positive space

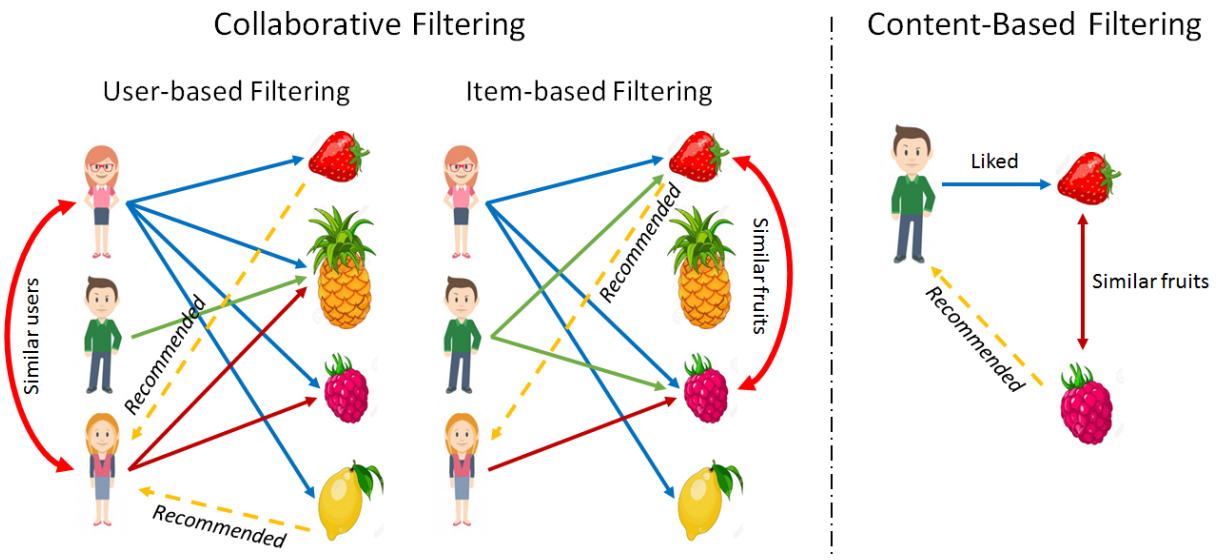


Figure 1.5: Collaborative and content-based filtering recommendation systems. **Left**: User-based filtering recommends (yellow dashed lines) strawberry and lemon to the customer, because of her similar taste to another customer. Item-based filtering calculate similarity between fruits according to the customers, and then recommends (yellow dashed lines) strawberry to the user. **Right**: Content-based filtering finds the similar fruits based on their characteristics, and then recommend (yellow dashed lines) raspberry to the customer based on his interest in strawberry.

and results in a similarity score limited between zero and one. It is worth mentioning that this range applies for any number of dimensions.

For example in information retrieval, documents are defined as vectors where where the values of dimensions correspond to the term frequency in the document multiplied by the inverse frequency of documents containing the term. Cosine similarity then gives a powerful gauge to understand and analyze the similarity between each two documents in terms of the relatedness of their subjects [Singhal, 2001, Muflikhah and Baharudin, 2009]. Furthermore, such a technique is used in the field of data mining and clustering [Alborzi et al., 2016, Tsiptsis and Chorianopoulos, 2011] to calculate unity between the members of clusters, and classification by neural networks to enhance systems accuracy and speed [Chunjie et al., 2017].

The cosine of two non-zero vectors, A and B, is defined as the Euclidean dot product mentioned in equation 1.5.

$$A.B = ||A||\ ||B||\ cos(\theta) \tag{1.5}$$

Therefore, given two vectors with their attributes, the cosine similarity $cos(\theta)$ is calculated using a dot product and magnitude as equation 1.6.

$$cos(\theta) = \frac{A.B}{||A||||B||} = \frac{\sum_{i=1}^{n} A_i B_i}{\sqrt{\sum_{i=1}^{n} A_i^2}\sqrt{\sum_{i=1}^{n} B_i^2}} \tag{1.6}$$

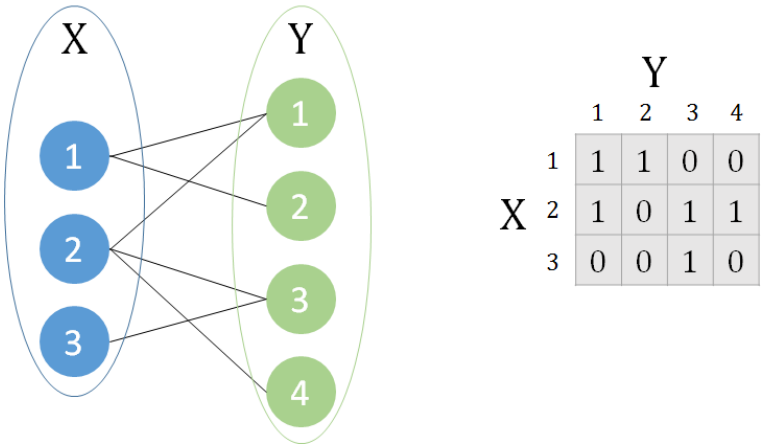


Figure 1.6: Translating a bipartite graph into a binary matrix. A bipartite graph sample comprise of two sets of items, X and Y. The items in X and Y are connected via edges which can be translated into a binary matrix. This binary matrix shows 1 if there is an edge between the item in X and the item in Y.

k-partite Graph

In mathematic and graph theory, a graph whose vertices are partitioned into k disjoint sets is called a k-partite (multi-partite) graph [Godsil and Royle, 2013, Bollobas, 2012]. In k-partite graph, vertices can be colored with k different colors and there is no edge between the vertices with a same color. If the graph vertices are divided into two independent sets it is called bipartite graph (bigraph) [Asratian et al., 1998]. Similarly, a tripartite graph is defined as a graph where vertices are partitioned into 3 independent sets. A k-partite graph can be translated into an adjacency matrix. The adjacency matrix of a bipartite graph $G = (U, V, E)$, with two disjoint sets of vertices ($|U|$ and $|V|$) is called biadjacency matrix. Biadjacency matrix is a binary matrix of size $|U| \times |V|$. This binary matrix has 1 for pairs of adjacent vertices and a 0 for pairs of non-adjacent vertices [Asratian et al., 1998]. Figure 1.6 shows a bipartite graph and its representative biadjacency matrix.

1.1.5 Statistical Validation of Extracted Pattern

Statistical data analysis is the process of the accumulating, analyzing, interpreting, presenting data based upon laws of probability in order to discover models and trends or validate patterns [Lindley, 2000]. The most popular type of statistical analysis is hypothesis testing. In statistical analysis, mathematical principles are used to calculate a probability that a sample results match the hypothesis about a population [Banerjee et al., 2009]. For instance, if an investigating hypothesis is that a coin is not fair, principles of statistics are used to estimate the probability of obtaining the samples if the investigating coin were unbiased (null hypothesis). If getting the sample results from a fair coin has very low probability, it is safe to reject the null hypothesis and deduce from the results that the coin is not fair. It should be noted that we can never say that the coin is certainly biased due to the fact that even using an unbiased coin might generate the sample results. However, we can come to the conclusion that the coin is biased by stating that our sample results oppose the null hypothesis with strong evidence [Banerjee et al., 2009, Wasserman, 2013].

In statistical hypothesis testing, a p-value (p for probability) is often reported to present that the sample results provide strong evidence against the null hypothesis or not. The p-value is indeed a

numerical measurement in hypothesis test to show its statistical significance. This measure presents the probability of getting our sample data (e.g., 20 tails) if the null hypothesis is true (e.g., unbiased coin). Conventionally, the p-value less than 5% (p-value < 0.05) supports the rejection of null hypothesis (i.e., the coin is biased). It means that there is strong evidence to assume that the null hypothesis is false if the $p < 0.05$, and it can be concluded that the results are statistically significant.

In computational research, methods usually produce a large set of results. In order to validate or filter the results from the view of statistical analysis, we could design a statistical significance test of the predictions. Thus, there is a set of hypotheses that we wish to test simultaneously. We could test each hypothesis separately, using the typical level of significance $\alpha = 0.05$. It is accepted [Mittelhammer et al., 2000] that the probability of observing at least one significant result only by chance is:

$$P\ (at\ least\ one\ significant\ result) = 1 - P\ (no\ significant\ results)$$

For example, if we consider a study where 20 hypotheses should be tested, we have a 0.64[1] chance of observing at least one significant result, even in the situation that all of the tests are indeed not significant. In biological related fields, the number of simultaneous tests often is greater than 20. Therefore, the probability of obtaining at least one significant result just by chance is very high. Bonferroni correction is a way to neutralize such a problem of multiple comparisons of independent test by reducing the significance threshold level to "α/n" [Mittelhammer et al., 2000]. In the example above, with 20 tests and $\alpha = 0.05$, we would reject a null hypothesis if the p-value of the test is less than 0.0025 (0.05/20). Thus, the probability of observing at least one significant result while using the Bonferroni correction is reduced to 0.049[2]. It should be noted that the Bonferroni correction could be exceedingly conservative, depending upon the correlation in the structure of the tests, it often leads to label too many predictions as false negatives [Benjamini and Hochberg, 1995].

1.2 Biological Context - Protein Function, Domain, and Interaction

1.2.1 Protein Sequence and Structure

Proteins are the most versatile macromolecules in living organisms responsible for functions in biological processes. At their most basic level, they are made up of a sequence of amino acids, specified by the sequence of nucleotides in a gene. There are 20 various amino acids in living organisms for construction of proteins. Amino acids includes both an acidic carboxyl group ("-COOH") and a basic amino group ("-NH2"). Different amino acids attach to each other in long chains. They form peptide bonds, amide bonds between the "-COOH" of one amino acid and the "-NH2" of another amino acid (Primary structure of protein). The terms protein or polypeptide are referred to sequences longer than 50 amino acids while sequences with fewer amino acids are usually called peptides. A protein can be formed by one or more polypeptides. Each peptide or protein sequence has two ends. The end of the sequence with a free carboxyl group terms the carboxy-terminus (or C-terminus) and the end of the sequence with a free amino group is called amino-terminus (or N-terminus).

Proteins fold into a three-dimensional structure based upon their amino acid sequences (amino acids have different biochemical properties) and their environment. This allows them to have interaction with other molecules and proteins and carry out their functions (Figure 1.7).

[1] $1 - (1 - 0.05)^{20}$
[2] $1 - (1 - 0.0025)^{20}$

Proteins or peptides strands have unique characteristic secondary structure. Depending on hydrogen bonding, the two principal secondary structures are the "α-helix" and the "β-sheet". The general three-dimensional organisation of all the secondary structure elements of a protein constitutes its tertiary structure. Thus, a protein molecule bends and twists in order to attain lowest energy state or maximum stability. Although, the amino acid sequence constitutes the primary structure of proteins, the chemical and biological properties of proteins highly depends on the three-dimensional (or tertiary structure). The tertiary structure can be described by a single polypeptide chain called backbone, with one or more protein secondary structures elements In many cases, a protein can be divided into several structural domains. Such domains can be described as a "fold" composed of a succession of secondary elements (α-helices or β-sheets) arranged in a particular 3D shape. Similar structural domains can be recognized in different proteins. They correspond to conserved subsequences that can be found in various proteins. Finally, many proteins consist of multiple polypeptides, referred to as protein subunits. The quaternary structure is a large aggregated protein complex that is formed by interactions between these subunits.

UniProtKB

UniProt Knowledgebase (UniProtKB) is a biological repository of protein sequences and their functional information which are curated by experts to a limited extent [Apweiler et al., 2017]. These protein sequences are principally obtained from genome sequencing projects and a major part of the functional information of proteins acquired from the scientific publications. UniProtKB is composed of two parts called "UniProtKB/SwissProt" and "UniProtKB/TrEMBL".

UniProtKB/SwissProt is a section of the UniProtKB containing non-redundant, manually annotated protein sequences [Boutet et al., 2007]. UniProtKB/SwissProt annotations derive from information extracted from biological literature merged with computational analysis evaluated by biocurator. UniProtKB/SwissProt aims to accumulate all known information with detailed analysis related to a specific protein in the database.

In the UniProtKB/SwissProt database, to eliminate data redundancy, differences between protein sequences from the identical gene and species (e.g. incorrect initiation sites or exon boundaries, natural variation, and alternative splicing) are documented and then merged into one entry.

Protein sequences are often annotated with gene and protein names, protein functions, enzymatic activity information such as cofactors and catalytic residues and activity, subcellular localization, protein interactions, expression pattern, protein domains and families, and etc.

As of July 2017, the increasing amount of protein sequences in UniProtKB/SwissProt database over thirty years is shown in Figure 1.8 (A) [3]. It shows that during three years from 2007 to 2010 the size of database was doubled, however, it expanded only by 10% during the past five years. Figure 1.8 (B) displays that majority of sequences in the UniProtKB/SwissProt database are Bacteria proteins.

UniProtKB/TrEMBL is the automatically annotated section of the UniProtKB containing computationally analyzed protein sequences [Bateman et al., 2014]. UniProtKB/TrEMBL came to existence in response to the burst of data generation by genome projects and incapability of manual curation process of UniProtKB/SwissProt to address all protein sequences. Translated coding sequences from the EMBL-Bank [Stoesser et al., 2002], GenBank [Benson et al., 2017], and DDBJ [Tateno et al., 2000] databases in addition to the protein sequences from the Protein Data Bank [Berman et al., 2006], and from gene prediction databases such as Ensembl [Yates et al., 2016] are processed and inserted into the UniProtKB/TrEMBL in a completely automatic fashion.

[3] `http://www.uniprot.org/statistics/Swiss-Prot`

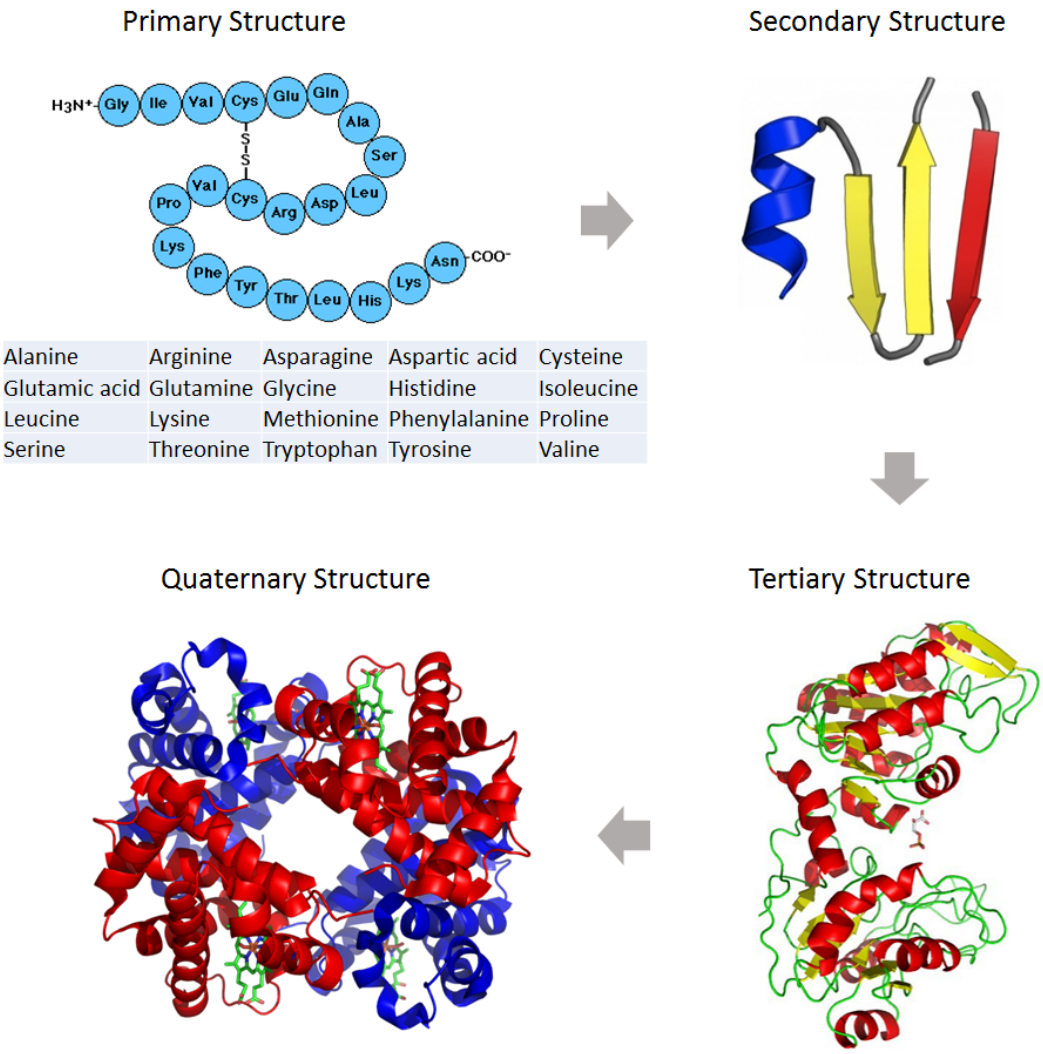


Figure 1.7: List of 20 amino acids, primary, secondary, tertiary, and quaternary protein structures.

As of July 2017, the number of protein sequences in UniProtKB/TrEMBL database is illustrated in Figure 1.9 (A) [4]. It explicitly shows the huge increase in the number of protein sequences which is mainly because of the high-throughput genome sequencing. Similar to the UniProtKB/SwissProt, Figure 1.9 (B) shows that most of the available sequences in the UniProtKB/TrEMBL database are Bacteria proteins.

In the rest of this thesis, we use SwissProt term for the UniProtKB/SwissProt database, and TrEMBL for the UniProtKB/TrEMBL database.

UniRef

The UniProt Reference Clusters (UniRef) [Suzek et al., 2007, Suzek et al., 2014] is a resource divided into three databases containing clustered protein sequences from UniProtKB (both UniProtKB/SwissProt and UniProtKB/TrEMBL) and selected UniParc records [Leinonen et al., 2004]. UniRef100 is one of

[4] https://www.ebi.ac.uk/uniprot/TrEMBLstats

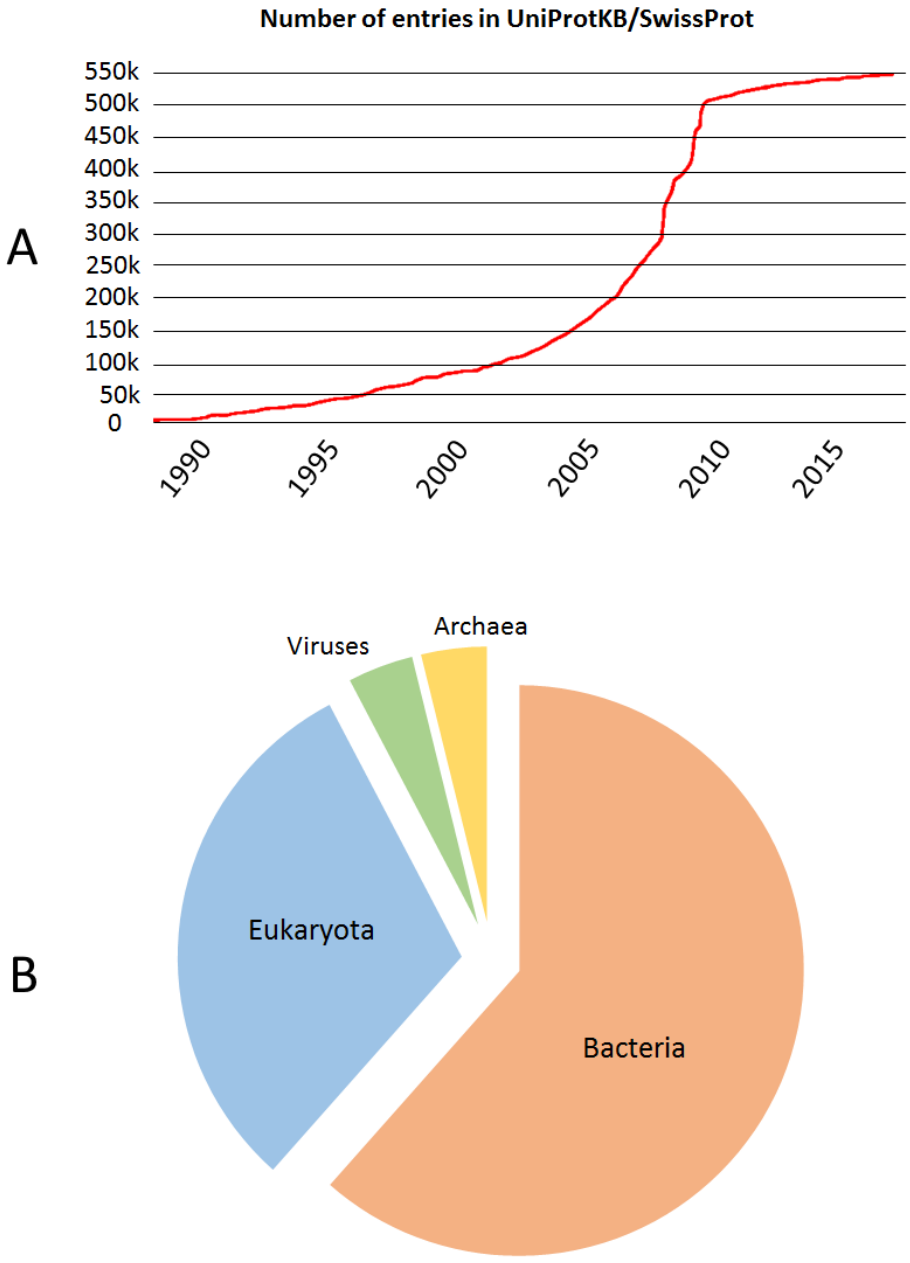


Figure 1.8: The number of entries in UniProtKB/SwissProt. A: Number of proteins sequences in UniProtKB/SwissProt database over time. There is an intensive growth in the number of proteins sequences from 2007 to 2010. B: Proportions of SwissProt entries per taxonomic kingdom.

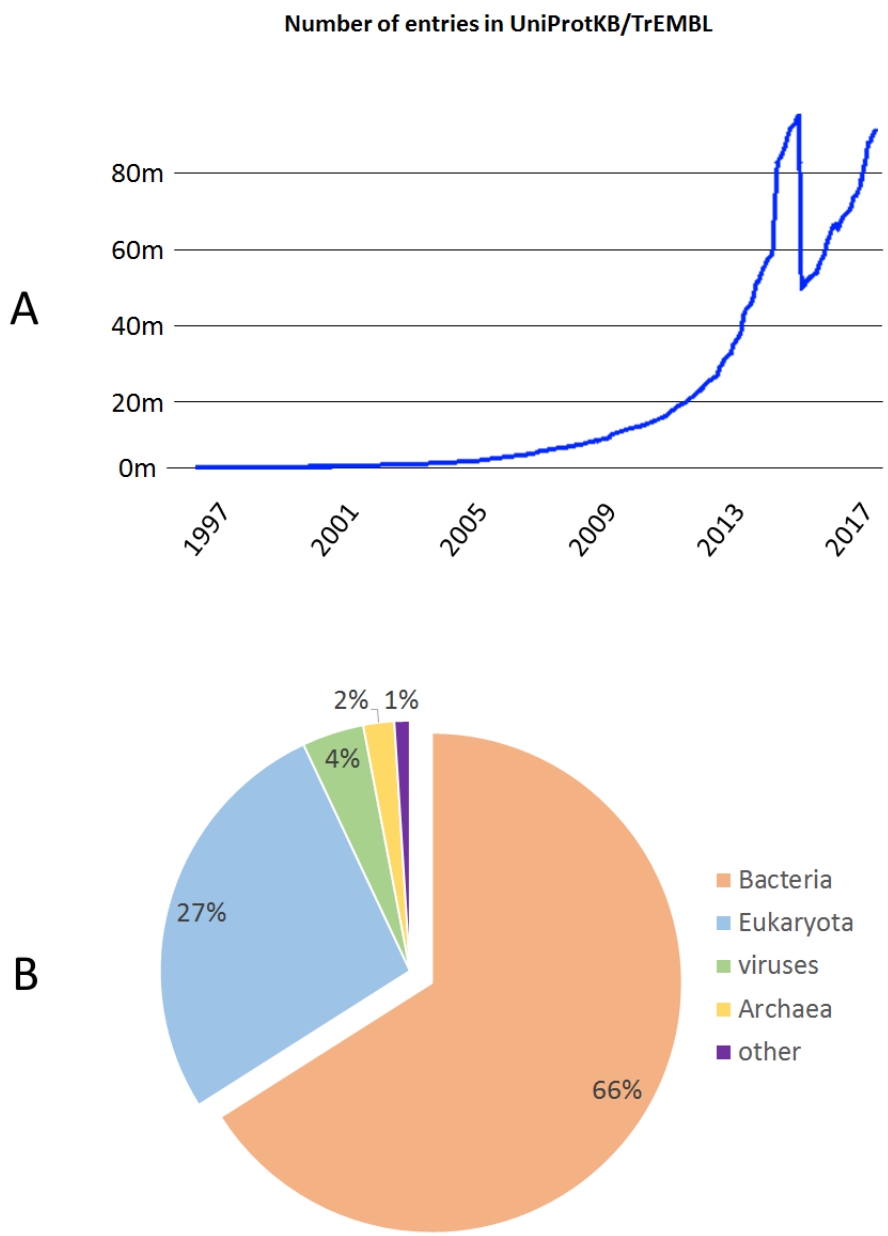


Figure 1.9: The number of entries in UniProtKB/TrEMBL database. A: Number of proteins sequences in UniProtKB/TrEMBL database over time. In mid 2015, TrEMBL size is dropped due to the proteome redundancy. This caused removing a large number of entries deemed as redundant [Bursteinas et al., 2016, Apweiler et al., 2017]. B: Proportions of the TrEMBL entries per taxonomic kingdom.

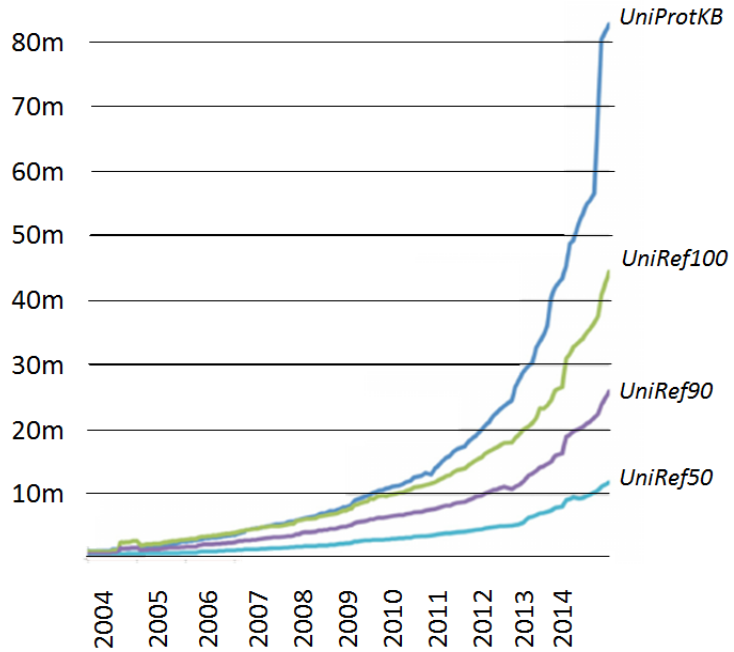


Figure 1.10: Growth of UniProt and UniRef databases.

databases that incorporates identical sequences (or sequence fragments) from any organism into a single UniRef record. Each UniRef entry has the UniProtKB accession numbers of all combined entries, as well as the sequence of a representative protein and the corresponding records in UniProtKB and UniParc. Furthermore, UniRef100 sequences are clustered using the CD-HIT algorithm [Li and Godzik, 2006] to construct new clusters of protein sequences with less similarity called UniRef50 and UniRef90 [Li et al., 2001]. In UniRef50 and UniRef90, each cluster includes protein sequences with at least 50% and 90% sequence similarity, respectively, to the longest protein sequence. Figure 1.10 shows how the number of entries in the UniRef100, UniRef90 and Uniref50 are increasing in comparison to the increase in the UniProtKB from 2004 until 2015 [Bateman et al., 2014].

PDB

The Protein Data Bank (PDB) [Bernstein et al., 1977, Berman et al., 2006, Gutmanas et al., 2014] is a database containing the three-dimensional structural data of biological molecules, like nucleic acids and proteins, acquired from experimental methods. Figure 1.11 shows different types of protein structures from the PDB[5].

The PDB database is an important resource in structural biology research and currently holds more than 127 thousand structures. These 3D structures are obtained and submitted by biologists and biochemists using mainly NMR spectroscopy, X-ray crystallography, and recently increasing cryo-electron microscopy (Cryo-EM). As of 14 March 2017, a categorization of the available PDB data based on its properties is shown in Table 1.1.

The PDB database is supervised by Worldwide Protein Data Bank (wwPDB - https://www.wwpdb.org/)

[5] https://commons.wikimedia.org/wiki/File:Protein_structure_examples.png

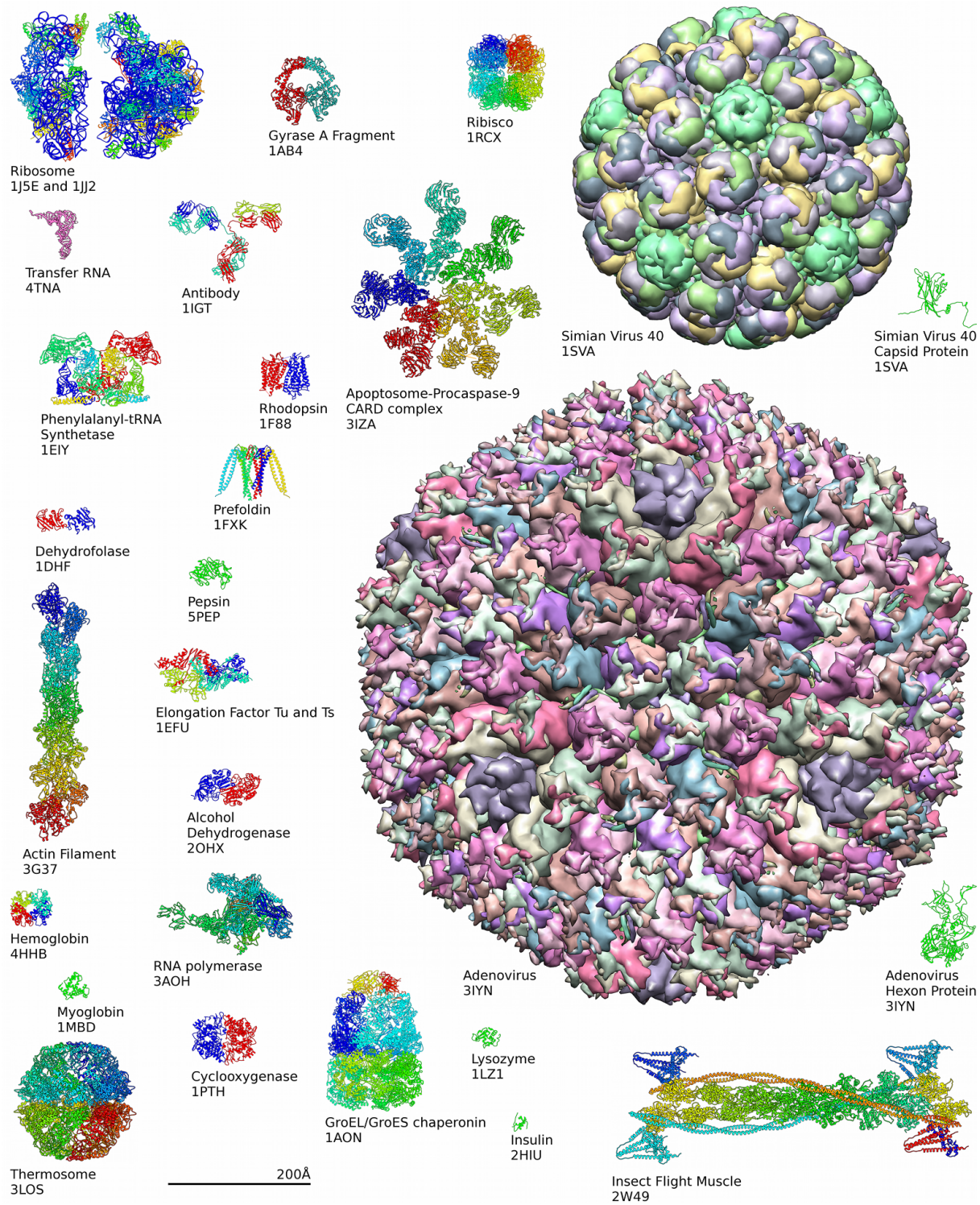


Figure 1.11: Examples of protein structures from the PDB

Experimental Method	Proteins	Nucleic Acids	Protein/Nucleic Acid complexes	Other	Total
X-ray crystallography	106595	1820	5471	4	113890
NMR spectroscopy	10296	1190	241	8	11735
Cryo-EM	1021	30	367	0	1418
Hybrid	99	3	2	1	105
Other	181	4	6	13	204
Total	118192	3047	6087	26	127352

Table 1.1: The PDB data (July 2017).

and structural data are accessible via the three member organizations called PDBe (https://www.ebi.ac.uk/pdbe/), PDBj (https://pdbj.org/), and RCSB (https://www.rcsb.org/).

SIFTS

The Structure Integration with Function, Taxonomy and Sequences resource [Velankar et al., 2012] is a database that cross-references PDB entries to biological resources. These resources are protein domain and family classifications such as Pfam, SCOP, CATH, and InterPro, or functional ontologies such as Gene Ontology (GO), and Enzyme Commission Numbers or the NCBI taxonomy database. Moreover, it maintains cross-reference information to UniProt entries, for PDB entries existing in the UniProt database. SIFTS database is updated weekly in close collaboration between the PDBe and UniProt using a semi-automated process. The SIFTS pipeline has two main phases. First, a semi-automated procedure cross-references the most recent UniProtKB entries for protein chains in the PDB. Second, an automated process produces correlations between proteins in the PDB and the corresponding UniProtKB sequence at the residue-level. In the second phase, cross-reference information to other biological databases are generated. SIFTS database is available at http://pdbe.org/sifts/.

1.2.2 Protein Function

Proteins carry out a large number of functions within living organisms. These functions vary from catalysing metabolic reactions and DNA replication to responding to stimuli, and transporting molecules from one location to another. Although protein functions can be described in multiple ways, researchers mainly define them with ontology terms from classification schemes provided by the Gene Ontology (GO) Consortium [Harris et al., 2004] and numerical classification scheme designed only for enzymatic functions called Enzyme Commission number (EC) [Webb et al., 1992].

Gene Ontology

Gene ontology (GO) provides a collection of controlled vocabularies in structured way to unify the representation and annotation of gene and gene products [Ashburner et al., 2000, Harris et al., 2004]. The main objectives of GO is to develop and maintain the controlled vocabularies in a way which is easily readable by machines as well as being unified across all species. Gene Ontology is divided into three ontologies as follows.

- **Molecular Function**: The functions of proteins at the molecular level such as catalysis or binding.

- **Biological Process**: Molecular events or operations with a determined start and finish, which is relevant to the function of living components such as cells, tissues.
- **Cellular Component**: The parts of a cell or the extracellular environment.

Information about each GO term within the GO ontology is organized into several items:

- **Term name**: A word or string of words.
- **Alphanumeric ID**: An accession number for accessing the information.
- **Namespace**: Ontology that the term belong to.
- **Synonyms**: Exactly equivalent, broader, narrower, or related names.
- **Reference**: Equivalent concept in other databases.
- **Comment**: Term meaning or usage.
- **Alternate ID**: Another ID of the term, mainly obsolete ones.
- **Relationship**: For relating a term with its ancestors and descendants. "is_a" relations operate between terms in the same category of GO ontology. "part_of" and "regulates" operate between different GO categories.

The GO ontology is built as a rooted Directed Acyclic Graph (DAG), and each GO term has one or many defined relationships to other GO terms. This relationship can be intra-ontology (using is_a) or inter-ontology (using part_of or regulates). One of the significant design feature of the GO vocabulary is to be species-neutral. It contains terms applicable to eukaryotes and prokaryotes which can be either single and multi-cellular organism. Figure 1.12 displays one example in GO hierarchy.

In this figure, each box represents a GO term ID with its name. Colored arrows show the relations between the GO terms while the colored lines express the types of relations (black: is_a, blue: part_of, and yellow: regulates, etc). It should be noted that GO terms are more specific going down the graph toward the leaf nodes and more general terms at the top of the graph toward the root nodes (molecular function, biological process, cellular component). GO terms may be linked to more than one parent GO term via different types of relations.

In order to show the GO terms information, there are several online resources with different criteria and features such as Amigo [Carbon et al., 2008] and QuickGO [Binns et al., 2009].

Enzyme Commission Numbers

The Enzyme Commission (EC) number is a numerical classification scheme for enzymatic activities. EC numbers are introduced in regards with the chemical reactions they catalyze [Webb et al., 1992]. Based on the naming system for EC, a recommended name for the respective enzyme is assigned to each EC number. It should be noted that EC numbers specify only reactions that enzymes are involved in. In other words, they define the function of the protein but they do not present any information about the protein itself. Therefore, diverse proteins (from different organisms) that catalyze the same reaction receive the same EC number [Omelchenko et al., 2010].

Each enzymatic activity as a code consists of the letters "EC" followed by four numbers which are separated by dots. These progressive numbers represent a more and more detailed classification of the enzyme. For example, the oxalate oxidase has the code "EC 1.2.3.4", whose components point out the following levels of information for the enzymatic activity:

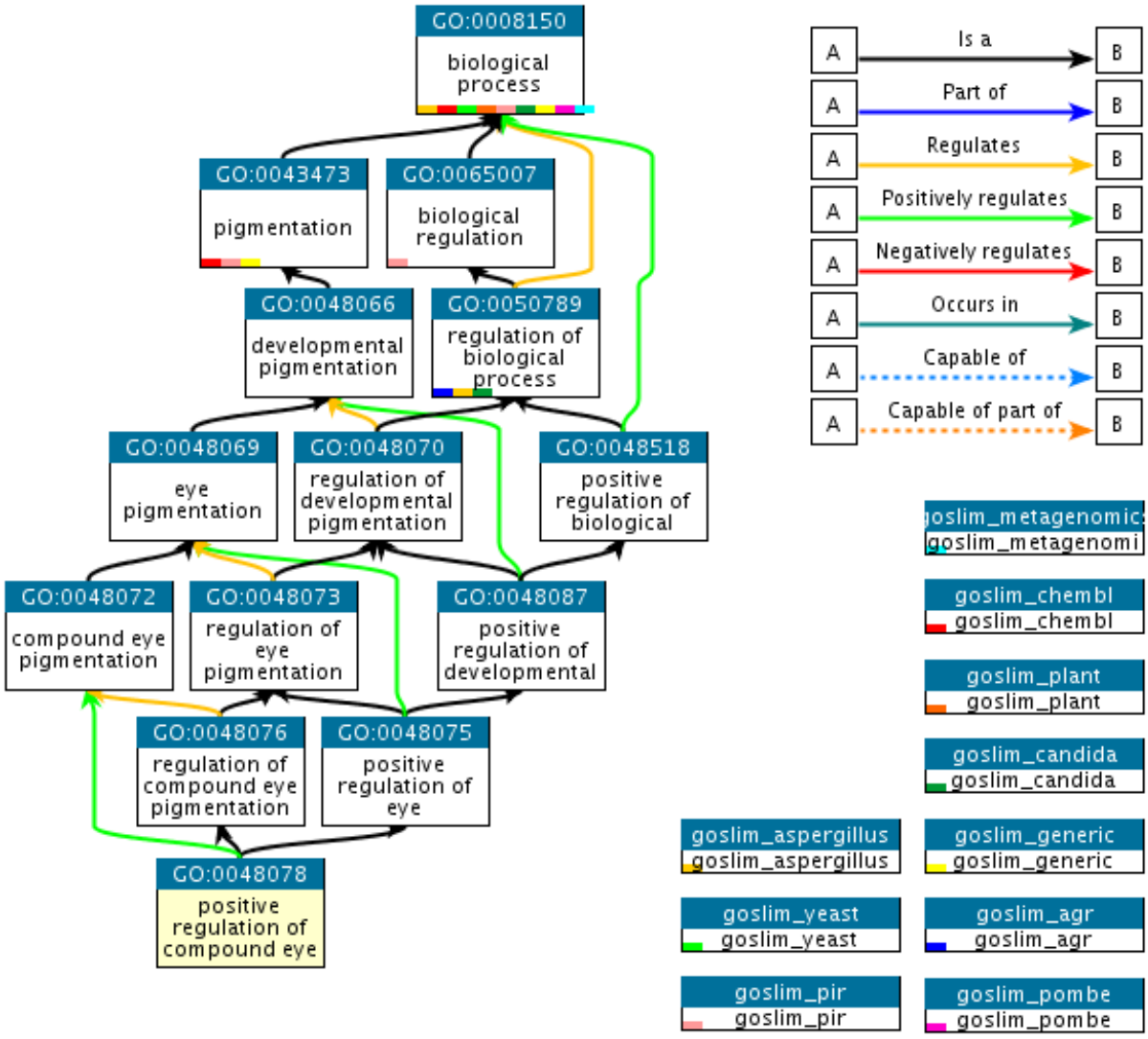


Figure 1.12: A hierarchical view of the relations between GO:0048078, Positive Regulation of Compound Eye.

- EC 1: Oxidoreductases.
- EC 1.2: Acting on the aldehyde or oxo group of donors.
- EC 1.2.3: With oxygen as acceptor.
- EC 1.2.3.4: Oxalate oxidase

Overall, there are six primary classifications of enzymes in the Enzyme Commission Codes as follow.

- EC 1: Oxidoreductases,
- EC 2: Transferases,
- EC 3: Hydrolases,
- EC 4: Lyases,
- EC 5: Isomerases,
- EC 6: Ligases.

In order to show the EC number information, there are online databases such as BRENDA [Schomburg et al., 2002], IntEnz [Fleischmann et al., 2004], KEGG Enzyme [Kanehisa et al., 2016], and ExPASy Enzyme [Bairoch, 2000].

UniProt General Annotations (Comments)

A large amount of useful biological information is available in the "Comments section" of the UniProt protein entries [Apweiler et al., 2010]. These annotations which are mostly biological knowledge are regularly added as a free text, however, UniProt is inclined to standardize them more and more using an in-house controlled vocabulary. There are more than 26 types of general comments introduced by UniProt. Following is the list of General annotations which are possible to be used for automatic prediction systems:

- **Function**: A general function of a protein.
- **Catalytic activity**: A reaction catalyzed by an enzyme.
- **Cofactor**: Non-protein substance needed for an enzymatic activity.
- **Subunit**: The protein quaternary structure and interaction(s) with other proteins.
- **Pathway**: Associated metabolic pathways.
- **Subcellular location**: Subcellular location of the mature protein.
- **Similarity**: The sequence similarity with other proteins and family.
- **Interaction**: Interaction with other proteins.

Additional information is available at http://www.uniprot.org/help/general_annotation/

1.2.3 Protein Domains and Families

Based on the structure or sequence similarity, proteins are grouped into categories. These categories mostly include proteins which are functionally characterized. Therefore, for a newly discovered protein, its functional characteristics can be identified according to the category in which it belongs. Although, these categories such as domains and families are broadly used in different biological contexts, their definitions usually vary in the view of each source. Protein family is a group of proteins whose evolutionary origin is the same. Proteins in a same family share similar functions in addition to similarities in their sequences or structures. Protein families are often hierarchically organized. If a protein shares a common ancestor with another bunch of proteins, they constitute a smaller and narrower related group. Superfamily and subfamily concepts are used in some classifications and mean a large group of distantly related proteins and a smaller group of closely related proteins, respectively. Figure 1.13 (top) displays a hypothetical family hierarchy in proteins and Figure 1.13 (bottom) illustrates that with the GPCR hierarchy. This figure shows the usefulness of protein family because of the amount of specific functional information that we can infer from hierarchy.

In computer science perspective, a domain is an abstract class that possesses several instances which are parts of particular proteins. In biology, domains are conserved functional and structural regions of a given protein that can evolve and exist independently in various proteins. They are structural and sequential building blocks of proteins. They are generally responsible for a specific function (or interaction) of their proteins. Domains often form functional units, however, similar domains can be found in proteins with different functions [Richardson, 1981]. Moreover, several domains can work together to create a multi-domain and multi-functional protein with a vast number of possibilities [Chothia, 1992]. In a multi-domain protein, each domain may carry out its individual function, or interact with its neighbors to fulfill a collective function. For example, Src homology 3 (SH3) domains are small domains with nearly 50 amino acids. Figure 1.14 shows its 3D structure. They occur in various proteins involved in diverse functions, like phosphatidylinositol 3-kinases, myosins, adaptor proteins, and phospholipases.

Domains are often grouped into domain families (or families) because of the very similar domains found in distinct proteins. Families can be considered as other classes whose instances are similar domains described in various proteins. However, it should be clarified here that most domain classifications use the same term, domain, to designate either the domain family or some instances of it (This will happen also in this thesis).

Classifications based on family or domain are always overlapping, because proteins are occasionally assigned to families based on domains they contain. For example Regulator of G-protein signalling (RGS) domains are building blocks of proteins that trigger GTPases function and belong to the RGS protein family. A RGS domain is present in all the RGS protein family members with the difference that some RGS proteins such as RGS1 are single domain whereas others such as RGS6 are multi-domains. RGS domains are also detected in several proteins from families other than RGS family such as axins and beta-adrenergic receptor kinases. Domain combination and family groupings of some RGS proteins and beta-adrenergic receptor kinase is depicted in Figure 1.15 below.

There are different classifications of protein domains and families. These classifications discover domains and families in automatic or semi-automatic fashion, mainly using Hidden Markov Model [Eddy, 1998] and multiple sequence alignment [Thompson et al., 1994] techniques. These classifications can be categorized based on their biological entities or their predictive models known as protein signatures. Figure 1.16 shows an overview of the well-known domain classifications and their categorization based on their biological entities and signature methods.

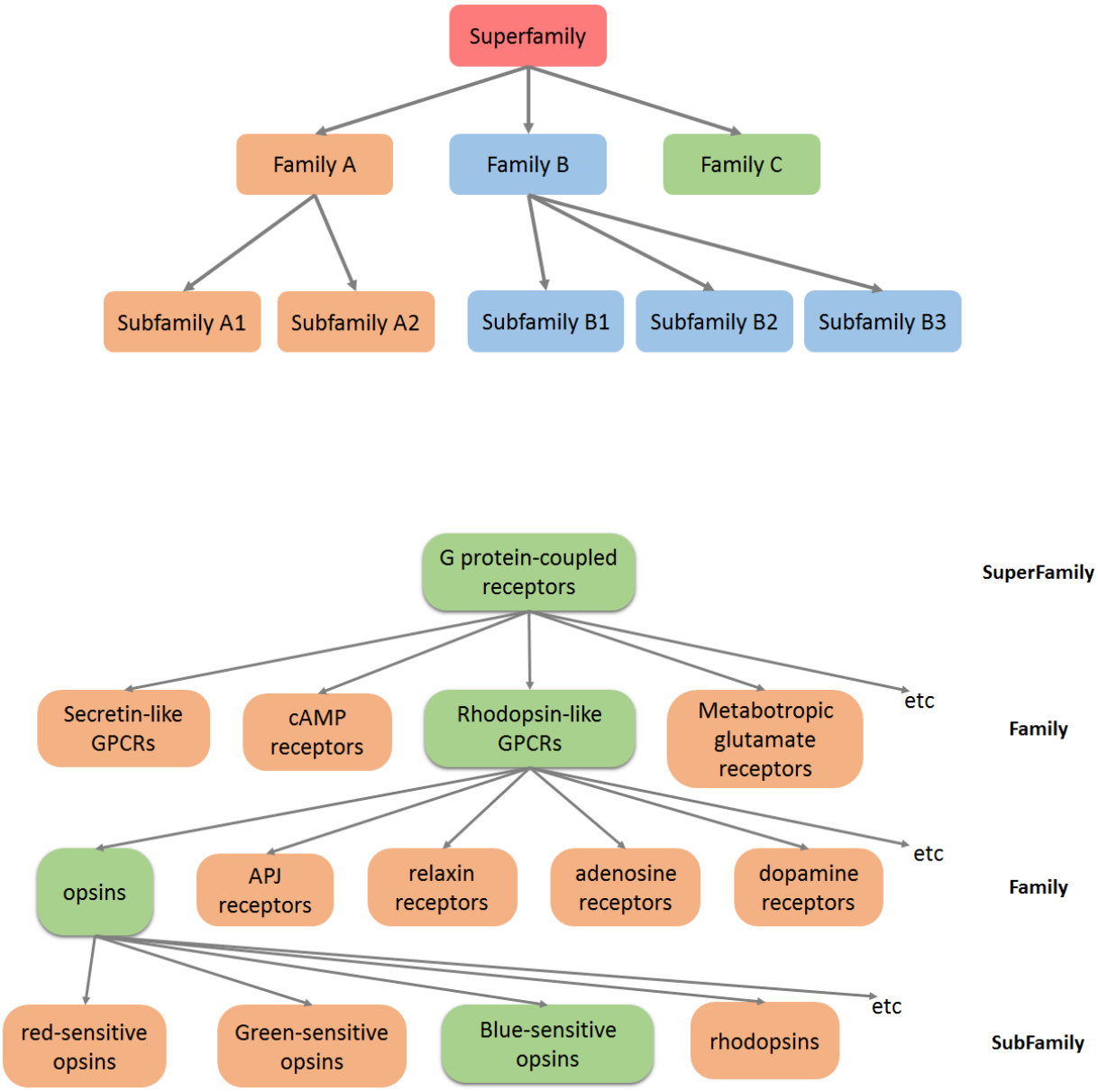


Figure 1.13: A hierarchy of superfamily, family and subfamily in protein. **Top**: This hierarchy in protein family expressing the relationships between superfamily, family and subfamily. Direction in the relation suggests a group is a subgroup of another group. **Bottom**: The GPCR hierarchy that highlights short-wave-sensitive opsins protein in green.

Figure 1.14: 3D structure of SH3 domain which is a component in several distinct proteins with different functions.

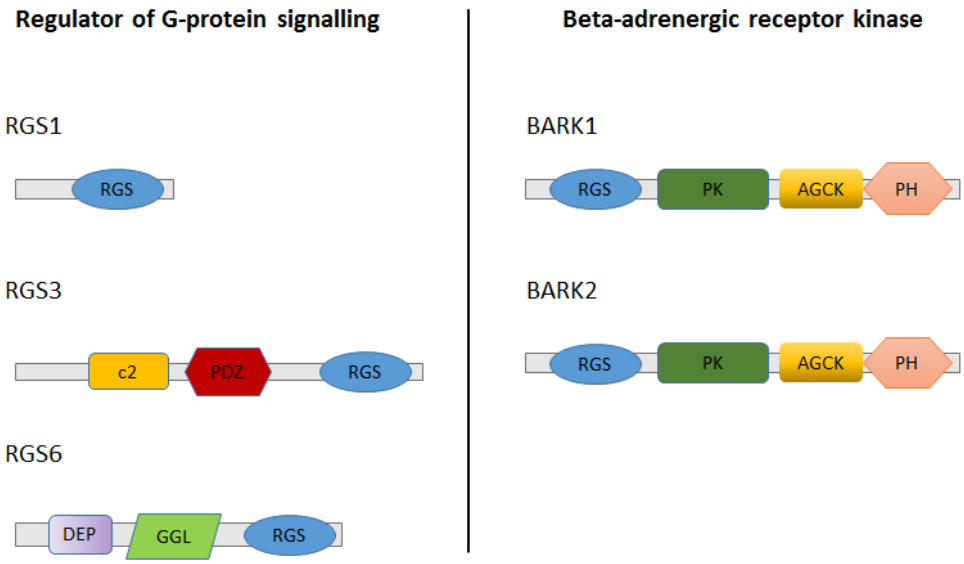


Figure 1.15: **Left**: Domain combination of RGS1 and RGS6 proteins from RGS protein family. **Right**: RGS domains in proteins from beta-adrenergic receptor kinases family.

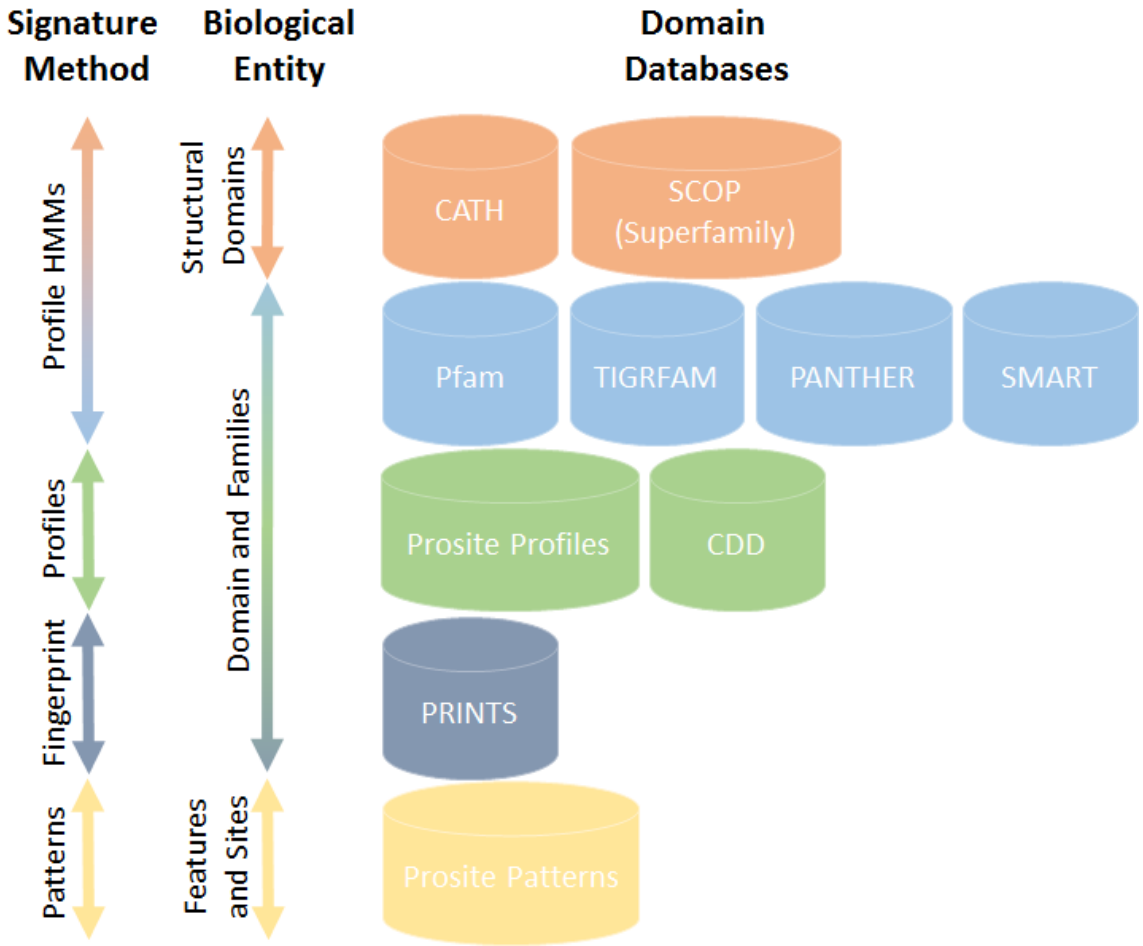


Figure 1.16: Overview of different domain classifications. Signature methods are divided into hidden Markov models (HMMs), profiles, fingerprints, and patterns. HMMs are powerful statistical models that convert multiple sequence alignments into position-specific scoring systems by modeling insertions and deletions. Profiles are constructed by converting conserved motifs from multiple sequence alignments into position-specific scoring systems (PSSMs). Fingerprints are generated using multiple profiles. Patterns are created by building regular expressions from identified conserved motifs in multiple sequence alignments.

In the following, the most commonly used protein domain and family classifications are introduced.

Gene3D:CATH

The CATH Protein Classification database [Orengo et al., 1997, Pearl et al., 2003] is a structural classification of domains which provides information on the evolutionary relationships of protein domains. CATH has many wide specifications in common with the SCOP database. Nevertheless, there are many details in which these structural classifications differ greatly [Hadley and Jones, 1999]. Experimentally-identified 3D structures of proteins are acquired from the PDB. If applicable, these structures are split into their successive polypeptide chains. Using a combination of several automatic and manual methods, protein domains are determined within these PDB chains. Then, protein domains are classified into the CATH structural hierarchy. The four principle levels of the CATH hierarchy are:

- **Class** is the type of the secondary-structure content of the domain.
- **Architecture** is high structural similarity without confirming homology.
- **Topology/fold** is categorization of topologies which have certain structure properties in common.
- **Homologous superfamily** is grouping based on evolutionary relationship.

It is worth it to mention that Class level in CATH and SCOP classifications are equivalent, while Architecture and Homologous superfamily in CATH are the counterparts of Fold and Superfamily in SCOP, respectively. CATH database is available at http://www.cathdb.info/

SCOP (Superfamily)

The Structural Classification Of Proteins (SCOP) database [Murzin et al., 1995] is a manual classification of structural domains of proteins based on the structure and sequence similarities. The overall goal of this classification is to specify proteins which are evolutionarily related. The unit of structural classification in SCOP is the protein domain. Based on a definition suggested in SCOP, small and most medium-sized proteins have only one domain while by the observation two SCOP domains are assigned for the human hemoglobin, one for the α and one for the β subunit.

The levels of SCOP are described as:

- **Class** is the types of fold, for example, beta sheet.
- **Fold** is the different forms of domains within a class.
- **Superfamily** distinguishes groups of domains within a fold, on the basis of a sometimes hypothetical distant common ancestor.
- **Family** distinguishes groups of domains within a superfamily on the basis of a more recent common ancestor.
- **Protein domain** is a group of domains within a family.
- **Species** is a group of domains in protein domains based on species.
- **Domain** is the smallest level (unit) of this classification.

SCOP database is available at http://scop.mrc-lmb.cam.ac.uk/scop/. SCOP stopped updating in 2010 but a successor called SCOP2 has been introduced [Andreeva et al., 2013]. SCOP2 similarly focuses on structurally characterized proteins in the PDB and structural and evolutionary relationships of proteins. SCOP2 also establishes a complex network of nodes instead of a tree-like hierarchy. Each node in the network demonstrates a specific relationship and is represented by a structural and sequential region of protein. SCOP2 database is available at http://scop2.mrc-lmb.cam.ac.uk/

Pfam

Pfam is a database of protein domain and family classification generated using multiple sequence alignments and hidden Markov models [Bateman et al., 2002, Finn et al., 2016b]. The main motivation of Pfam is to provide general and complete classification of protein domains and families [Sammut et al., 2008]. The Pfam domains and families are extensively used by researchers due to its broad coverage of proteins and realistic way of naming domains [Xu and Dunbrack Jr, 2012]. For example, Pfam was used for functional annotation of genomic data in the human genome project. Pfam has also been utilized as the basis of protein-protein interaction resources such as iPfam [Finn et al., 2013] and 3did [Stein et al., 2005]. Pfam entry types are as follows:.

- **Family** is the default type, defines that members of the family are related.
- **Domain** is described as an independent structural or sequential unit found in multiple proteins.
- **Repeat** is another type of Pfam entries which is not independently stable. Repeats are usually required to be combined to create tandem repeats in order to form a domain.
- **Motif**s, unlike Repeats, are usually shorter sequence units which are stable in isolation and found outside of globular domains.

The recent version of Pfam database is 31.0 and it was released in March 2017. It contains 16,712 domains and families so that around 76% of protein sequences in UniprotKB matched to at least one Pfam. Pfam database is available at http://pfam.xfam.org/

TIGRFAMs

TIGRFAMs is a database of manually curated protein families designed to support both manual and automated curated genome annotation [Haft et al., 2003, Haft et al., 2012]. TIGRFAMs entries consist of multiple sequence alignments and hidden Markov models. TIGRFAMs have models of full-length or small regions of proteins at three levels which are listed below.

- **Superfamily** is the complete set of proteins having homology over essentially their whole length.
- **Subfamily** is grouping based on distinct clade (a group of organisms evolved from a common ancestor) within a superfamily.
- **Equivalog** is sets of homologous proteins conserved in function since their last common ancestor.

The objective of this classification is to provide domains possessing maximum utility for the annotation purposes. Therefore, TIGRFAMs is a complementary collection to the Pfam, in which models widely cover across distant homologs but end at the boundaries of conserved structural domains. Figure 1.17 shows one striking difference between TIGRFAMs and Pfam.

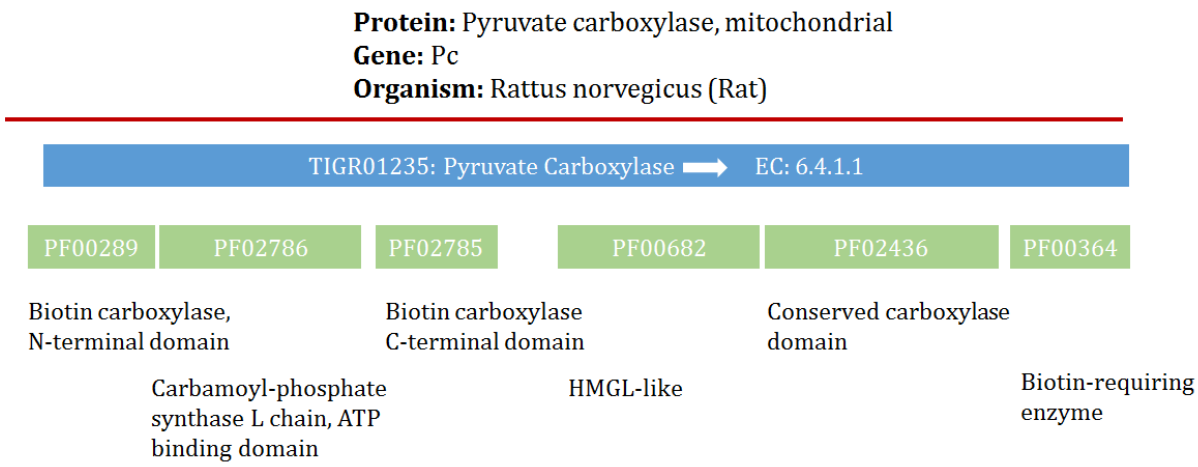


Figure 1.17: Six Pfam domains are covered with one TIGRFAMs entry ($\Sigma Pfam_i = TIGRFAMs_j$). The red line represents the protein sequence and blue and green boxes represent regions for different HMMs hits by TIGRFAMs and Pfam, respectively. Larger number of Pfam domains compared to the TIGRFAMs implies that Pfam domains are spread among various sequences and can also be found in shorter sequences.

Six separate domains from Pfam illustrate the architecture of the "rat pyruvate decarboxylase". However, they are not singly responsible for the function (or a full name) of the protein. In contrary with each of these Pfam domains which describe regions shared by proteins with various functions, an individual equivalog model of TIGRFAMs provides annotation for the whole protein. TIGRFAMs is available at http://www.jcvi.org/cgi-bin/tigrfams/index.cgi.

PANTHER

Protein ANalysis THrough Evolutionary Relationships (PANTHER) classification system [Thomas et al., 2003, Mi et al., 2017] is a manually curated biological database of protein families and their functional annotations. PANTHER families can be utilized identifying the function of proteins, ontology, and pathways. In PANTHER, proteins are classified based on different attributes such as families and subfamilies, Gene Ontology, and pathways. The most substantial feature of PANTHER is to infer functions of uncharacterized proteins based on their evolutionary relationships to protein with known functions. For each protein family in PANTHER, there is a phylogenetic tree. Using this phylogenetic model, PANTHER is able to predict the functions of an uncharacterized protein through inheritance from its ancestors in its tree. [Mi et al., 2017]. PANTHER database is available at http://pantherdb.org/.

SMART

Simple Modular Architecture Research Tool (SMART) database is a biological resource to detect and annotate domains and their architecture within protein sequences [Schultz et al., 1998, Letunic et al., 2014]. SMART like many other domain databases uses hidden Markov models built from multiple sequence alignments to identify protein domains. Data from SMART has been used to create the Conserved Domain Database (CDD) collection. SMART is available at http://smart.embl.de/.

CDD

The Conserved Domain Database (CDD) [Marchler-Bauer et al., 2005] provides annotation of protein sequences using the location of conserved domains as footprints. These footprints are then used to infer the functions sites in protein sequences. CDD combines several protein domain and full-length protein model collections, and maintains an active curation effort that aims at providing fine grained classifications for major and well-characterized protein domain families, as supported by available protein three-dimensional (3D) structure and the published literature. So far, the majority of protein three-dimensional structures are represented by models tracked by CDD, and CDD curators are characterizing novel families that emerge from protein structure determination projects. CDD is accessible via http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml.

PROSITE

PROSITE is a database [Hulo et al., 2007, Sigrist et al., 2002] that consists of entries describing the protein domains and families. It also contains functional sites and amino acid patterns and profiles. PROSITE entries are manually curated and then integrated into the UniProtKB/SwissProt database. PROSITE procedure identifies functions of recently discovered proteins and analyzes known proteins for functions which are formerly uncharacterized. It propagates properties of well-studied proteins to the proteins of biologically related organisms or predicts functions based on similarities for poorly know proteins [Hulo et al., 2007]. ProRule is another database builds on top of the domain descriptions in PROSITE [Sigrist et al., 2005]. It supplies further information about functionally critical amino acids. Such information can help creating automatic annotation based on PROSITE. PROSITE is available at http://prosite.expasy.org/.

PRINTS

PRINTS is a database of fingerprints [Attwood et al., 2003] that provides an annotation list for protein families as well as a diagnostic tool for newly discovered protein sequences. A fingerprint is a group of conserved motifs found by a multiple sequence alignment. The motifs create a special signature for the protein families which are aligned. The motifs mainly come together in three-dimension to determine interaction surfaces or binding sites in the molecules. The main strength of fingerprints is discerning differences in protein sequences at four levels of clan, superfamily, family and subfamily. This allows more accurate functional predictions for uncharacterized sequences. The database is accessible at http://www.bioinf.man.ac.uk/dbbrowser/PRINTS/.

InterPro

InterPro is a rich integrated collection of protein domains families as well as protein functional sites. It exploits features which are distinguishable in characterized proteins to apply on new protein sequences which are functionally unknown [Apweiler et al., 2001, Finn et al., 2016a].

The contents of InterPro include diagnostic signatures and proteins which are remarkably similar. The signatures are composed of several models such as regular expressions or hidden Markov models, which describe protein domains families or functional sites. Models are often constructed from amino acid sequences of characterized protein domains and families. Afterward, uncharacterized protein sequences (such as proteins introduced by new genome sequencing) are also aligned against models and distributed in the different classes. Three main entities: proteins, signatures (also known as “methods” or “models”) and

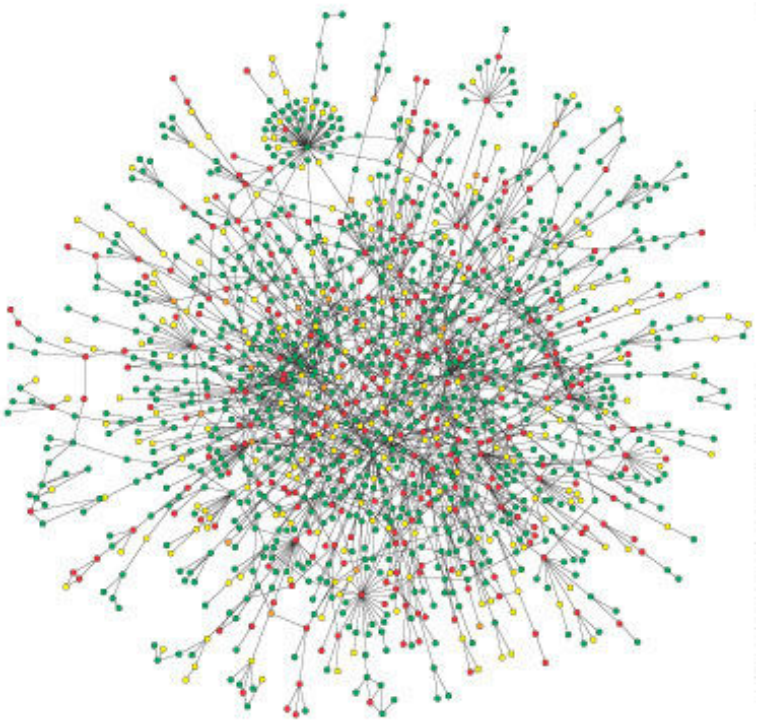


Figure 1.18: Yeast interactomes obtained using the yeast-two hybrid method [Jeong et al., 2001].

entries are stored in InterPro. The InterPro signatures are from member databases, the most important ones are listed below.

The main intention of InterPro is to integrate protein domain classifications. InterPro database stores all the signatures from different member databases into the InterPro entries (All domains classifications in Figure 1.16 are integrated into the InterPro database). Signatures from different domain databases corresponding to equivalent domains, families or functional sites are gathered into the same entry. Additionally, applicable information such as a description, consistent names, cross-reference to function ontologies like Gene Ontology (GO) are also associated with each InterPro entry.

InterProScan [Zdobnov and Apweiler, 2001, Jones et al., 2014] is a scanning software that searches the protein signatures of the above-mentioned member databases inside a given protein sequence. InterProScan and InterPro are accessible from https://www.ebi.ac.uk/interpro/interproscan.html/ and https://www.ebi.ac.uk/interpro/, respectively.

1.2.4 Protein Interaction

Protein-protein interactions (PPIs) exists in nearly all biological process in a single cell. Thus, knowing how proteins interact is a substantial study to recognize functions and behaviors of living organisms and their parts in normal and abnormal conditions. It is also crucial in the development of drugs due to the fact that drugs can influence PPIs. Protein-protein interaction networks (PPIN) illustrate the physical contacts amongst proteins in organisms mathematically. A sample map of protein-protein interactions of yeast is shown in Figure 1.18. These contacts occur between particular binding sites in the interacting proteins, and represent a specific biological meaning such as a determined function.

Interactions between proteins can be indicated as both stable and transient. On the one side, stable interactions are organized in protein complexes such as ribosome and haemoglobin while on the other side, transient interactions are short-time interactions that alter or transport a protein and lead to subsequent changes such as protein kinases and nuclear pore importins. Transient interactions contain majority of the dynamic sector of the interactome. Knowledge about PPIs can be applied to the function prediction of uncharacterized proteins, enhancement of the details about a signaling pathway, and characterization of the proteins relationships that establish poly-molecular complexes (e.g. proteasome).

Molecular interaction can be discovered by various methods and techniques. It is important to realize that they all have their own weaknesses and strengths and no individual methodology can precisely generate a complete list of protein-protein interactions. Information about protein interactions can be acquired using experimental or computational approaches. Interaction data discovered by experimental methods are generally more accurate than computationally predicted interactions. The most frequent experimental methods with the striking contribution to the growing of PPIs are Yeast Two-Hybrid [Brückner et al., 2009], Affinity Purification Mass Spectrometry [Bauer and Kuster, 2003], Protein-fragment Complementation Assays (PCA) [Morell et al., 2009], Co-immunoprecipitation (Co-IP) [Isono and Schwechheimer, 2010], X-ray crystallography [Kobe et al., 2008], and Fluorescence Resonance Energy Transfer (FRET) [Kenworthy, 2001]. Nonetheless, these techniques are time-consuming and expensive in terms of money and manpower. Therefore, experimental methods furnish only a small part of the available interactions data [Pitre et al., 2008, Valencia and Pazos, 2002]. Moreover, in the same organism, there are considerable discrepancies between the PPI data acquired by the same or different methods. All these issues encourage the emergence of computational techniques for PPI prediction.

Protein-protein interaction prediction using computational methods uses the combination of structural biology and bioinformatics to find physical interactions between proteins. Computationally predicted interactions have an important role in completing the list of experimental interactions. Similar to the experimental discovery of protein interactions, computationally predicted interactions can be used to gain insights into intracellular signaling pathways and protein complex structures.

Protein-protein interactions can be studied at the domain level. In general, protein interactions occur through their domains instead of the entire protein molecules [González and Liao, 2010]. Protein domains interact physically with other protein domains to carry out the functions which their corresponding proteins are supposed to perform [Deng et al., 2002]. Thus, understanding protein-protein interactions at the level of domain gives a better view of the protein functions and the protein interaction network. Protein-protein interaction at the domain level is called domain-domain interaction (DDI). There are a variety of approaches to infer DDIs which are listed below. It also should be noted that interaction between proteins can be predicted using the discovered DDIs. In the following, four databases of observed and predicted DDIs in addition to six databases of the most commonly used PPIs are briefly introduced.

3did

Database of three-dimensional interacting domains (3did) is a database of protein-protein interactions with a known three-dimensional structure [Stein et al., 2005, Stein et al., 2010]. 3did uses the Pfam protein domain and family classification for identifying protein domains inside the protein structures. It classifies all possible DDIs models in the PDB database and adds molecular characterizations to each DDI. Recently, 3did clusters similar interfaces into a group in order to include annotations [Mosca et al., 2013]. 3did is available for download and browsing at http://3did.irbbarcelona.org.

KBDOCK

KBDOCK provides a three-dimensional biological database which systematically identifies and spatially clusters proteins binding sites for template-based (knowledge-based) protein docking [Ghoorah et al., 2013a]. KBDOCK incorporates the Pfam protein domain and family classification with their structures from PDB coordinate data in order to investigate the arrangements of DDIs in the three-dimensional space. This procedure ends with a set of structural templates for protein docking [Ghoorah et al., 2011, Ghoorah et al., 2013b]. KBDOCK database is accessible for downloading and querying at http://kbdock.loria.fr/.

DOMINE

DOMINE provides a database of predicted and observed domain-domain interactions amassed from diverse sources [Raghavachari et al., 2007, Yellaboina et al., 2010]. DOMINE accommodates DDIs observed in the PDB database, as well as predicted DDIs from eight computational methods. This database serves as a reference and robust dataset of DDIs for testing new methods of predicting protein and domain interactions and for analysis of the topological structure of interaction networks. DOMINE is accessible at http://domine.utdallas.edu.

INstruct

INstruct is a three-dimensional database that structurally identifies protein interaction networks in human and six model organisms [Meyer et al., 2013]. INstruct incorporates available protein-protein interactions with atomic-resolution information derived from co-crystal structures. Its web interface is designed to allow for flexible search based on standard and organism-specific protein and gene-naming conventions, visualization of protein architecture highlighting interaction interfaces and viewing and downloading custom 3D structurally resolved interactome datasets. INstruct is available for viewing and downloading at http://instruct.yulab.org.

IntAct

IntAct is a protein interaction software and database which houses protein-protein interactions models and their analysis [Kerrien et al., 2006, Kerrien et al., 2011]. In the IntAct resource, data are accumulated from peer-reviewed journals and are manually annotated by expert curators. In its website, protein interactions are textually represented and graphically visualized for protein interaction networks. It also provides additional information for the interacting proteins such as GO annotations and pathways. IntAct data and software are available at http://www.ebi.ac.uk/intact.

MINT

The Molecular INTeraction database (MINT) [Zanzoni et al., 2002, Licata et al., 2011] is a repository of protein-protein interactions curated from experimental details of biomedical literature. MINT prepares the curation work on physical interactions between proteins and does not include any genetic or computationally inferred interactions. Interaction data alongside the annotations are explorable in the MINT website. The dataset can be accessed online at http://mint.bio.uniroma2.it/mint/.

DIP

The Database of Interacting Proteins (DIP) is a biological database that aims to maintain protein-protein interactions which are experimentally determined [Xenarios et al., 2000]. This database is intended to provide a comprehensive and integrated tool for browsing and efficiently extracting information about protein interactions and interaction networks [Xenarios et al., 2002]. Researchers are able to analyze, visualize and integrate their experimental data with the protein interacting information in the DIP database using the DIP tools. Moreover, the DIP database is beneficial for studying the features and relationships in protein interaction networks, benchmarking predictions of protein-protein interactions and studying the evolvement of protein–protein interactions. The database is accessible at http://dip.doe-mbi.ucla.edu.

BioGRID

The Biological General Repository for Interaction Datasets (BioGRID) is a database aimed to store genetic and protein interactions extracted from the primary published scientific literature [Stark et al., 2006]. These interactions are obtained for all major model organism species and humans. Interaction Management System (IMS) organizes curation in BioGRID. This system facilitates the compilation of interaction entries through gene annotation, phenotype ontologies, and structured evidence codes. The BioGRID architecture supports the representation of more complex multi-gene or multi-protein interactions to account for cellular phenotypes via structured ontologies [Oughtred et al., 2016]. BioGRID is available at http://thebiogrid.org/.

HPRD

Human Protein Reference Database (HPRD) is a database of manually curated proteomic information only for human proteins [Peri et al., 2003]. HPRD database has detailed information regarding to different facets of human proteins such as protein interactions and post-translational modifications obtained from manual investigation of literature as well as analyses of protein sequences. HPRD resource can be accessed at http://www.hprd.org/.

STRING

The STRING database is a repository of interacting proteins which collects and integrates functional interactions between proteins, by combining predicted and known protein-protein association data for many diverse organisms. Protein associations in STRING comprise two types of physical and functional interactions and are collected from experimental data from curated databases as well as predicted protein interactions. Predicted interactions are extracted from identification of shared signals among genomes, co-expression analysis in a systematic way, automatic text-mining on the biomedical literature, transferring interaction knowledge across organisms computationally. An interesting scoring system in STRING allows users to gather and categorize the most reliable interactions having a score greater than a desired threshold. The STRING database is accessible at http://string-db.org/.

In this chapter, a rapid overview of the computational and biological backgrounds of this thesis has been presented. Very important research challenges take place at the crossing of these two domains: data mining and knowledge discovery on the one hand, protein structure, function and interactions on the other hand. In particular, the huge amount and complexity of biological data accumulating in databases today makes it necessary to develop knowledge-based computational approaches to make sense of these data and facilitate their use for various applied purposes in biology and health. The next four chapters will describe four applications based on two major computational approaches (CODAC and CARDM) that constitute the contribution of this thesis to this interdisciplinary field.

Chapter 2

Discovering Hidden Associations between Enzyme Commission Numbers and Pfam Domains

Contents

Many entries in the protein Data Bank (PDB) and UniProtKB are annotated to show their component protein domains according to the Pfam classification, as well as their biological function through the enzyme commission (EC) numbering scheme. However, despite the fact that the biological activity of many proteins often arises from specific domain-domain and domain-ligand interactions, current on-line resources rarely provide an explicit relationship between individual EC numbers and Pfam domains. Since the PDB now contains many tens of thousands of protein chains, and since protein sequence databases can dwarf such numbers by orders of magnitude, there is a pressing need to develop automatic method to find direct mapping between EC numbers and Pfam domains.

2.1 Introduction

Proteins perform many essential biological functions such as catalysing metabolic reactions and mediating signals between cells. These functions are often carried out by distinct "domains", which may be identified as highly conserved regions within a multiple alignment of a group of similar protein sequences, as in the Pfam classification [Finn et al., 2016b]. It is widely accepted that such protein domains often correspond to distinct and stable three-dimensional (3D) structures, and that there is often a close relationship between protein structure and protein function [Berg et al., 2002]. Indeed, it is well known that protein structures are often more highly conserved than protein sequences [Chothia and Lesk, 1986], and this suggests that proteins with similar structures will have similar biological functions [Martin et al., 1998]. The Protein Data Bank (PDB) [Bernstein et al., 1977, Gutmanas et al., 2014] now contains over 107,000 3D structures, most of which have been solved by X-ray crystallography or NMR spectroscopy.

As well as sequence-based and structure-based classifications, proteins may also be classified according to their function. For example, the Enzyme Commission [Webb et al., 1992] uses a hierarchical four-digit numbering system to classify the enzymatic function of many proteins. The first digit, or top-level "branch" of the hierarchy, selects one of six principal enzyme classes (oxidoreductase, transferase, hydrolase, lyase, isomerase, and ligase). The second digit defines a general enzyme class (chemical substrate type). The third digit defines a more specific enzyme-substrate class (e.g. to distinguish methyl transferase from formyl transferase), while the fourth digit, if present, defines a particular enzyme substrate. However, it should be noted that because EC numbers are assigned according to the reaction catalyzed, it is possible for different proteins to be assigned the same EC number even if they have no sequence similarity or if they belong to different structural families.

Furthermore, there are several ways in which a protein might provide one or more enzymatic functions, as illustrated in Figure 2.1. In the simplest case (Figure 2.1 (A)), a protein contains just one domain, and there is a one-to-one association between that domain and a particular enzymatic function. In this case, it is reasonable to suppose that the catalytic site is located entirely on that domain. Similarly, a protein may have two or more distinct domains, each of which provides a distinct enzymatic (or non-enzymatic) function (Figure 2.1 (B)). On the other hand, a protein domain could be involved in more than one catalytic activity, as illustrated in Figure 2.1 (C). Finally, a catalytic site may be at the interface between two domains, or one domain serves as a necessary co-factor for the other (Figure 2.1 (D)). It is biologically relevant to be able to distinguish all such cases. However, except for the simplest case (Figure 2.1 (A)), it can be seen that finding domain-EC associations automatically is a non-trivial task. Several groups have described approaches or resources that can associate entire PDB protein chains with enzyme EC numbers [Reichert et al., 2000, de Beer et al., 2014, Laskowski, 2001, Martin, 2004]. Probably the most up-to-date and exhaustive association between PDB chains and EC numbers is provided by SIFTS [Velankar et al., 2012], which is a collaboration between the Protein Data Bank in Europe and UniProt [Apweiler et al., 2010]. SIFTS incorporates a semi-automated procedure which links PDB chain entries to external biological resources such as Pfam, and IntEnz [Fleischmann et al., 2004].

While all of the above mentioned approaches can provide associations between PDB protein chains and enzyme EC numbers, to our knowledge, very few approaches have been published for automatically assigning EC numbers to structural domains. SCOPEC [George et al., 2004] uses sequence information from SwissProt and PDB entries that have been previously annotated with EC numbers in order to assign EC numbers to SCOP domains [Murzin et al., 1995]. It first looks for PDB chains that fully map to SwissProt entries (to within up to 70 residues) and that match on at least the first three EC number digits. In this way, SCOPEC identifies single domain structures that can be associated unambiguously

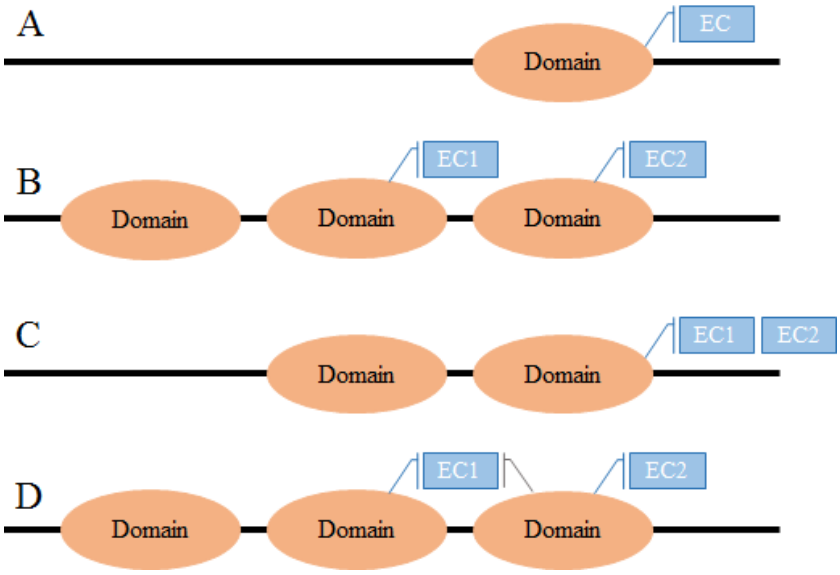


Figure 2.1: A graphical representation of different situations of EC-Domain association in a protein sequence or structure.

with an EC number. Although SCOPEC can subsequently propagate a known EC-domain association to a matching domain in a multi-domain protein, it is generally not able to resolve cases where multiple ECs are associated with multi-domain chains (parts B, C, and D in Figure 2.1).

Furthermore, it appears that the SCOPEC database is no longer available on-line.

In contrast, the dcGO ontology database for protein domains produced in 2012 is still available online and provides several ontological annotations (Gene Ontology: GO, EC, pathways, phenotype, anatomy and disease ontologies) for more than 2,000 SCOP domain families [Fang and Gough, 2013].

The dcGO approach follows the principle that if a GO term tends to be attached to proteins in UniProtKB that contain a certain domain, then that term should be associated with that domain. The statistical significance of an association is assessed against a random chance association using a hypergeometric distribution followed by multiple hypotheses testing in terms of false discovery rate. The dcGO approach addresses the issues of hierarchical structure of most biological ontologies and the nature of domain composition for multi-domain proteins. However, a mapping onto Pfam domains is proposed only for GO terms. Here, we describe a recommender-based approach call "ECDomainMiner" for associating Pfam domains with EC numbers, which builds on our previously described statistical approach [Alborzi et al., 2015]. Recommender systems are a class of information filtering system [Hanani et al., 2001, Ricci et al., 2011] which aim to present a list of items that might be of interest to an on-line customer. There are two main kinds of recommender systems. Collaborative filtering approaches make associations by calculating the similarity between activities of users [Sarwar et al., 2001, Koren and Bell, 2015]. Content-based filtering aims to predict associations between user profiles and description of items by identifying common attributes [Robillard et al., 2014, Ricci et al., 2011]. Such an approach has recently been applied to a quite different problem of discovering novel cancer drug combinations [Huang et al., 2014]. Here, we use content-based filtering to associate EC numbers with Pfam domains from existing EC-chain and Pfam-chain associations from SIFTS, and from EC-sequence and Pfam-sequence associations from SwissProt and TrEMBL, where protein chains and sequences serve as the common attributes through which

EC-Pfam associations are made. Note that our approach *does not* attempt to identify catalytic sites or catalytic residues. Rather, we aim to detect frequent co-ocurrences of Pfam domains and EC numbers in order to deconvolute the often complex EC-Pfam relationships within multi-domain and multi-function protein chains. We assess the performance of our approach against a "Gold Standard" dataset derived from InterPro [Finn et al., 2016a], and we compare our results with the Pfam-EC associations derived from the dcGO database. We also show how our database of more than 20,000 EC-Pfam associations can be exploited for automatic annotation purposes.

2.2 Methods and Materials

2.2.1 Data Preparation

Our data sources are SIFTS for EC number and Pfam domain annotations of PDB chains, and Uniprot for EC number and Pfam domain annotations of protein sequences. UniProt is divided into three parts: (i) the non-redundant, high quality, manually curated SwissProt part, (ii) the TrEMBL data that are annotated using Unified Rules [Pedruzzi et al., 2013], called here UniRule, and (iii) the rest called here TrEMBL.

In addition, in order to parameterize and evaluate ECDomainMiner, we use the InterPro database [Finn et al., 2016a] which contains a large number of manually curated EC-Pfam associations. Flat data files of SIFTS (July 2015), Uniprot (July 2015), and InterPro (version 53.0) were downloaded and parsed using in-house Python scripts. From the SIFTS data, associations between EC numbers and PDB chains, and associations between PDB chains and Pfam domains were extracted. Associations between Uniprot sequence accession numbers (ANs) and EC numbers, and AN-Pfam associations were then extracted from the SwissProt section of Uniprot to give a dataset of Swissprot associations. For the TrEMBL entries, we collected and stored the corresponding AN-EC and AN-Pfam associations which had been annotated by UniRule, and those associations lacking UniRule annotations to give two further sequence-based datasets of associations, which we call the UniRule and TrEMBL association datasets.

To avoid bias due to duplicate structures or sequences in the four source datasets, all PDB chains and Uniprot sequences were grouped into clusters having 100% sequence identity using the Uniref non-redundant cluster annotations [Suzek et al., 2007], and each cluster was assigned a chain unique identifier (CID). Note that since just a few point mutations can dramatically change an enzyme's substrate specificity, making clusters of identical rather than highly similar sequences avoids the risk of falsely clustering proteins that share highly similar folds but which have quite different substrates. Moreover, Pfam families display highly conserved residues in their amino acid signature. Clustering sequences according to sequence similarity less than our strict condition may create a chance of mutating the conserved residues' in Pfam families and lead to incorrect mapping. The source EC-chain and EC-AN associations were then mapped to the corresponding CID in order to make four sets of EC-CID associations. A similar mapping was applied to the source Pfam-chain and Pfam-AN associations to give four sets of Pfam-CID associations.

For the reference data, we extracted from InterPro a total of 1,515 EC-Pfam associations in which each EC number had all four digits and each Pfam accession code (AC) referred either to a Pfam domain or a Pfam family (i.e. Pfam motifs and repeats were excluded). These associations were considered to be "positive examples", and were randomly divided into two equal "training" and "test" subsets. However, for training purposes, we also needed some "negative examples". We therefore created a set of "false" EC-Pfam associations by first shuffling the CID-EC and CID-Pfam associations from SIFTS dataset, and

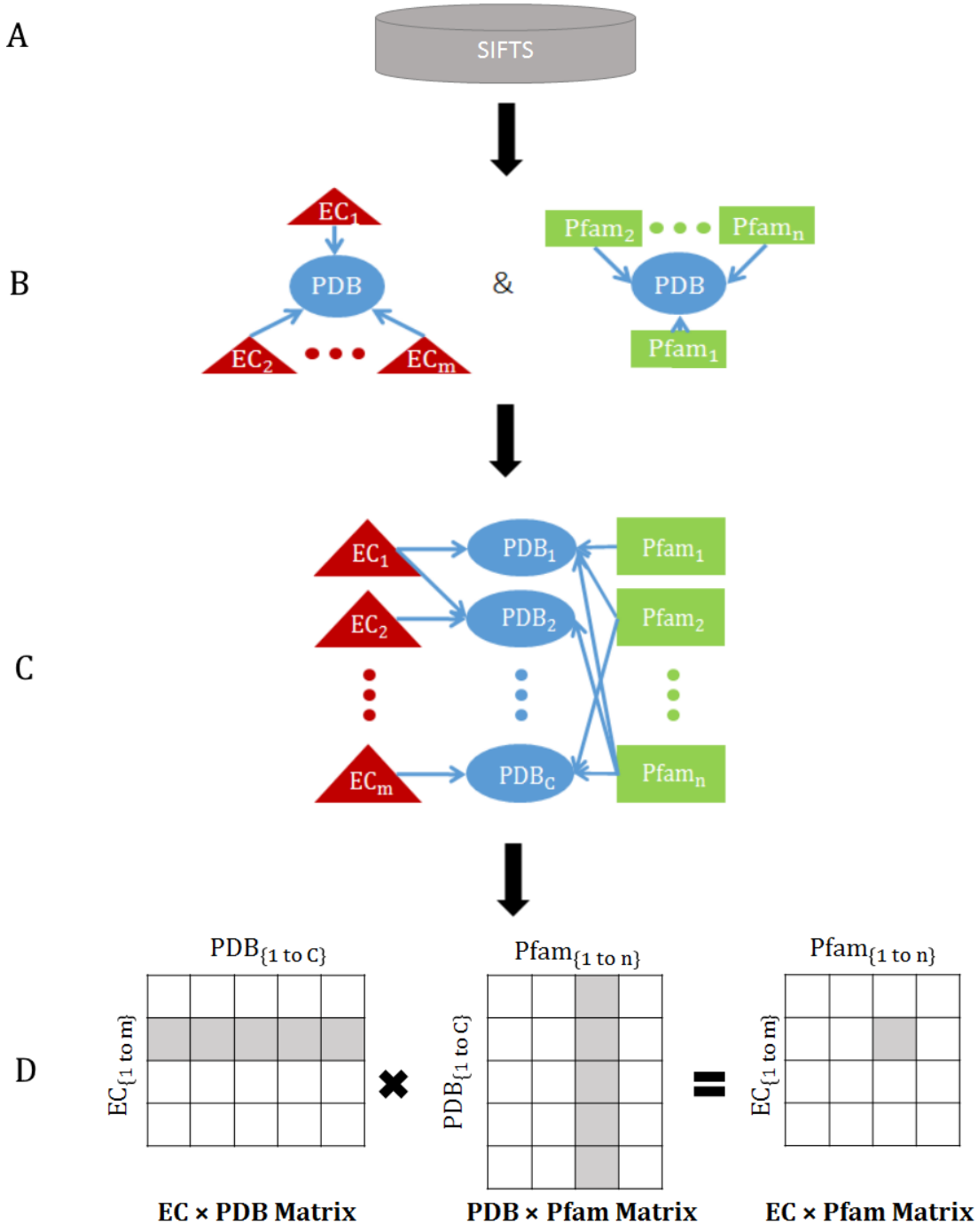


Figure 2.2: A graphical illustration of calculating raw EC-Pfam association scores from existing SIFTS EC-CID and Pfam-CID associations.

by then randomly collecting 1,515 wrong EC-Pfam associations from the shuffled datasets. In the rest of this article, we will refer to the combined set of 758 randomly chosen positive examples from InterPro and 758 randomly chosen negative examples as our "training dataset" and the remaining 1,513 positive and negative examples as our "test dataset".

2.2.2 Inferring EC-Pfam Domain Associations

The main idea underlying the discovery of hidden EC-Pfam associations is to represent EC numbers and Pfam domains as feature vectors, with one feature per PDB or UniProt CID, and to score any inferred EC-Pfam association with the cosine similarity between its EC and Pfam vectors.

The various steps of our content-based filter approach for finding associations between 4-digit EC numbers and Pfam domains are illustrated in Figure 2.2 for the SIFTS dataset. First, all relations between PDB CIDs and EC numbers, and between PDB CIDs and Pfam domains are extracted from SIFTS, as described above. Joining these two lists of relations then yields a complex many-to-many graph that contains relations between EC numbers, PDB CIDs, and Pfam domains.

After this join operation, all EC-CID relations are encoded in a binary matrix, where a 1 represents the presence of an association and a 0 represents no association. This matrix is then row-normalized such that each row has unit magnitude when considered as a vector. Similarly, all PDB CID-Pfam relations are encoded in a second binary matrix which is column-normalized. Consequently, the product of the two normalized matrices corresponds to a matrix of cosine similarity scores between the rows of the first matrix and the columns of the second matrix. Thus, each element, $S(ec, d)$, of the product matrix represents a raw association score between an EC number, ec, and a Pfam domain, d.

Similarly, raw EC-Pfam association scores are calculated from EC-CID and Pfam-CID relations extracted from SwissProt, TrEMBL and Unirule. Then, because we wish to draw upon the relations from all four datasets, we combine the four raw scores as a weighted average to give a single normalized confidence score, $CS_{ec,d}$:

$$CS_{ec,d} = \frac{\sum_i w_i S_i(ec, d)}{\sum_i w_i} \tag{2.1}$$

where $i \in \{SIFTS, Swissprot, TrEMBL, UniRule\}$ enumerates the four datasets, w_i are weight factors, to be determined, and where an individual association score, $S_i(ec, d)$, is set to zero whenever there is no data for the (ec, d) pair in dataset i.

In order to find the best values for the four weight factors, receiver-operator-characteristic (ROC) curves [Fawcett, 2006] were calculated using the positive examples of our Interpro-based training dataset, against the rest of associations (background associations).

Each weight was varied from 0.0 to 1.0 in steps of 0.1, and for each combination of weights a ROC curve of the ranked association scores was calculated. The combination of weights that gave the largest area under the curve (AUC) of the ROC curve was selected.

2.2.3 Defining a Confidence Score Threshold

Having determined the best weight for each data source, we next wished to determine an overall threshold for the confidence score. To do this in an objective way, we used the training dataset, then scored and ranked the members of the dataset, and labeled them true or false according to a threshold value that was varied from 0.0 to 1.0 in steps of 0.01. For each threshold value, we counted the number of positive examples above the threshold (TPs), negative examples above the threshold (FPs), negative examples

below the threshold (TNs), and positive examples below the threshold (FNs). We then calculated the recall, R, precision, P, and their harmonic mean in order to obtain a "F-measure" using:

$$R = \frac{TP}{TP + FN}, \qquad P = \frac{TP}{TP + FP}, \qquad F = \frac{2RP}{P + R}. \tag{2.2}$$

The score threshold that gave the best F-measure was checked on the Test subset and selected as the best threshold to use for accepting inferred associations [6].

2.2.4 Exploiting the EC Number Hierarchy

The above approach has focused on finding explicit co-occurrences between Pfam domains and 4-digit EC numbers. However, it is possible to find more associations by relaxing the criteria for co-ocurrences of EC-Pfam annotations by looking for matches only at the 3-digit EC level. Indeed, we have observed several cases where true associations according to the InterPro training dataset were assigned confidence scores below the threshold value because they had too few (4-digit EC number) instances to provide sufficient support. Therefore, the above procedure was repeated using 3-digit EC numbers to give a 3-digit scoring scheme (with different weight factors and a different score threshold). Then, any 4-digit EC-Pfam association below the 4-digit threshold, but consistent with a 3-digit EC-Pfam association above the 3-digit threshold, was added (i.e. "rescued") to the final list of accepted 4-digit EC-Pfam associations. It should be clarified that "consistent" means here that the 4-digit EC number is a descendant of the 3-digit EC number and that the Pfam domains are the same.

2.2.5 Hypergeometric Distribution p-Value Analysis

While the above procedure provides a systematic way to infer EC-Pfam associations, we wished to estimate the statistical significance, and thus the degree of confidence, that might be attached to those predictions. More specifically, we wished to calculate the probability, or "p-value", that an EC number and a Pfam domain might be found to be associated simply by chance. For example, it is natural to suppose such associations can be predicted at random if ec or d are highly represented in the structure/sequence CIDs. In principle, in order to estimate the probability of getting our EC-Pfam associations by chance, one could generate random datasets by shuffling the relations between EC numbers and CIDs on the one hand, and between Pfam domains and CIDs on the other hand. However, this is quite impractical given the very large numbers of CIDs, EC numbers, and Pfam domains, and the complexity of the filtering procedure that would have to be repeated for each shuffled version of the dataset. Therefore, as in [Fang and Gough, 2013], we rather assume that the random distribution of the number of CIDs associated with both ec and d follows an hypergeometric law.

Letting N denote the total number of CIDs, N_d the number of CIDs related to the Pfam domain d, and N_{ec} the number of CIDs related to the EC number ec, the hypergeometric probability distribution is given by

$$p(X_{ec,d} \geqslant K_{ec,d}) = \frac{\sum_{i=K_{ec,d}}^{\min(N_d, N_{ec})} \binom{N_{ec}}{i} \binom{N - N_{ec}}{N_d - i}}{\binom{N}{N_d}}, \tag{2.3}$$

where $p(X_{ec,d} \geqslant K_{ec,d})$ represents the probability of having a number $X_{ec,d}$ equal to or greater than the observed number $K_{ec,d}$ of CIDs associated with both d and ec. Traditionally, a p-value of less than 0.05

[6] F-measure is chosen as the performance measure because it considers true and false positive and negative instances classified by our system. Furthermore, our test dataset is balanced, thus, other performance measures such as MCC which also take TP, TN, FP and FN into account, provide the similar results.

is taken to be statistically significant. However, because this test is applied to a large number of EC-Pfam associations, we apply a Bonferroni correction which takes into account the so-called family-wise error rate (FWER) [Cui et al., 2003]. We therefore consider any p-value less than $0.05/T$ as denoting a statistically significant inferred EC-Pfam association in a dataset, with T the total number of tested EC-Pfam associations for this dataset, In order to distinguish EC-Pfam associations using both confidence scores and p-values, we classify them into three classes, "Gold", "Silver", and "Bronze". An association is assigned to the Gold class if both its EC-Pfam score is greater than the determined threshold and all its p-values (in all datasets) are statistically significant. An association is labeled Silver if its score is above the threshold but one or more of its p-values is not statistically significant, or if its score is below the threshold ("rescued" associations, see Section 2.2.4) but all its p-values are statistically significant. All other associations are labeled Bronze.

Please note that the above-mentioned method will be generalized in the next chapter. Figure 3.9 exemplifies a workflow of the algorithm to map protein functions and domains.

2.3 Results and Discussion

2.3.1 Data Source Weights and Score Threshold

After clustering identical structures and sequences, and calculating raw association scores (Figure 2.2), our merged dataset contains 6,306 SIFTS, 18,917 SwissProt, 124,699 TrEMBL, and 141,990 UniRule candidate EC-Pfam associations, giving a total of 262,571 distinct EC-Pfam associations to draw from (Table 2.1). In our ROC-based training procedure (Section 2.2.2), the best AUC value of 0.985 was obtained with weights $w_{SIFTS} = 0.1$, $w_{SwissProt} = 1.0$, $w_{TrEMBL} = 0.1$, and $w_{UniRule} = 0.6$. These weights indeed give greater importance to the candidate associations in SwissProt and UniRule, respectively, compared to those in SIFTS and TrEMBL. This is mainly due to fact that data in SwissProt and UniRule datasets has higher quality compared to the TrEMBL. However, the small amount of data in SIFTS dataset results in low weight.

The optimal score threshold was determined according to the F-measure training procedure using our training dataset (Section 2.2.3). This gave a score threshold of 0.04 for a maximum F-Measure of 0.9476. Applying this threshold to the test dataset yielded a comparable F-measure of 0.935, and precision and recall values of 0.99 and 0.893, respectively.

2.3.2 Global Analysis of Inferred EC-Pfam Associations

The results of the ECDomainMiner approach are summarized in Table 2.1. This table shows the numbers of 4-digit EC-Pfam associations along with the numbers of distinct EC numbers and Pfam entries involved in those associations for the four sources and the merged datasets before filtering. After applying the 0.04 score threshold, the number of EC-Pfam associations falls to 8,256 with an overlap of about 96% of InterPro reference associations. Using the relaxed 3-digit association approach (Section 2.2.4), the final ECDomainMiner dataset contains 20,728 EC-Pfam associations that overlap by 99.3% the InterPro reference dataset. These numbers show that our approach efficiently retrieves the InterPro reference EC-Pfam associations, including a small percentage (about 3.3%) that have a low confidence score.

Table 2.1 also shows that our ECDomainMiner set of EC-Pfam associations represents a 13.7 fold-increase (20,728 / 1,515) in EC-Pfam associations with respect to InterPro. Moreover, the list of EC-Pfam associations produced by ECDomainMiner contains 6.4 times more EC numbers and 2.8 times more Pfam

	Dataset	EC-Pfam associations	Distinct 4-digit EC numbers	Distinct Pfam entries
Source Datasets	SIFTS	6,306	2,648	2,611
	SwissProt	18,917	4,013	3,101
	TrEMBL	124,699	3,751	5,703
	UniRule	141,990	1,020	2,907
	Merged	262,571	4,648	6,639
Reference	InterPro	1,515	688	1,284
ECDomainMiner Results	With CS above threshold	**8,256**	**3,701**	**3,022**
	(Overlap with InterPro)	(1,461)	(688)	(1,245)
	Including low CS	**20,728**	**4,455**	**3,613**
	(Overlap with InterPro)	(1,498)	(688)	(1,273)

Table 2.1: Statistics on the source datasets and calculated EC-Pfam associations. CS is the Confidence Score.

domains than InterPro. Figure 2.3 shows how this increase in EC-Pfam associations distributes across the 6 top-level branches (i.e. 1-digit codes) of the EC classification. The greatest ECDomainMiner scale-

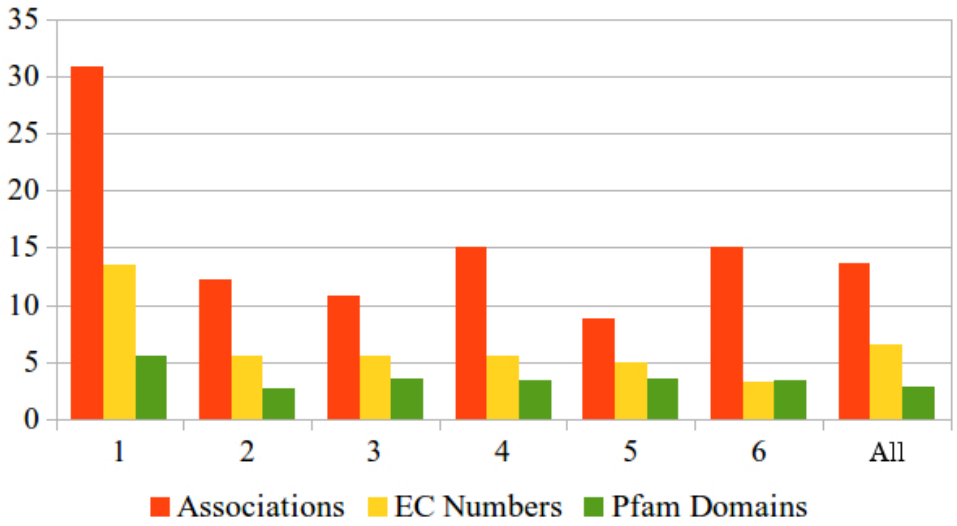


Figure 2.3: Scale-up factors for ECDomainMiner compared with InterPro. Ratios between the numbers in ECDomainMiner and in InterPro have been calculated for associations (red), EC numbers (yellow), and Pfam domains (green) after dividing the dataset according to each EC branch represented in the associations (1 to 6) and for all the dataset (All). 1: oxydoreductases; 2: transferases; 3: hydrolases; 4: lyases; 5: isomerases; 6: ligases

up factor occurs for associations involving the oxydoreductases (EC branch 1). The smaller scale-up factor observed for Pfam domains (2.8 versus 6.4 for EC numbers) can be explained by the fact that not all Pfam domains display an enzymatic activity. Thus there is a natural limit in the coverage of Pfam database by our EC-Pfam associations, whereas there is no such limit for the coverage of EC numbers. Combining the confidence scores with the calculated p-values as described in Section 2.2.5 gave 4,552 Gold associations (having scores above the threshold and significant p-values in all source datasets), 11,426 Silver associations (with either scores above the threshold and one or more non-significant p-values, or

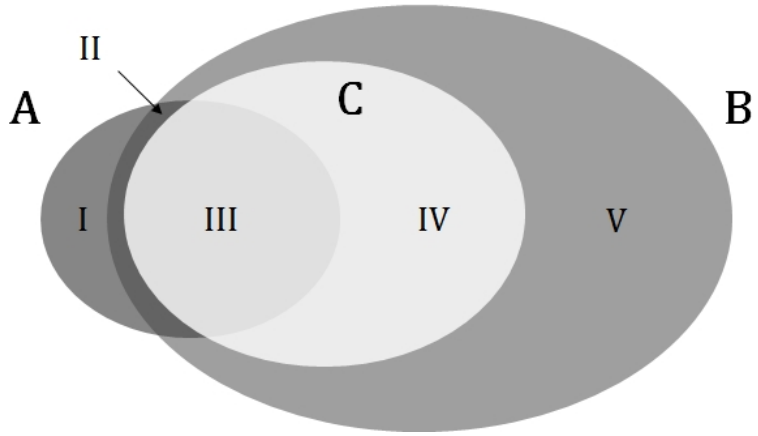


Figure 2.4: Venn diagram showing the intersection between (A) Pfam2EC (2,500 associations) from dcGO, (B) All-Merged (262,571 associations), and (C) ECDomainMiner (20,728 associations). Region I (480 associations) is the portion of (A) for which there is no data in any of our four source datasets. Region II (128 associations) is the portion of (A) that exists in (B) but is not retained in ECDomainMiner (C). Region III (1,892 associations) is the overlap between (A) and (C). Region IV (18,836 associations) is the portion of ECDomainMiner associations that are not available from SCOP2EC. Region V (241,363 associations) is the rest of the merged set of EC-Pfam source associations that are absent from (A) and not retained as Gold, Silver, or Bronze associations by ECDomainMiner.

with a score below the threshold but with significant p-values in all source datasets), and 4,201 Bronze associations.

2.3.3 Comparison with dcGO

In order to compare ECDomainMiner with the dcGO approach [Fang and Gough, 2013], we extracted SCOP2EC associations from the Domain2EC file available from the dcGO database [7]. The Domain2EC file includes 7,249 associations with 4-digit EC numbers, of which 3,774 are related to SCOP "Families" and 3,475 to SCOP "SuperFamilies". Because InterPro only tabulates SCOP family domains, we limited our comparison to the set of 3,774 SCOP2EC family associations. The SCOP families were mapped to Pfam families according to InterPro mapping files in order to generate a set of 2,500 "Pfam2EC" associations (*i.e.* EC-Pfam associations which may be deduced directly from the SCOP2EC data). This set (shown as set A in Figure 2.4) was compared with the set of all 262,571 merged EC-Pfam associations found by ECDomainMiner (set B in Figure 2.4).

This comparison showed that a total of 480 Pfam2EC associations from SCOP2EC are not present in our merged dataset. The remaining 2,020 Pfam2EC associations were then compared with the 20,728 associations calculated by ECDomainMiner (set C in Figure 2.4). This comparison (the intersection of sets A and C) produced a total of 1,892 EC-Pfam associations which are common to Pfam2EC and ECDomainMiner, indicating that ECDomainMiner agrees with 75.7% of the Pfam2EC associations from dcGO. Furthemore, this comparison also shows that ECDomainMiner result set contains 18,836 $(20,728 - 1,892)$ additional EC-Pfam associations that are not available through dcGO.

[7] `http://supfam.org/SUPERFAMILY/dcGO`

2.3.4 Selecting plausible associations in multi-domain proteins

Because ECDomainMiner finds many new EC-Pfam associations, it is important to ask to what extent it also might produce false associations. Firstly, we recall that ECDomainMiner eliminated more than 92% (241,843 out of 262,571) of low-scoring associations from the merged source dataset. This suggests that most of the eliminated associations involve Pfam domains that are not catalytically active. Indeed, if a Pfam domain is not regularly associated with protein chains or sequences having an enzymatic activity, the ECDomainMiner score for that domain is very low, and hence no EC number is assigned to that domain. This applies in particular for accessory domains that can co-occur with various catalytic domains in multi-domain proteins. A good example of such an accessory domain is PF00188 (the CAP protein family) which is a part of 216 different architectures. Among these architectures, there are 3 and 5 different architectures, which additionally contain PF00112 (Peptidase C1 domain) and PF00069 (Protein kinase domain), respectively. According to Pfam website, PF00188 is catalytically inactive but PF00112 and PF00069 are active. In fact, ECDomainMiner assigns PF00112 to 26 different EC numbers with a majority of EC 3.4.22 (Cysteine endopeptidases), and PF00069 to 28 different EC numbers that all start with 2.7 (Transferring phosphorus-containing groups). However, ECDomainMiner does not assign PF00188 to any EC number. This is because a large number of protein chains and sequences containing either PF00112 or PF00069 and associated with the above-mentioned EC activities, do not contain PF00188. In other words the catalytic activities of PF00112 and PF00069 are not strictly dependent on the presence of PF00188. Moreover, the SIFTS and UniProt databases indicate that PF00188 is associated with 43, and 5,197 different protein chains and sequences, respectively. However, none of those protein chains are associated with a EC number in SIFTS and only 31 protein sequences (24 in TrEMBL and 7 in UniRule) are associated with at least one 4-digit EC number. Consequently, the association score of PF00188 with any EC number is zero for both the SIFTS and SwissProt datasets and is very small (less than 0.02) for both the TrEMBL and UniRule datasets. Thus, the confidence scores of all of the associations involving PF00188 in ECDomainMiner are lower than our threshold of 0.04, and so these candidate associations are filtered out. This mechanism explains how an accessory domain is not assigned to an EC number by ECDomainMiner, and suggests that most of the retained associations are proper candidates for domain functional annotation.

2.3.5 Single and Multiple EC-Pfam Associations

Exploring the ECDomainMiner results readily reveals that a given EC number or Pfam domain can be involved in one or more distinct EC-Pfam associations. Figure 2.5 shows the relative distribution of EC numbers and Pfam domains according to the number of EC-Pfam associations they are involved in. This figure shows that 1,576 out of 4,393 EC numbers and 1,280 out of 3,542 Pfam domains are involved in a single EC-Pfam association. Although this represents rather high proportions of the total number of EC numbers and Pfam domains in ECDomainMiner (35.9% and 36.1%, respectively), the intersection of the concerned EC-Pfam single associations yields a list of only 97 one-to-one EC-Pfam associations, of which 62, 34, and 1 are Gold, Silver, and Bronze associations, respectively. Comparison with the InterPro reference dataset reveals that two thirds (65) of these one-to-one associations are novel compared to InterPro. Interestingly, we confirmed in our source datasets that all of these associations involve single-domain proteins. Thus, these unambiguous associations constitute the most reliable novel associations calculated by ECDomainMiner.

The complete list of one-to-one EC-Pfam associations found by ECDomainMiner may be downloaded from the ECDomainMiner web site. Interestingly 14 of these associations (8 Gold, of which 2 match

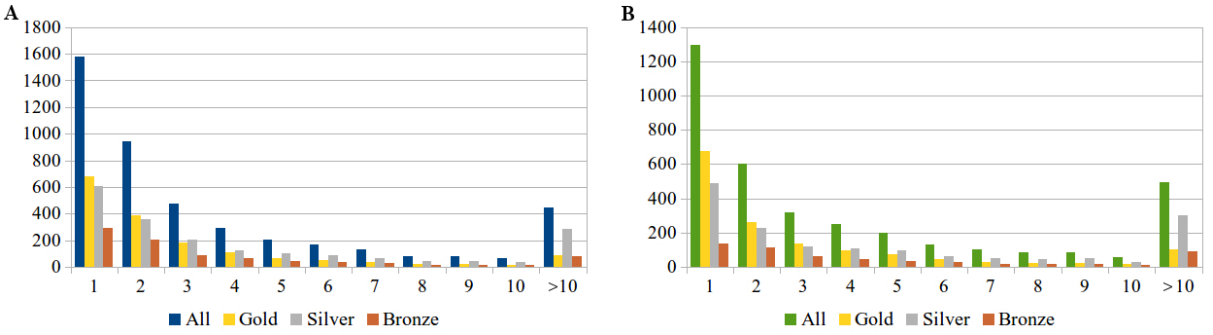


Figure 2.5: Distribution of EC numbers (A) and Pfam domains (B) in multiple associations. Numbers (1 to 10 and >10) represent the arity of the association in which a given EC number, respectively Pfam domain, is involved. In addition, for each arity, the normalized number of Gold, Silver, and Bronze associations is plotted. It can be observed that for arities equal to or greater than 4, the proportion of Silver associations is always the highest but significant amounts of Gold associations remain present even for high arity numbers.

InterPro reference associations, and 6 Silver) concern "DUF" (domain of unknown function) or "UPF" (uncharacterized protein family) Pfam entries. They are listed in part (A) of Table 2.2 according to decreasing confidence score.

These examples demonstrate that ECDomainMiner can be used to enrich domain annotation. Visual inspection of the one-to-one EC-Pfam associations indicates that about one quarter of them (23) could have been retrieved simply by comparing the names associated with the EC number and the Pfam identifier, which are nearly identical (see example in Table 2.2(B)). However, only 10 of these associations were in fact already known in InterPro. As it is shown in the table, minor and unpredictable spelling differences impair the automatic retrieval of such similar but non-identical EC and Pfam names. Nonetheless, while these associations could be found by clever text matching, we emphasise that ECDomainMiner's confidence scores and p-values provide a level of support for each association that would be very difficult to obtain from text mining alone.

The multi-partner associations calculated by ECDomainMiner provide many more complex EC-Pfam associations. As a first analysis of such multiple associations, we looked for obligate pairs or tuples of Pfam domains that are always associated with a given EC number. Briefly, for any pair of Pfam domains, (d_1, d_2), associated with the same EC number, *ec*, (i) we reject those pairs for which at least one *ec*-annotated CID (in any source dataset) occurs in relation with d_1 and not d_2 or with d_2 and not d_1, (ii) for all other pairs we calculate for each source dataset the ratio of the number of *ec*-annotated CIDs related to d_1 and d_2, to the total number of *ec*-annotated CIDs. A support ratio of 1 means that all CIDs annotated with *ec* in a dataset are also related to d_1 and d_2. A similar algorithm was used for triplets and quadruples of Pfam domains. For a support ratio of 1 in at least one source dataset, we found 907, 191 and 47 obligate associations between an EC number and a pair, a triplet or a quadruplet of Pfam domains. These associations are available from the ECDomainMiner website. Two examples are given in part (C) of Table 2.2.

Interestingly, filtering the names of the Pfam domains with the expressions "N-terminal" and "C-terminal" yielded 58 obligate pairs containing both a N-terminal and a C-terminal domain of the same function. This indicates that our approach is finding enzymes in which the catalytic function is provided

	EC	Pfam	Score	EC name	Pfam name	Quality	PDBs (SIFTS)
A	2.7.8.28	PF01933	0.972	2-phospho-L-lactate transferase	Uncharacterized protein family UPF0052	Gold	9/0/11
	4.1.99.5	PF11266	0.944	Aldehyde oxygenase (deformylating)	Protein of unknown function DUF3066	Gold	18/0/0
	2.1.1.286	PF11968	0.889	25S rRNA (adenine(2142)-N(1))-methyltransferase	Putative methyltransferase DUF3321	Gold	0/0/0
	1.13.99.1	PF05153	0.667	Inositol oxygenase	Family of unknown function DUF706	Gold	4/0/0
	2.4.1.155	PF15027	0.611	Alpha-1,6-mannosyl-glycoprotein 6-beta-N-acetylglucosaminyltransferase	Domain of unknown function DUF4525	Gold	0/0/0
	4.2.3.130	PF10776	0.611	Tetraprenyl-beta-curcumene synthase	Protein of unknown function DUF2600	Gold	0/0/0
	2.3.1.78	PF07786	0.609	Heparan-alpha-glucosaminide N-acetyltransferase	Protein of unknown function DUF1624	Gold	0/0/0
	3.1.4.45	PF09992	0.584	N-acetylglucosamine-1-phosphodiester alpha-N-acetylglucosaminidase	Predicted periplasmic protein DUF2233	Gold	0/0/1
	1.13.12.20	PF08592	0.556	Noranthrone monooxygenase	Domain of unknown function DUF1772	Gold	0/0/0
	2.1.1.312	PF11312	0.556	25S rRNA (uracil(2843)-N(3))-methyltransferase.	Protein of unknown function DUF3115	Gold	0/0/0
	2.1.1.313	PF10354	0.556	25S rRNA (uracil(2634)-N(3))-methyltransferase	Domain of unknown function DUF2431	Gold	0/0/0
	2.5.1.128	PF01861	0.556	N4-bis(aminopropyl) spermidine synthase	Protein of unknown function DUF43	Gold	0/0/1
	5.2.1.14	PF13225	0.556	Beta-carotene isomerase	Domain of unknown function DUF4033	Gold	0/0/0
	1.14.99.29	PF04248	0.333	Deoxyhypusine monooxygenase	Domain of unknown function DUF427	Silver	0/0/5
B	6.3.2.25	PF03133	0.610	Tubulin–tyrosine ligase	Tubulin-tyrosine ligase family	Gold	0/2/21
C	2.7.1.30	PF00370	0.847	Glycerol kinase	FGGY family of carbohydrate kinases, N-terminal domain	Gold	85/32/9
		PF02782	0.828		FGGY family of carbohydrate kinases, C-terminal domain	Gold	85/32/7
	6.3.4.23	PF06973	0.997	Formate-phosphoribosyl-amino-imidazol carboxamide ligase	DUF1297	Gold	16/3/0
		PF06849	0.997		DUF1246	Gold	16/3/0

Table 2.2: One-to-one examples of the EC-Pfam association. (A) Fourteen one-to-one EC-Pfam associations found by ECDomainMiner and involving domains of unknown function, (B) an example of one-to-one EC-Pfam association with very similar EC and Pfam descriptions, and (C) two examples of obligate Pfam pairs associated with an EC number. The 'PDBs (SIFTS)' column contains 3 counts of PDB chains containing the mentioned Pfam domains: in the first position, the count of PDB chains having in SIFTS the same EC annotation as recommended by ECDomainMiner, in the second position, the count of PDB chains with different EC annotation and in the third position, the count of PDB chains with no EC annotation in SIFTS. More detail and complete lists of PDB identifiers can be retrieved from the ECDomainMiner web server.

Association Type	ECDomainMiner Associations Concerned	PDB Chains Concerned
Any	14,573	58,722
Gold	3,591	41,246
Silver	7,796	44,406
Bronze	3,186	34,820
One-to-One	44	1,334

Table 2.3: The numbers of PDB protein chains that could be annotated by ECDomainMiner associations.

by the interface between two consecutive Pfam domains. Only 4 of these obligate pair associations are currently documented in InterPro.

2.3.6 Annotating PDB Chains with EC Numbers

Our analysis of the December 2015 release of the SIFTS database reveals that about 45% of PDB entries lack an EC number annotation. Of course, such an annotation is not expected to be present in all PDB entries because not all proteins have enzymatic activity. Nonetheless, it is interesting to use ECDomainMiner to analyze the number of PDB entries that contain Pfam domains which are present in EC-Pfam associations. Table 2.3 shows that a total of 58,722 PDB chains lacking EC annotations in SIFTS include at least one of the 3,542 Pfam domains present in ECDomainMiner. Overall, we calculated that these chains map to a total of 24,995 PDB entries that could benefit from the additional annotations inferred by ECDomainMiner. For those chains lacking EC annotations, ECDomainMiner finds Gold, Silver, and Bronze EC-Pfam associations for 41,246, 44,406 and 34,820 PDB chains, respectively. In particular, 1,334 PDB chains could benefit from our dataset of 97 non ambiguous one-to-one EC-Pfam associations.

In chapter 4, a more systematic way to predict protein functions, using taxonomic information and combinations of domains will be introduced.

2.3.7 The ECDomainMiner web server

The ECDomainMiner web server (Figure A.1) may be queried by EC number or Pfam domain. Thus, if one wishes to search for associations for a protein chain that currently lacks any EC annotation in the PDB (e.g. chain 2q7xA), one first needs to retrieve from the PDB the Pfam domain(s) that it contains (in this example, PF01933). Then, querying the ECDomainMiner server with each Pfam domain identifier will show the associated EC numbers (in this example, 2.7.8.28), along with the associated filtering scores and quality classes. In this example, ECDomainMiner finds a Gold quality association between PF01933, present in PDB chain 2q7xA, and EC number 2.7.8.28 (2-phospho-L-lactate transferase) which consequently can be associated with PDB entry 2q7x. Interestingly, PDB entry 2q7x is described as a putative phospho transferase from *streptococcus pneumoniae* tigr4, which is consistent with the enzymatic activity found by ECDomainMiner, and which could not be deduced from the Pfam domain name (UPF0052).

2.4 Conclusion

We have presented a filtering approach for associating EC numbers with Pfam domains. This approach has been shown to be able to infer a total of 20,728 non-redundant EC-Pfam associations, which corresponds to over 13 times as many EC-Pfam associations as currently exist in InterPro. Furthermore, thanks to

our calculated p-values, we have assigned an intuitive quality rating (Gold, Silver, or Bronze) to each EC-Pfam association found. These calculated associations are publicly available on the ECDomainMiner web site.

We believe that enriching protein chain annotations will facilitate a better understanding and exploitation of structure-function relationships at the domain level. While many of the associations calculated by ECDomainMiner are consistent with those recently made available by the domain-centric dcGO approach for finding EC-SCOP associations, the ECDomainMiner results set contains many more associations than dcGO. Indeed, the ECDomainMiner result set contains 18,836 EC-Pfam which are not available in dcGO. Our analysis of the simple one-to-one associations found by ECDomainMiner shows that several DUF or UPF entries in Pfam may be assigned functions from the EC classification, and that obvious inconsistencies in the annotation texts may easily be corrected or unified. However, only a relatively small number (less than 0.5 %) of EC-Pfam associations in our result set are simple one-to-one associations, indicating that there exist a large number of many-to-many relations between EC numbers and Pfam domains. Further analyses of these complex associations using graph database and machine-learning techniques could reveal many more hidden protein structure-function relationships.

In the next chapter, we show that our method can be generalized to other annotation vocabularies or ontologies, such as GO. However, it is worth mentioning that our findings include less noise for those ontologies whose terms are in average assigned to fewer protein sequences, like EC numbers.

Chapter 3

Computational Discovery of Direct Associations between Annotations using Common Content - CODAC

Contents

Families of related proteins and their different functions may be described systematically using common classifications and ontologies such as Pfam and GO (Gene Ontology), for example. However, many proteins consist of multiple domains, and each domain, or some combination of domains, can be responsible for a particular molecular function. Therefore, identifying which domains should be associated

with a specific function is a non-trivial task. We describe a general approach, based on our experience from chapter 2, for the computational discovery of associations between different sets of annotations by formalizing the problem as a bipartite graph enrichment problem in the setting of a tripartite graph. We call this approach "CODAC" (for COmputational Discovery of Direct Associations using Common Neighbors). As one application of this approach, we describe "GODomainMiner" for associating GO terms with protein domains.

We used GODomainMiner to predict GO-domain associations between each of the 3 GO ontology namespaces (MF, BP, and CC) and the Pfam, CATH, and SCOP domain classifications. Overall, GODomainMiner yields an average enrichment of 15-, 41- and 25-fold in GO-domain associations compared to the existing GO annotations in these 3 domain classifications, respectively. These associations could potentially be used to annotate many of the protein chains in the Protein Data Bank and protein sequences in UniProt whose domain composition is known but which currently lack GO annotation. The GODomainMiner result database is publicly available at http://godm.loria.fr/.

3.1 CODAC

In this chapter an approach (called CODAC) to directly associate two sets of items which are indirectly linked is described. Given two items (A and B) are joint through a set of items (I call them common contents (CC)), eliminating the common contents gets the two items associated directly with a similarity score. This similarity score shows either how strong is the connection between the two items or how similar these two items are. For the simplicity, linkage between two items and the common contents can be depicted as a tripartite graph and association between two items is a bipartite graph. A tripartite graph is a graph whose vectors are partitioned into three different independent sets. A tripartite graph can be colored with three colors while no two endpoints of an edge have the same color. Moreover, a bipartite graph (or bigraph) is a graph whose vertices can be divided into two disjoint sets such that every edge connects a vertex from one set to another. Here, the tripartite graph has one limitation which is defined based on the nature of future problems. There is no link between two sets of vertices. Equivalently, two sets of vertices are connected to each other only through the third set. CODAC removes the connector set of vertices and find direct links between the other two sets of vertices. Each link between the member of each vertex sets is presented with a corresponding score. This score demonstrates how good is the link between two vertices. Then, we can convert this link between two vertices into an association while the score shows the similarity between two associated vertices. Each association can statistically be analyzed in order to verify its statistical significance.

In the next section, the CODAC method is described in detail.

3.1.1 Tripartite Graph Model

In graph theory, a k-partite graph is a graph whose vertices can be partitioned into k disjoint subsets, such that in each subset no two vertices are connected. If $k = 2$, the graph is called a bipartite graph (or bigraph), and if $k = 3$ it is called a tripartite graph. The CODAC approach is designed to solve problems of bipartite graph enrichment within a tripartite graph framework. The main intuition is to calculate new weighted edges between two sets of items which already contain reliable but sparse associations, and which are indirectly connected through common associations with a third set of items.

Let $\mathcal{G}(X, Y, Z, E)$ be a tripartite graph where X, Y and Z are 3 sets of items and E is the set of all edges connecting X, Y and Z in the input configuration. Let us consider 3 bipartite subgraphs of

$\mathcal{G}$, denoted as $\mathcal{G}_1(X, Z, E_1)$, $\mathcal{G}_2(Y, Z, E_2)$, and $\mathcal{G}_3(X, Y, E_3)$. We now assume that the set of edges E_3 is incomplete, and that the aim is to compute new edges between items of X and items of Y in order to generate $\mathcal{G}_3^*(X, Y, E_3^*)$ which together with $\mathcal{G}_1$ and $\mathcal{G}_2$ will make the final tripartite graph, $\mathcal{G}^*(X, Y, Z, E^*)$, where E^* denotes an enriched set of edges. New edges may be discovered by exploiting the existing edge distributions in $\mathcal{G}_1$ and $\mathcal{G}_2$. For example, if items x_i of X and y_j of Y share the same (or almost the same) set of neighbors $\{z_k\}$ in Z, then it may be supposed that an edge might exist between x_i and y_j. Figure 3.1 illustrates the discovery of a candidate edge between x_2 and y_2 because these items are associated with the same subset of items $\{z_1, z_3, z_4\}$ from Z. Candidate edges found in this way are then scored and filtered, as described in more detail below.

It is now possible to instantiate our model with a set of MF GO terms (X), a set of Pfam domains (Y), and a set of UniProtKB/SwissProt sequences (Z). E_1 is the set of edges derived from the MF GO annotation of UniProtKB/SwissProt sequences, E_2 is the set of edges derived from the domain contents of UniProtKB/SwissProt sequences, and E_3 is the set of edges derived from the InterPro manually curated MF GO annotations of Pfam domains. In this case, our aim is to produce E_3^*, which will contain an enriched set of MF GO-Pfam associations weighted by their neighborhood similarity score.

3.1.2 Biadjacency Representation of bigraphs

While graphs allow complex relationships to be visualised easily, analysing graphs computationally can be very time-consuming. In our approach it is convenient to represent each bigraph as a bi-adjacency matrix, in which a matrix element has a value of 1 or 0 according to whether the corresponding pair of nodes is connected or not.

Given a tripartite graph $\mathcal{G}(X, Y, Z, E)$ as input, the core CODAC algorithm divides it into two bigraphs $\mathcal{G}_1(X, Z, E_1)$ and $\mathcal{G}_2(Y, Z, E_2)$. A procedure named *Cosine* calculates a cosine similarity matrix C between items of X and items of Y using the two biadjacency matrices M_1 (of dimension $|X| \times |Z|$) and M_2 (dimension $|Y| \times |Z|$), derived from $\mathcal{G}_1$ and $\mathcal{G}_2$, respectively. These matrices are then row-normalized to give matrices U_1 and U_2. Each element of the matrix $C = U_1 \times U_2^T$ thus represents a cosine similarity between an item x of X and an item y of Y, according to the number of common associations with the items in Z.

The main procedure called *PredictAssociations* determines a similarity threshold T for filtering the raw scores in C to produce C^*. The matrix C^* can be interpreted as the weighted biadjacency matrix of the enriched bigraph $\mathcal{G}_3^*(X, Y, E_3^*)$ and therefore used to predict new weighted associations between items of X and Y. Pseudocode for the core CODAC algorithm is presented in Algorithm 1.

3.1.3 Gold Standard of Positive and Negative Examples

In order to determine an edge similarity threshold, we need to define a "gold standard" set of positive and negative examples of associations. Here, we take all of the $P = |E_3|$ existing associations present in $\mathcal{G}_3$ as positive examples. To create negative examples, we shuffle the edges of $\mathcal{G}_1$ and $\mathcal{G}_2$ in order to rearrange in a random way all edges between X and Z, and between Y and Z. During shuffling, the node degrees of each x_i, y_j and z_k is kept constant, and the shuffled edges are constrained not to overlap the original edges. The shuffled graphs are denoted by $\mathcal{G}_1^\#$ and $\mathcal{G}_2^\#$, from which a new shuffled cosine similarity matrix, $C^\#$, may be calculated. This matrix is then used to select $|N| = |P|$ negative examples at random. Taken together, the P positive and N negative examples constitute our "Gold Standard" dataset.

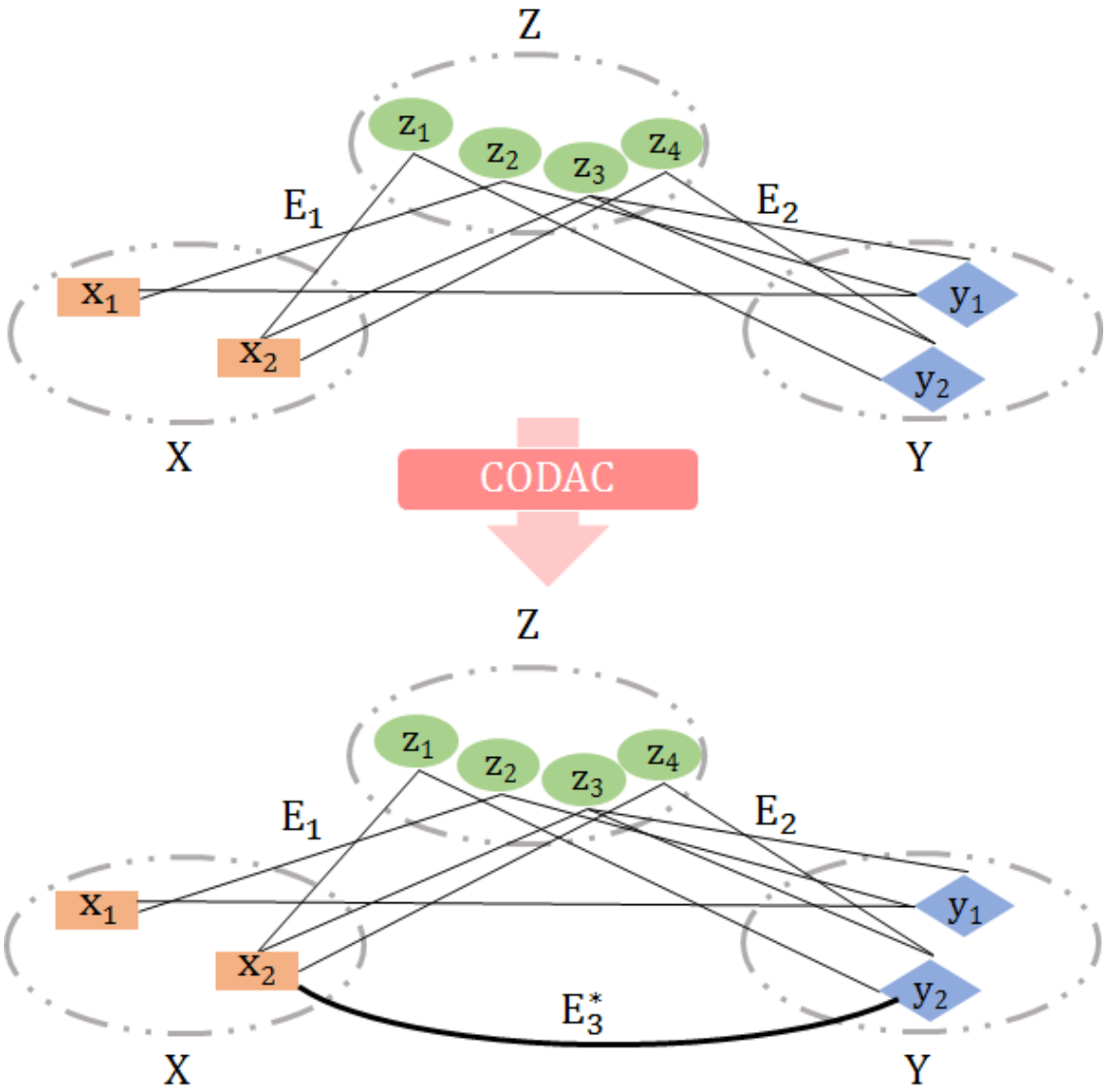


Figure 3.1: Schematic illustration of edge discovery. In a typical instantiation, X is a set of MF GO terms, Y a set of Pfam domains, and Z a set of UniProtKB/SwissProt sequences. E_1 are edges derived from the MF GO annotation of UniProtKB/SwissProt sequences, E_2 are edges derived from the domain contents of UniProtKB/SwissProt sequences, E_3^* is the enriched set of edges, derived from initial E_3 that included a limited number of edges (represented here by (x_1, y_1)), derived from the InterPro manually curated MF GO annotations of Pfam domains. E_3^* contains all newly discovered MF GO-Pfam associations represented here by (x_2, y_2).

Algorithm 1 The Core CODAC Algorithm

Input: $\mathcal{G}(X, Y, Z, E)$, a tripartite graph with $\mathcal{G}_1(X, Z, E_1)$, $\mathcal{G}_2(Y, Z, E_2)$, $\mathcal{G}_3(X, Y, E_3)$, 3 associated bigraphs
Output: $\mathcal{G}_3^*(X, Y, E_3^*)$, the enriched bipartite graph with new weighted edges.

procedure $PredictAssociations(\mathcal{G})$
  $C = Cosine(\mathcal{G}_1, \mathcal{G}_2)$
  $\mathcal{G}_1^{\#} = Shuffle(\mathcal{G}_1)$
  $\mathcal{G}_2^{\#} = Shuffle(\mathcal{G}_2)$
  $C^{\#} = Cosine(\mathcal{G}_1^{\#}, \mathcal{G}_2^{\#})$
  $P = CreatePositives(C, \mathcal{G}_3)$
  $N = CreateNegatives(C^{\#})$
  $GS = CreateGoldStandard(P, N)$
  $\{Training, Test\} = SplitGoldStandard(GS)$
  $T = \arg\max_t FMeasure(Threshold_t, Training)$
  $ReportFMeasures(T, Test, Training)$
  $C^*_{i,j} = C_{i,j}$ if $C_{i,j} > T$ or if an (x_i, y_j) edge already exists in input E_3, otherwise $C^*_{i,j} = 0$ forall $\{i, j\}$
  $AddEdge(x_i, y_j, E_3^*)$ if $C^*_{i,j} > 0$ forall $\{i, j\}$
  return$(\mathcal{G}_3^*, C^*)$
end procedure

procedure $Cosine(\mathcal{G}_1, \mathcal{G}_2)$
  $M_1 = CreateBiadjacency(\mathcal{G}_1)$
  $M_2 = CreateBiadjacency(\mathcal{G}_2)$
  $U_1 = RowNormalise(M_1)$
  $U_2 = RowNormalise(M_2)$
  $C = U_1 \times U_2^T$
  return(C)
end procedure

3.1.4 Determining the Score Threshold

We randomly split the Gold Standard dataset into two groups with equal numbers of positive and negative examples to give a "Training" and a "Test" subset. We then rank the scores of all members of the Training subset, and label them "positive" or "negative" according to a score threshold that is varied from 0.0 to 1.0 in steps of 0.001. This allows us to determine the numbers of true positive (TP), false positive (FP), true negative (TN), and false negative (FN) predictions for each threshold. We then calculate the recall, $R = TP/(TP + FN)$, precision, $P = TP/(TP + FP)$, and the F-measure, $F_1 = 2RP/(P + R)$. The similarity threshold T that gives the best F-measure with the Training subset is verified using the Test subset and retained to calculate a filtered cosine similarity matrix, C^*, according to $C^*_{i,j} = C_{i,j}$ if $C_{i,j} > T$ or if the (x_i, y_j) edge already exists in E_3, otherwise, $C^*_{i,j} = 0$.

3.1.5 Combining Multiple Datasets

There may often be more than one configuration for a graph $\mathcal{G}$, that has the same $\mathcal{G}_3$ but different Z, E_1, and E_2 in $\mathcal{G}_1$ and $\mathcal{G}_2$. In our instantiation this corresponds to the fact that GO terms and Pfam domains can be indirectly connected either through UniProtKB/SwissProt sequences [Apweiler et al., 2010] or through PDB chains in SIFTS [Velankar et al., 2012]. To handle multiple datasets, each input tripartite graph is processed separately to calculate its respective cosine similarity matrix C^d. The cosine similarity scores are then combined as a weighted average to give a consensus similarity matrix, CS. Whenever there is no data for a given pair (x, y) in an input graph, the corresponding score $C^d_{x,y}$ is set to zero.

Receiver-operator-characteristic (ROC) analysis provides an objective way to measure the performance of an information retrieval system to retrieve positive documents as first ranked, i.e. with the best scores [Mogotsi, 2010]. One advantage of ROC-based approaches is that they are rather insensitive to the particular numbers of the positive and negative instances used [Chawla et al., 2004]. Here, in order to find the best values for the dataset weights w_d, each weight is varied from 1 to 10 in steps of 0.1, and for each combination of weights a ROC performance curve is calculated using the complete ranked list of consensus scores and our Gold Standard set of positive examples. The combination of weights that gives the largest area under the curve (AUC) is selected and used to calculate the best consensus similarity matrix CS. Then, the *PredictAssociations* procedure determines the best threshold to filter the consensus similarity matrix CS and to deduce the resulting enriched bipartite graph $\mathcal{G}^*_3$ (refer to Algorithm 2).

3.1.6 Bipartite Graph Extension with Hierarchy of Classes

Ontologies are often described as taxonomic hierarchies of classes, as is the case for the GO gene ontology [Ashburner et al., 2000]. Thus, if one of the input graphs contains items from a hierarchical ontology, important relationships between the ancestors of a term and its neighbor(s) could be missed because they are generally not mentioned explicitly in the data. For example, if a vertex x from set X represents a term in an ontology and has a neighbor z in set Z, it is quite possible that all of the ancestors of x present in X should also have z as neighbor. If requested by the user, whenever an edge (x, z) is found where z is annotated with an ontology term x, then CODAC will add additional edges between item z and all parents of x present in X. This is illustrated in Figure 3.2.

Algorithm 2 Calculating a Consensus Similarity Matrix

Input: $\mathcal{Z} = \{\mathcal{G}_1^d(X, Z^d, E_1^d), \mathcal{G}_2^d(Y, Z^d, E_2^d), d = 1, ...D\}$, a set of input bipartite graphs.

Input: $\mathcal{G}_3(X, Y, E_3)$, the bipartite graph to be enriched.

Output: CS, a consensus similarity matrix with an optimal set of weights, W.

procedure $Consensus(Z, \mathcal{G}_3)$
  for each $d \in \{1, ..., D\}$ **do**
    $C^d = Cosine(\mathcal{G}_1^d, \mathcal{G}_2^d)$
  end for
  for each set of weights $w = \{w_d\}$ with $d \in \{1, ..., D\}$ and $w_d \in [1, 10]$ with steps of 0.1 **do**
    $CS_{i,j}^w = \frac{\sum_d w_d \times C_{i,j}^d}{\sum_d w_d}$
    $ROC^w = CreateROC(CS^w, P)$
  end for
  $W = \arg\max_w AUC(ROC^w)$
  return(W, CS^W)
end procedure

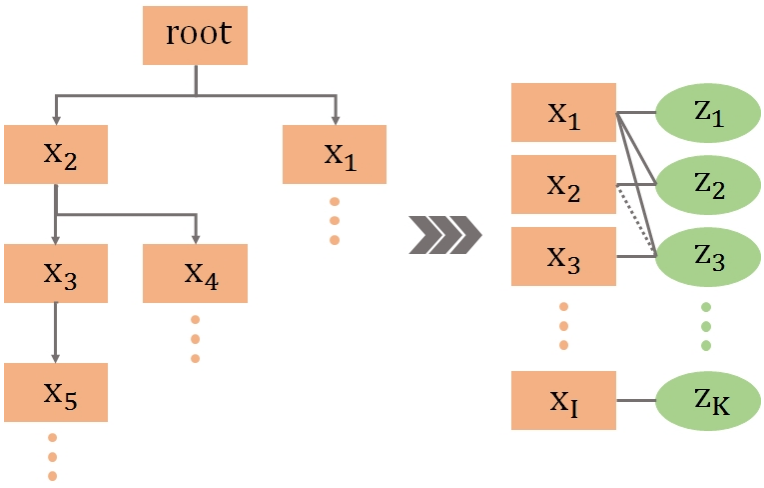


Figure 3.2: Edge enrichment using an ontology. Here, edge (x_2, z_3) is added (right, dashed link) because z_3 has an existing association with x_3, and x_2 is a parent term of x_2 in the ontology (left).

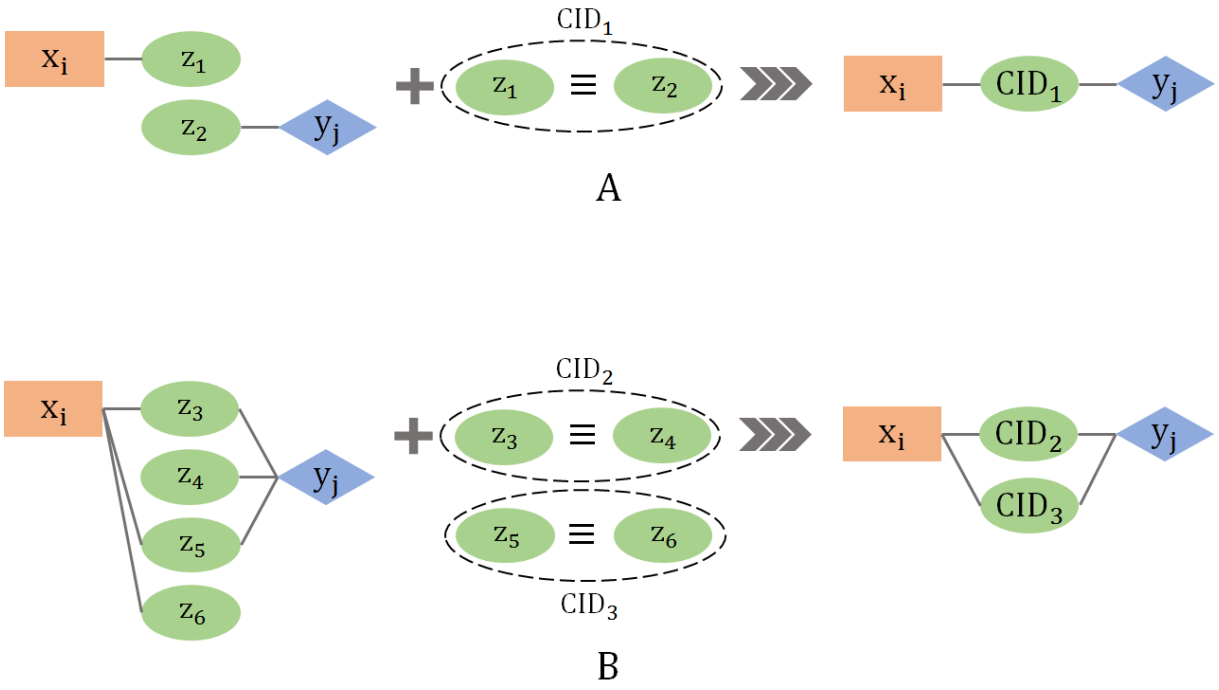


Figure 3.3: Clustering identical or highly similar items in Z. A: Clustering of items z_1 and z_2 of initial degree 1 induces a new association between x_i and y_j. B: Clustering reduces the complexity of initial multiple associations. In both cases, clustering will increase the cosine similarity scores of the associated items x_i and y_j.

3.1.7 Clustering Graph Edges

A possible source of bias in any data mining approach is the existence of redundant items in the input. This is especially the case for protein entries in UniProt where it is quite possible to have entries with different identifiers but identical amino-acid sequences. In order to deal with this possibility, CODAC groups all items in Z into clusters having 100% identity. Each cluster is represented by a unique cluster identifier (CID). As shown in Algorithm 3, all source edges (x, z_i) and (y, z_j) from E_1 and E_2 in which identical z_i and z_j belong to the same CID, are merged into unique (x, CID) and (y, CID) edges, producing $\mathcal{G}_1^{Cl}$ and $\mathcal{G}_2^{Cl}$, the reduced bipartite graphs that serve as input to the CODAC core approach. It should be noted that the 100% sequence identity threshold may be reduced to 99% or lower if desired. As illustrated in Figure 3.3, grouping identical items into clusters of 100% identity can be very beneficial for recovering missing edges.

3.1.8 Calculating Statistically Significant Edges in E_3^*

While our approach provides a systematic way to predict edges in $\mathcal{G}_3^*$, it is important to calculate a probability, or "p-value", for finding an edge simply by chance. For example, it is reasonable to suppose that an edge (x, y) might be predicted at random if x and y are each highly connected to many items in Z. In order to estimate the probability of finding edges by chance, one could generate multiple random graphs by shuffling the edges of a given input graph, as described above for constructing the Gold Standard *Negative* examples. However, this is quite impractical given the very large numbers of

Algorithm 3 Clustering Graph Edges

Input: $\mathcal{G}_1(X, Z, E_1)$ and $\mathcal{G}_2(Y, Z, E_2)$, two bipartite graphs having redundant items in Z.

Output: $\mathcal{G}_1^{Cl}$ and $\mathcal{G}_2^{Cl}$, the reduced bipartite graphs in which all items of Z are grouped by the cluster of identical items (CID).

```
1: procedure Cluster(G1, G2)
2:     Build Z^Cl = {CID_k}
3:     E_1^Cl = ∅
4:     for each (x, z) ∈ E_1, such that z ∈ CID do
5:         if (x, CID) ∉ E_1^Cl then Add (x, CID) to E_1^Cl
6:         end if
7:     end for
8:     E_2^Cl = ∅
9:     for each (y, z) ∈ E_2, such that z ∈ CID do
10:        if (y, CID) ∉ E_2^Cl then Add (y, CID) to E_2^Cl
11:        end if
12:    end for
13:    return(G1 = G_1^Cl, G2 = G_2^Cl)
14: end procedure
```

items in X, Y, and Z and the complexity of the filtering procedure that would have to be repeated for each shuffled version of the dataset. Instead, we assume that the probability for finding an edge (x, y) by random chance is given by a hypergeometric distribution of the number of common neighbors (x, z) and (y, z). Letting N_z denote the total number of items in Z, N_x the number of neighbors of x in Z, and N_y the number of neighbors of y in Z, the hypergeometric probability distribution is given by

$$p(K \geqslant K_{x,y}) = \sum_{v=K_{x,y}}^{\min(N_x,N_y)} \binom{N_x}{v}\binom{N_z - N_x}{N_y - v} / \binom{N_z}{N_y}, \tag{3.1}$$

where $p(K \geqslant K_{x,y})$ is the predicted probability of having a number, K, equal to or greater than the observed number $K_{x,y}$ of common neighbors z of both x and y. Because this p-value test is applied to a large number of (x, y) edges in $\mathcal{G}_3^*$, we apply a Bonferroni correction to take into account the so-called family-wise error rate [Cui et al., 2003]. Therefore, letting $|E_3^*|$ denote the total number of edges tested, we consider any p-value less than $0.05/|E_3^*|$ as denoting a statistically significant edge.

3.1.9 Classification into Gold, Silver, and Bronze Associations

While the above consensus scores and p-values give objective measures of the quality of predicted associations, from a user's point of view it is often convenient to provide a simple and memorable quality scale. Therefore, we classify a predicted association as "Gold" if all of the individual data source p-values for this association are statistically significant. A predicted association is classed as "Silver" if more than half of the data source p-values are statistically significant. Otherwise, it is classed as a "Bronze" association.

3.2 GODomainMiner: Computational Discovery of Direct Associations between GO terms and Protein Domains

Proteins are macromolecules which carry out many biological functions in living organisms. At the molecular level, protein functions are often performed by highly conserved structural regions identified from sequence or structure alignments, which may be classified into families of domains. Because many protein domains fold into characteristic three-dimensional (3D) structures, there is often a close relationship between protein structure and protein function [Berg et al., 2002]. Currently, the Pfam database is one of the most widely used sequence-based classifications of protein domains and domain families [Finn et al., 2016b]. The CATH [Orengo et al., 1997] and SCOP [Murzin et al., 1995] databases are examples of structural domain classifications.

As well as sequence-based and structure-based classifications, proteins may also be classified according to their function. For example, the Gene Ontology (GO) [Ashburner et al., 2000] consists of a controlled vocabulary of GO terms which describe the gene products in a cell. Each GO term has a name, a distinct alphanumeric identifier, and a "namespace" (ontology) which has one of the following 3 values: biological process (BP), molecular function (MF), or cellular component (CC). The GO ontology is structured as a rooted Directed Acyclic Graph (rDAG) in which terms are nodes connected by different hierarchical relations. However, most protein domain classification systems annotate domains only according to the entire protein to which it belongs. One interesting exception is the dcGO database [Fang and Gough, 2013] which provides multiple ontological annotations (such as GO) for protein domains. Nonetheless, we found that there are several manually curated GO-Pfam associations from InterPro [Finn et al., 2016a] which are not present in dcGO. Indeed, from the results of a previous version of our approach [Alborzi et al., 2015, Alborzi et al., 2017c], we estimated that dcGO associations can only annotate 43% of the unannotated structures in the Protein Data Bank (PDB) [Gutmanas et al., 2014].

More generally, there are many millions of protein sequences that currently lack GO annotations. On the other hand, only a relatively small number of distinct protein domain families exist, which are re-used and combined in different ways in different proteins. Indeed, compared to the vast number of different sequences that exist, current domain classifications contain of the order of only 15,000 distinct protein domain families. Therefore, it is natural to suppose that if known protein structure and sequence annotations could be assigned GO terms at the domain level, many of these annotations could be transferred to a potentially very large number of unannotated proteins. However, we emphasize here that our aim is to discover functional annotations for protein domains themselves rather than entire protein sequences, in order to improve domain description and classification by combining structural and functional features. Nonetheless, even the task of associating GO terms with protein domains is a non-trivial problem because, except for single-domain proteins where the mapping is obvious, many different kinds of relationships can occur (see Figure 3.4).

We described an early version of the approach presented here for assigning Enzyme Commission (EC) numbers to Pfam domains [Alborzi et al., 2017c]. Because our new GODomainMiner approach [Alborzi et al., 2017b] aims to answer a similar problem, with GO terms replacing EC numbers, we decided to generalise the overall approach under the name of CODAC (for COmputational Discovery of Direct Associations using Common Neighbors). Firstly, the problem is formalized as a bipartite graph enrichment problem in the setting of a tripartite graph. The core CODAC algorithm solves this problem using the vector cosine similarity model, from which it creates new weighted edges between items of the bipartite graph on the basis of their graph neighborhood similarity. This approach is augmented using

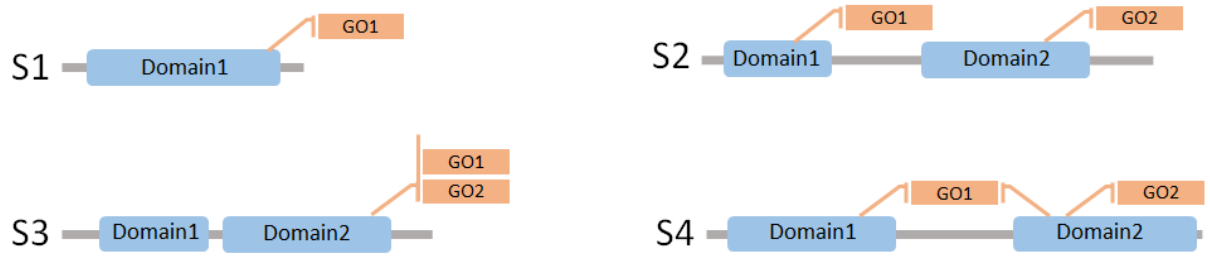


Figure 3.4: Graphical representation of the different kinds of relationships that may exist between GO terms and protein domains. S1: A protein with one domain providing one function; S2: Two domains of the same protein provide different functions; S3: A protein with two domains, where one domain provides two different functions, and the second domain has no known function; S4: A protein having one domain that provides one function, and a second domain which acts as a co-factor with the first domain to provide an additional function.

techniques to handle the problems of multiple data sources, bias due to identical items, the influence of the hierarchical organisation of the GO ontology, and statistical significance.

Here, the overall approach is applied to 9 different bipartite graphs involving the 3 GO ontologies (BP, MF, and CC) and 3 popular protein domain classifications (Pfam, CATH, and SCOP). Our results show that the GO-domain associations discovered by this approach represent an average of 15-, 41- and 25-fold increase in the number of edges on the concerned bipartite graphs. These newly discovered associations are compared with existing associations from InterPro and those predicted by dcGO, and a selected subset of one-to-one associations is analyzed from a biological point of view.

3.2.1 GODomainMiner Data Preparation

In this section, the CODAC approach is applied to discover new weighted GO-domain associations (the workflow is illustrated in Figure 3.9). In our $\mathcal{G}(X, Y, Z, E)$ tripartite graph model, the set X corresponds to one of the MF, BP or CC GO namespaces, and Y corresponds to one of the Pfam, CATH, or SCOP protein domain classifications. For each of the 9 combinations of X and Y, 3 data sources were selected to provide common neighbors (Z) of the items in X and Y, namely: (i) SIFTS providing curated PDB chain associations, (ii) UniProtKB/SwissProt (SP) providing curated UniProt entries, and (iii) UniProtKB/TrEMBL (TR) providing non-curated automatically annotated UniProt sequences.

Flat data files of SIFTS (June 2017), Uniprot (June 2017), and InterPro (version 63.0) were downloaded and parsed using in-house Python scripts. Associations between PDB chains and GO terms, and associations between PDB chains and protein domains (Pfam, CATH, and SCOP) were extracted from the SIFTS data. All CATH and SCOP domain families were transformed into their corresponding superfamilies, and all Pfam "repeat" and "motif" domain types were discarded. Associations between Uniprot sequence accession numbers (ANs) and GO terms and AN-Pfam associations (as well as AN-CATH and AN-SCOP associations) were extracted from the UniProtKB/SwissProt and UniProtKB/TrEMBL sections of Uniprot to give two datasets of UniProtKB/SwissProt associations and UniProtKB/TrEMBL associations, respectively. Then, using the evidence code of the GO term, the associations in the SIFTS, UniProtKB/SwissProt, and UniProtKB/TrEMBL datasets were divided into two groups, namely one group for which the GO term evidence code indicated manual curation, and one group for GO terms

with evidence code "inferred from electronic annotation" (IEA). We do not make any distinction between the various possible manual evidence codes. However, we note that the GO_REF field for IEA currently covers 12 an-notations sources, namely InterPro2GO, UniProt Keywords2GO, UniProt Subcellular Location2GO, EC2GO, UniRule2GO, UniPathway2GO, Ensembl Compara, Ensembl Fungi, Ensembl Metazoa, Ensembl Plants, Ensembl Protists, and the GeneOntology Consortium. Of these, the largest number of annotations come from InterPro2GO and UniProt Keywords2GO, which each provide around 169 million associations in UniProtKB. Moreover, only 34%, 4%, and 5% of the InterPro2GO annotations are GO-Pfam, GO-CATH, and GO-SCOP associations, respectively.

Here, the resulting 6 datasets are called SIFTS, SIFTS-IEA, SP, SP-IEA, TR, and TR-IEA. Thus, there are 6 input tripartite graphs for each of the 9 combinations of the X and Y source datasets. All PDB chain IDs and Uniprot ANs having identical sequences were clustered using the Uniref non-redundant cluster annotations [Suzek et al., 2007].

3.2.2 Dataset Weights and Threshold Scores

Using our Training set of InterPro-based positive associations and random negative associations, the best ROC-plot AUC values and optimal weights for each input source were calculated. Table 3.1 shows a summary of the obtained dataset weights, AUCs, F-measures of the Test and Training sets, and consensus score thresholds found from these calculations. This table shows that our procedure gives greater weight to GO-Pfam associations from the IEA sections of the SIFTS, UniProtKB/SwissProt, and UniProtKB/TrEMBL than to associations from the experimental and manually curated sections of SIFTS and UniProtKB/SwissProt datasets.

In order to investigate this further, we re-calculated the AUC-based weight optimization with all IEA weights forced to zero. This caused our optimal AUC to fall from around 0.96 to less than 0.60. This reflects the fact that in this setting, we do not consider the propagated InterPro2GO annotations in UniProtKB, and consequently the GODomainMiner retrieves less amount of Gold-Standard associations. As IEA annotations are not uniquely propagated from InterPro2GO, we also miss the contribution of the other annotation sources (refer to previous section). We therefore decided to incorporate IEA datasets into our approach for the rest of this study.

3.2.3 Analysis of Calculated GO-Pfam Associations

Summaries of our calculated GO MF-domain, BP-domain, and CC-domain associations are shown in Tables 3.2, 3.3, and 3.4, respectively. These tables show the numbers of distinct GO terms and domain entries (in units of thousands) involved in associations for the 6 source datasets, the filtered GODomainMiner predictions and the InterPro dataset of positive associations. In these tables, the total numbers of GO-Pfam associations found by GODomainMiner refer only to most-specific GO terms in each branch of a GO hierarchy. In other words, if a domain is associated to a GO term and to one or more of its parent terms, only the most-specific (non-parent) term is counted as a found association.
The overlap between the GODomainMiner predictions and InterPro is shown in the last row of these tables (here, a match at any GO level is counted as a common association). The high percentage of overlap between GODomainMiner and InterPro (from 91 % to more than 99%) reflects the fact that our method is calibrated to recover as many as possible correct InterPro associations. Nevertheless it also shows that a small percentage of the InterPro associations have consensus scores below our calculated score threshold, revealing the role of human rather than data-driven knowledge in the definition of such associations.

Dataset		AUC	Optimal Weights						F-measure		Threshold
						IEA					
			SIFTS	SP	TR	SIFTS	SP	TR	Training	Test	
MF	GO-Pfam	0.9605	1	1	6	10	10	10	0.926	0.924	0.005
	GO-CATH	0.9710	1	1	10	10	1	9	0.935	0.943	0.004
	GO-SCOP	0.9693	1	1	10	10	1	2	0.954	0.931	0.004
BP	GO-Pfam	0.9546	1	1	1	10	1	8	0.898	0.903	0.008
	GO-CATH	0.9726	1	1	1	10	1	5	0.922	0.938	0.007
	GO-SCOP	0.9756	1	1	1	10	1	3	0.943	0.939	0.007
CC	GO-Pfam	0.9228	1	1	6	10	1	10	0.871	0.866	0.003
	GO-CATH	0.9741	1	1	1	10	1	9	0.955	0.932	0.003
	GO-SCOP	0.9684	1	1	1	10	1	6	0.927	0.906	0.005

Table 3.1: Calculated AUCs, dataset weights, F-measures, and score thresholds for GO-domain associations for the 3 GO ontologies and 3 domain classifications studied here. Data source abbreviations are: SP for UniprotKB/SwissProt and TR for UniProtKB/TrEMBL.

Overall, our approach yields a total of 32,881 MF GO-Pfam associations (shown as 33×10^3 in Table 3.2) that include 3,968 associations already present in InterPro (2,657 specific term matches plus 1,311 parent term matches). This corresponds to an enrichment of about 8-fold in MF GO-Pfam associations. Similar calculations reveals enrichemnts of about 22 and 14-fold for MF GO terms associations with CATH and SCOP domain superfamilies, respectively. For BP GO terms we get 21-, 51- and 31-fold enrichments in associations with Pfam, CATH and SCOP domains, respectively, and for CC GO terms 19-, 62- and 32-fold enrichments, respectively.

3.2.4 Distribution of GO-Domain Associations per GO term or per domain

Figure 3.5(A) shows the average numbers of MF, BP, and CC GO-Pfam associations per GO term and Pfam entry, for associations in InterPro (green) and those calculated by GODomainMiner when counting the most-specific GO terms assigned to a domain (purple).

GODomainMiner generally predicts more associations per GO term and per Pfam domain than exist in InterPro. For example (top panel), GODomainMiner predicts that each MF GO term and each Pfam entry are associated with an average of 5.2 domains and 4.0 MF GO terms, respectively, compared to averages of 3.9 domains and 1.3 MF GO terms in InterPro, respectively. For BP and CC GO terms we see similar enrichments from GODomainMiner compared with InterPro, with ratios of 5.4 versus 3.5 and 16.9 versus 6.8 associations per GO term, and 8.2 versus 1.17 and 4.5 versus 1.1 associations per Pfam, respectively. These results demonstrate that GODomainminer produces a considerable enrichment in the number of annotations compared with InterPro. They also support the notion that many Pfam domains participate in different functions, either as singleton domains or as components of multi-domain proteins.

The bar charts in Figure 3.5(B) show the distributions of GO terms (shown in orange) and Pfam entries (in blue) according to the number of associations they are involved in. For example, considering the first two bars in part B, it can be seen that some 2,100 MF, 3,500 BP, and 320 CC GO terms and 2600, 2300, and 2,800 Pfam domains are involved in only one GO-Pfam association. The remainder of this figure shows that many GO terms and Pfam domains are involved in two or more associations, which supports the notion that complex many-to-many relationships exist between GO terms and domains

Dataset	GO-Domain Associations			MF GO Terms			Domain Entries		
	Pfam	CATH	SCOP	Pfam	CATH	SCOP	Pfam	CATH	SCOP
SIFTS	31	16	9.9	44	22	17	2.8	1.1	0.8
SIFTS-IEA	69	36	23	26	29	23	4.8	2.0	1.5
SwissProt	194	72	73	6.3	5.4	5.6	7.4	1.2	1.1
SwissProt-IEA	225	79	79	4.8	4.2	4.3	8.1	1.4	1.2
TrEMBL	215	104	96	4.0	3.4	3.5	7.4	1.2	1.0
TrEMBL-IEA	756	240	208	6.4	5.7	5.8	13	1.6	1.4
Merged	917	306	266	7.9	7.2	7.3	14	2.5	1.8
GODomainMiner	33	13	9.7	6.3	4.5	4.0	8.3	2.1	1.6
InterPro	4.226	0.607	0.743	1.076	0.273	0.301	3.300	0.466	0.584
Overlap	3.968	0.594	0.713	1.059	0.273	0.300	3.101	0.457	0.560

Table 3.2: The numbers of given and predicted MF GO-domain associations in thousands ($\times 10^3$).

Dataset	GO-Domain Associations			BP GO Terms			Domain Entries		
	Pfam	CATH	SCOP	Pfam	CATH	SCOP	Pfam	CATH	SCOP
SIFTS	182	90	53	9.8	8.5	6.8	2.7	1.1	0.7
SIFTS-IEA	197	109	70	7.6	6.8	5.7	4.9	2.1	1.5
SwissProt	1336	461	465	20	18	19	8.6	1.2	1.2
SwissProt-IEA	844	267	302	14	12.5	13	9.4	1.4	1.3
TrEMBL	837	360	337	13	12	12	8.3	1.2	1.1
TrEMBL-IEA	1756	623	548	18	17	17	12	1.6	1.3
Merged	2436	872	764	21	20	20	13	2.4	1.8
GODomainMiner	75	23	18	14	8.6	7.8	9.1	2.1	1.6
InterPro	3.829	0.461	0.586	1.094	0.206	0.244	3.265	0.388	0.491
Overlap	3.518	0.448	0.572	1.077	0.205	0.244	3.028	0.376	0.480

Table 3.3: The numbers of given and predicted BP GO-domain associations in thousands ($\times 10^3$).

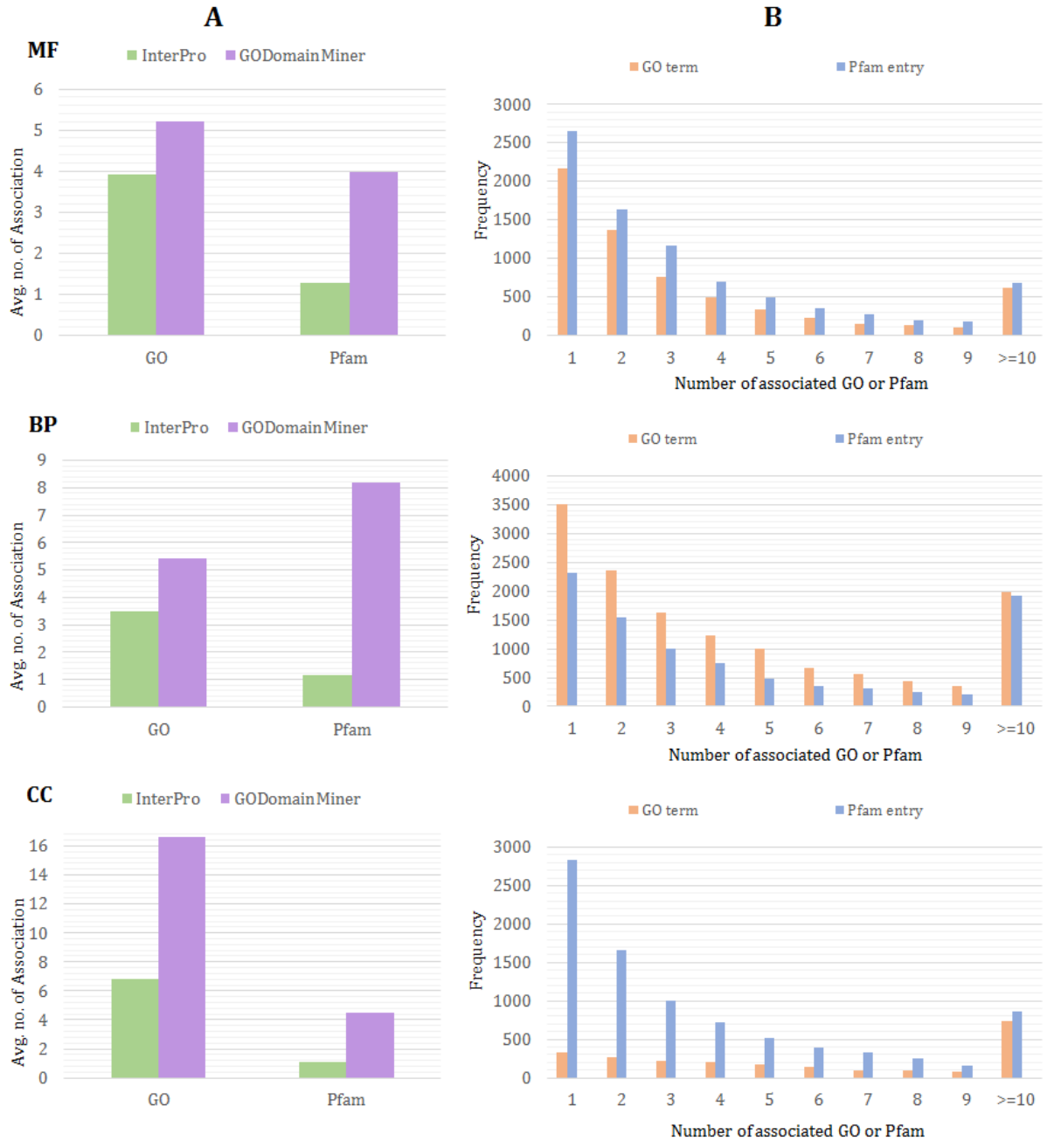


Figure 3.5: Distribution of GO-Pfam associations for the 3 GO ontologies (MF: top; BP: middle; CC: bottom). A: Average number of GO-Pfam associations per GO term and per Pfam entry for InterPro (green), and GODomainMiner (purple). B: Numbers of GO terms (orange) according to their numbers of associations with Pfam entries, and numbers of Pfam entries (blue) according to their numbers of associations with GO terms.

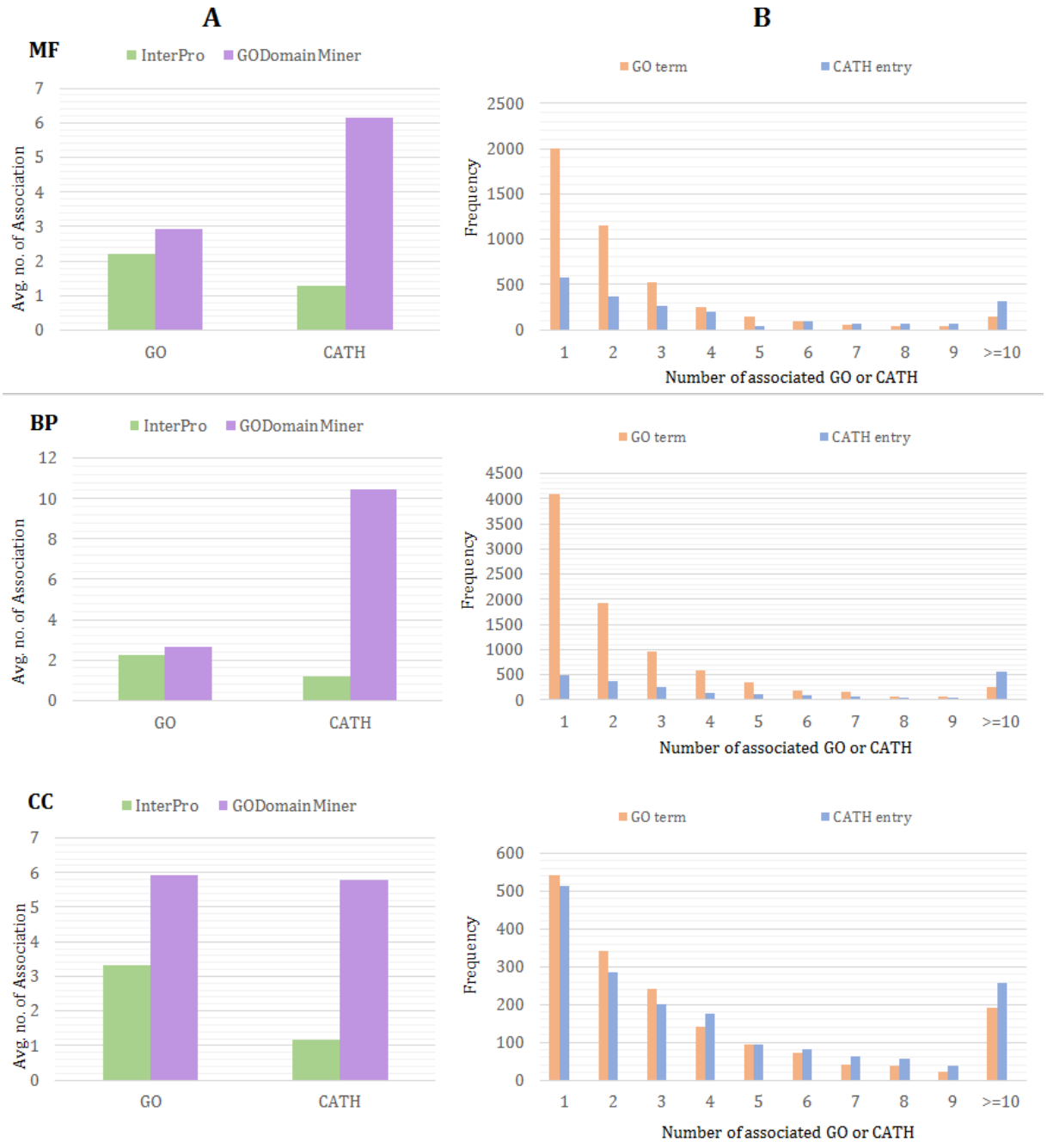


Figure 3.6: Distribution of GO-CATH associations for the 3 GO ontologies (MF: top; BP: middle; CC: bottom). A: Average number of GO-CATH associations per GO term and per CATH entry for InterPro (green), and GODomainMiner (purple). B: Numbers of GO terms (orange) according to their numbers of associations with CATH entries, and numbers of CATH entries (blue) according to their numbers of associations with GO terms.

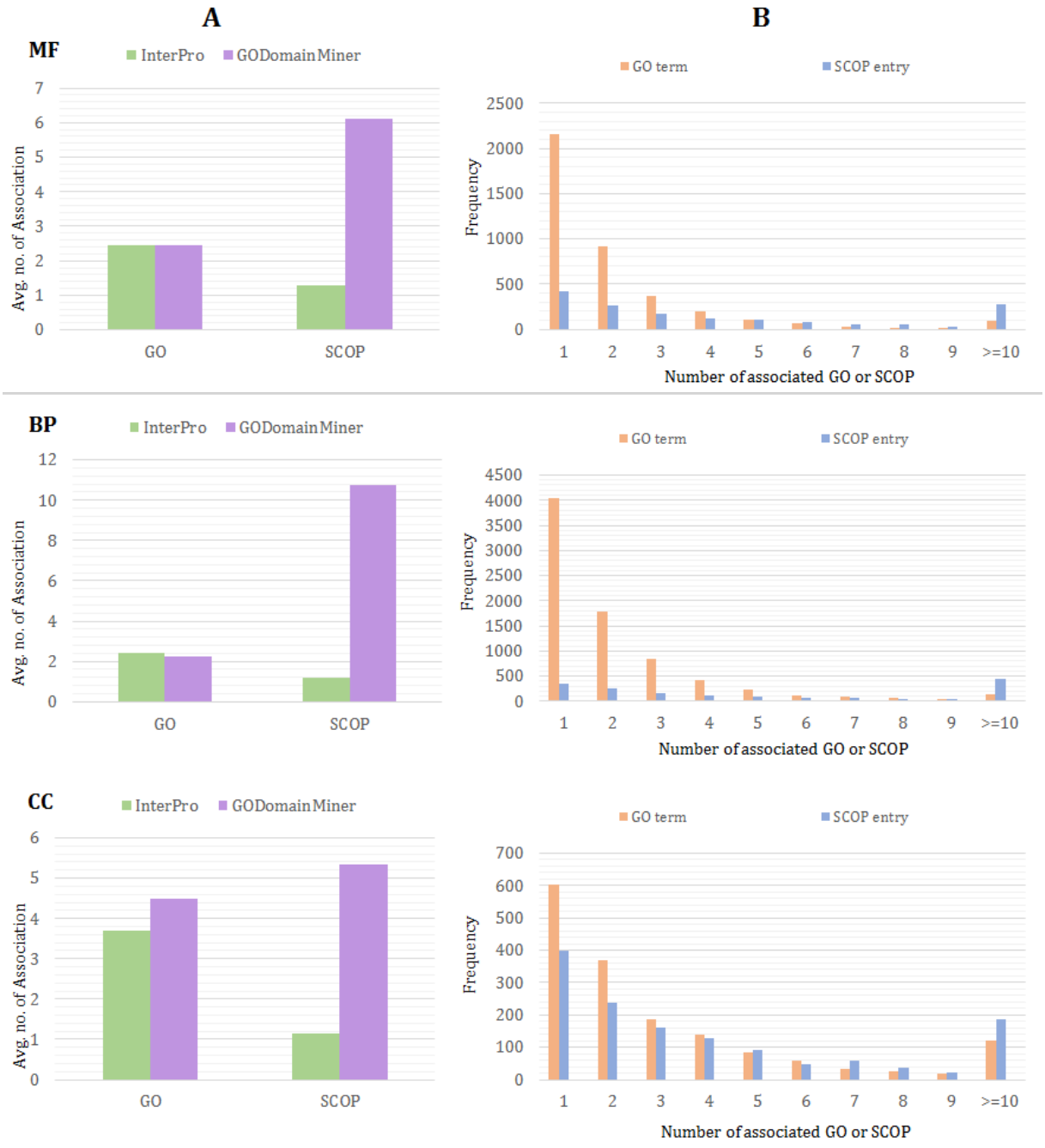


Figure 3.7: Distribution of GO-SCOP associations for the 3 GO ontologies (MF: top; BP: middle; CC: bottom). A: Average number of GO-SCOP associations per GO term and per SCOP entry for InterPro (green), and GODomainMiner (purple). B: Numbers of GO terms (orange) according to their numbers of associations with SCOP entries, and numbers of SCOP entries (blue) according to their numbers of associations with GO terms.

Dataset	GO-Domain Associations			CC GO Terms			Domain Entries		
	Pfam	CATH	SCOP	Pfam	CATH	SCOP	Pfam	CATH	SCOP
SIFTS	37	17	10	1.4	1.1	0.9	2.6	1.0	0.7
SIFTS-IEA	38	19	13	1.0	0.8	0.7	3.9	1.6	1.2
SwissProt	251	74	74	2.5	2.3	2.4	8.4	1.2	1.2
SwissProt-IEA	185	55	54	1.8	1.6	1.7	10	1.4	1.3
TrEMBL	179	67	61	1.7	1.6	1.6	7.9	1.2	1.1
TrEMBL-IEA	360	111	94	2.3	2.1	2.1	14	1.6	1.4
Merged	479	151	129	2.7	2.5	2.6	15	2.3	1.8
GODomainMiner	39	10	7.3	2.3	1.7	1.6	8.7	1.8	1.4
InterPro	2.289	0.192	0.237	0.336	0.058	0.064	2.042	0.163	0.208
Common with InterPro	2.085	0.191	0.230	0.335	0.058	0.064	1.878	0.163	0.202

Table 3.4: The numbers of given and predicted CC GO-domain associations in thousands ($\times 10^3$).

(Figure 3.4). Similar results for GO-CATH and GO-SCOP associations are shown in Figures 3.6 and 3.7, respectively.

Finally, Table 3.5 shows the distribution of GODomainMiner predicted associations according to our Gold, Silver, and Bronze classification, along with the degree of overlap with the InterPro reference dataset. Since the Gold class represents associations with statistically significant p-values, it is interesting

	GODomainMiner			Overlap with InterPro		
Class	MF	BP	CC	MF	BP	CC
Gold	15,605	24,782	12,967	1,815	1,378	887
Silver	11,098	31,920	17,062	778	865	628
Bronze	6,178	18,060	8,939	64	116	124
Total	32,881	74,762	38,968	2,657	2,239	1679

Table 3.5: The distribution of all most-specific GO-Pfam associations from GODomainMiner, and their overlap with InterPro, in the Gold, Silver, and Bronze categories.

to see that the majority (68%) of our predicted MF GO-Pfam associations common with InterPro fall in this class. Overall, we calculate that 47% of the GODomainMiner MF GO-Pfam associations and 33% of the predicted BP and CC associations are of Gold quality. The quality of GO predictions for CATH and SCOP classifications also follow very similar paths (see Tables 3.6 and 3.7).

3.2.5 Comparison with GO-Domain Associations from dcGO

In order to compare the GODomainMiner results with those obtained from dcGO [Fang and Gough, 2013], we extracted the Pfam2GO associations from the dcGO website (http://supfam.org/SUPERFAMILY/dcGO/). To avoid the complexity of comparing GO annotations at different levels in the rDAG, our comparison mainly focuses on GO-domain associations in which GO terms are leaves of the GO rDAG. GODomainMiner contains a total of 515,582 GO-Pfam associations regardless of their level in GO hierarchy, of which 79,589 involve leaf GO terms (comprising 21,410 MF, 36,814 BP, and 21,365 CC GO-Pfam asso-

	GODomainMiner			Overlap with InterPro		
Class	MF	BP	CC	MF	BP	CC
Gold	7,238	9,248	3,774	257	174	84
Silver	4,256	8,525	4,139	92	80	67
Bronze	1,558	5,020	2,288	9	16	7
Total	13,052	22,793	10,201	358	270	158

Table 3.6: The distribution of all most-specific GO-CATH associations from GODomainMiner, and their overlap with InterPro, in the Gold, Silver, and Bronze categories.

	GODomainMiner			Overlap with InterPro		
Class	MF	BP	CC	MF	BP	CC
Gold	5,181	6,219	2,723	278	189	99
Silver	3,452	7,315	3,159	133	123	83
Bronze	1,070	4,182	1,455	9	24	6
Total	9,703	17,716	7,337	420	336	188

Table 3.7: The distribution of all most-specific GO-SCOP associations from GODomainMiner, and their overlap with InterPro, in the Gold, Silver, and Bronze categories.

ciations). The Pfam2GO dataset from dcGO contains a total of 720,534 associations, of which 62,779 involve leaf GO terms (comprising 5,939 MF, 24,334 BP, and 32,506 CC associations). Thus, the numbers of associations in GODomainMiner and Pfam2GO are broadly comparable. However, when considering the leaf levels of all 3 ontologies, Figure 3.8 shows that only 11,138 GO-Pfam associations are common between GODomainMiner and dcGO (overlap region B, about 14% of the GODomainMiner set and 18% of the dcGO set). Looking at the overlap with InterPro, which contains 2,799 leaf level GO-Pfam associations, GODomainMiner shares 2,744 associations (98%) with InterPro, while dcGO shares only 724 associations (26%; overlap C). This shows that GODomainMiner gives a greater coverage of the InterPro reference set than dcGO. Although this is perhaps not surprising since InterPro was used to calibrate GODomainMiner, the high agreement between GODomainMiner and InterPro gives a good indication of the reliability of other associations predicted by GODomainMiner.

We also compared GO-SCOP associations predicted by GODomainMiner with the SCOP2GO database from dcGO and with InterPro. Overall, GODomainMiner calculates a total of 19,708 leaf GO-SCOP associations, compared to 2,445 such associations in SCOP2GO and 422 in InterPro. Of these, 845 GO-SCOP associations are common to GODomainMiner and SCOP2GO. Also, 421 (i.e. *99.75%* of InterPro set) GODomainMiner associations overlap with InterPro, whereas only 55 (*13%* of InterPro set) SCOP2GO associations from dcGO are found in InterPro. This confirms the trend observed for GO-Pfam associations, in favor of a much better coverage by GODomainMiner than by dcGO, of the InterPro reference set.

3.2.6 Biological Assessment of New Discovered GO-Pfam Associations

It would certainly be a very tedious task to validate manually the huge number of new GO-domain associations proposed by the GODomainMiner approach. For this reason, we decided to check manually a small subset of these associations, namely the one-to-one GO-domain associations in which the GO

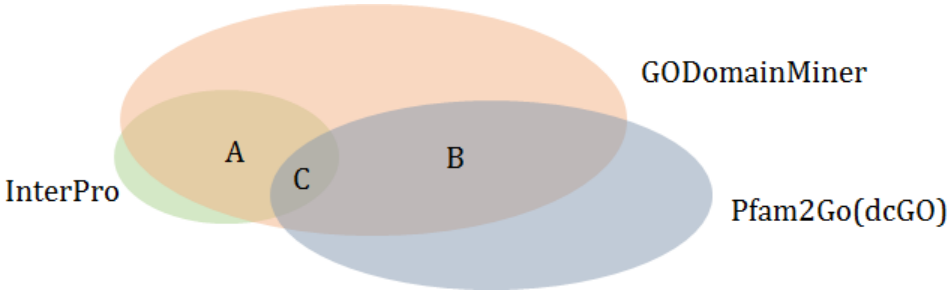


Figure 3.8: Venn diagram showing the intersections between leaf GO-Pfam associations from Pfam2GO (62,779 associations), GODomainMiner (79,589), and manually curated associations from InterPro (2,799). Region A (2,744 associations) is the overlap between GODomainMiner and InterPro. Region B (11,138 associations) is the overlap between GODomainMiner and Pfam2GO. Region C (724 associations) is the overlap between Pfam2GO and InterPro.

term is uniquely associated with one domain, which is itself uniquely associated with that GO term. Such one-to-one associations can easily be used to assess the novelty and biological consistency of knowledge discovered through our approach. All lists of one-to-one associations found in the 9 settings of this study are available on the GODomainMiner website.

For the sake of brievity, we review here only the MF GO-Pfam one-to-one associations. We obtained 125 one-to-one MF GO-Pfam associations with consensus scores ranging from 0.9704 to 0.0052, 75 associations in the gold category (all p-values significant), 30 and 20 in the silver and bronze categories, respectively. From the 125 associations, 30 are already known in InterPro (21 from the gold category) and 95 are new (54 from the gold category). Manual checking of the MF GO terms and Pfam domain names led us to distinguish 5 situations (see the examples in Table 3.8). (i) The MF GO terms and Pfam domains descriptions are almost identical (34 associations). Such associations are trivial but only 16 of them are reported in InterPro, probably because the remaining 18 escaped automatic retrieval due to unpredictable spelling differences. (ii) The MF GO term is more specific than the Pfam domain description (21 associations including 3 from InterPro). (iii) The Pfam description is more specific than the MF GO term (11 associations including 3 from InterPro). (iv) The MF GO term and the Pfam descriptions are quite different (51 associations including 8 from InterPro). Such associations are likely the most interesting to provide to the expert for further analyses. (v) The Pfam domain has no known function (8 associations not present in InterPro). These 8 associations are listed in Table 3.8 as examples of new knowledge discovered by the CODAC approach.

We expect that many further novel associations between MF GO terms and yet uncharacterized domains may be mined from the complete MF GO-Pfam dataset which contains more than 3,400 associations concerning so-called DUF (Domain of Unknown Function) or UPF (Uncharacterized Protein Family) Pfam domains.

Concerning the strict many-to-one MF GO-Pfam associations, we identified 30 such Pfam domains, most of which have only two associated GO terms. This results in 55 associations of which 7 are known in InterPro (6 gold and 1 bronze)and 48 are new (33 gold, 8 silver and 7 bronze). For one Pfam domain only (CobS,PF02654) the two GO terms are known already in InterPro. For 5 other Pfam domains, one of the GO terms is known in InterPro and the other one is new. New MF GO-Pfam associations generally give lower scores than known InterPro associations. However, in some cases this suggests an

MF GO ID	MF GO term	Pfam ID	Pfam description	Consensus Score	Class
Case (i) : Trivial but not in InterPro					
GO:0008437	thyrotropin-releasing hormone activity	PF05438	Thyrotropin-releasing hormone (TRH)	0.0638	gold
Case (ii) MF GO term more specific than Pfam description					
GO:0098640	integrin binding involved in cell-matrix adhesion	PF09085	Adhesion molecule, immunoglobulin-like	0.0752	gold
Case (iii) Pfam description more specific than MF GO term					
GO:1990919	nuclear membrane proteasome anchor	PF08559	Cut8, nuclear proteasome tether protein	0.0309	gold
Case (iv) MF GO term and Pfam description differ					
GO:0047991	hydroxylamine oxidase activity	PF13447	Seven times multi-haem cytochrome CxxCH	0.2654	gold
Case (v) Domains of yet unknown function					
GO:1990838	poly(U)-specific exoribonuclease , activity producing 3' uridine cyclic phosphate ends	PF09749	Uncharacterized conserved protein	0.0235	gold
GO:0030144	alpha-1,6-mannosylglycoprotein 6-beta-N-acetylglucosaminyl transferase activity	PF15027	Domain of unknown function (DUF4525)	0.5273	silver
GO:0030735	carnosine N-methyltransferase activity	PF07942	N2227-like protein	0.2705	silver
GO:0010340	carboxyl-O-methyltransferase activity	PF04301	Protein of unknown function (DUF452)	0.0201	silver
GO:0016772	transferase activity, transferring phosphorus-containing groups	PF01989	Protein of unknown function DUF126	0.0137	silver
GO:0071617	lysophospholipid acyltransferase activity	PF10998	Protein of unknown function (DUF2838)	0.0072	silver
GO:0015666	restriction endodeoxyribonuclease activity	PF12102	Domain of unknown function (DUF3578)	0.0111	bronze
GO:0016841	ammonia-lyase activity	PF11807	Domain of unknown function (DUF3328)	0.0066	bronze

Table 3.8: Selected examples of new one-to-one MF GO-Pfam associations. All of these examples are absent in InterPro; additional examples are available from the GODomainMiner website for cases (i) to (iv)).

alternative substrate for the domain activity which may be interesting to investigate. For example, for Pfam domain Mqo (PF06039 Malate:quinone oxidoreductase), GO:0052589 (malate dehydrogenase (menaquinone) activity) is found in addition to GO:0008924 (malate dehydro-genase (quinone) activity). The remaining 24 Pfam domains all have new GO MFannotations that do not exist in InterPro. Interestingly, in some cases a different more general InterPro annotation exists, as in the case of PF07722 domain Peptidase_C2 which GODomainMiner associates with GO:0034722 (gamma-glutamyl-peptidase activity) and with GO:0033969 (gamma-glutamyl-gamma-aminobutyrate hydrolase) activity, whereas the InterPro annotation is simply GO:0016787 (hydro-lase activity).

3.3 Implementation

The CODAC method was written mainly with the Python script. Python is a widely used high-level programming language for general-purpose programming which is very suitable for quick prototyping as well as creating a robust application software. Linux shell scripts were additionally used for handling certain modules such as downloading and extracting files. MySQL, an open-source relational database management system (RDBMS), database was used to store the inferred associations in a structure way. Database queries are principally processed by MySQL Connector developed by the MySQL community.

HTML, CSS, PHP and Javascript languages are used for online presentation of the discovered associations. PHP, a server-side scripting language designed primarily for web development is used for processing data and querying database from MySQL. jQuery, a cross-platform JavaScript library designed to simplify the client-side scripting of HTML, is also used for result presentation in ECDomainMiner and GODomainMiner. HTML and CSS are used for building the general structure of the web-servers. ClustrMaps is free embedded plug-in in our websites to instantly discover where our visitors are accessing. It has several features such as audience geo-location heatmap to highlight the areas in which our websites are popular, and the total number of visits originated from there.

The web interface (Figure A.2) has been tested using several popular browsers for the Windows, Linux, and Mac OS X operating systems.

Here, the algorithm complexity of different phases of the GODomainMiner is presented. We separate the GODomainMiner algorithm into five phases. First, reading phase which its complexity is calculated based on the reading time of SIFTS, SwissProt and TrEMBL. Because reading of a flat file is carried out with a linear algorithm, the reading phase complexity is linear, highly depends on the size of the largest input sources ($\mathcal{O}(s)$). Second, enrichment phase of the input sources including hierarchy usage and the clustering of identical common neighbors. The complexity of using hierarchy is $\mathcal{O}(n \times s)$ where the n is the size of the available GO terms, and s is the size of the sequences (common neighbors). The clustering complexity is $\mathcal{O}(s \times c)$, where the s is the size of the sequences, and c is the size of clusters. Due to the huge size of the clusters and sequences, this it the most time-consuming (bottleneck) phase of the system. Third, the complexity of the consensus score computation is based on the size of the GO terms and domains, ($\mathcal{O}(n \times m)$). Similarly, the complexity of fourth phase to calculate p-values is $\mathcal{O}(n \times m)$. Last but not least, fifth phase is classification which is carried out in linear time size of the associations $\mathcal{O}(a)$.

It should be mentioned that running one time GODomainMiner to find associations between MF GO terms and Pfam domains takes $\approx$ 8 hours with only one processor.

3.4 Conclusion

We have presented a systematic approach called CODAC for mining associations from datasets that can be represented as tripartite graphs. We have presented one implementation of this approach called GODomainMiner, for predicting associations between GO terms and protein domains (Figure 3.9).

This was achieved by first collecting existing Pfam, CATH, and SCOP domain annotations of protein chains and sequences on one hand and MF, BP, and CC GO term annotations on the other. We then applied our method to find a list of direct associations between GO terms and domains. Considering only the most-specific GO terms, our approach yields an enrichment of about 15-fold in the number of GO-Pfam associations that currently exist in InterPro. A selected subset of one-to-one associations has been analyzed from a biological point of view, and these all appear to be highly meaningful and consistent with available knowledge. We believe that the large numbers of GO-domain associations calculated here can enrich the existing annotations of UniProt sequences and protein chains in the PDB, and that this will facilitate a better understanding and exploitation of protein structure-function relationships at the domain level.

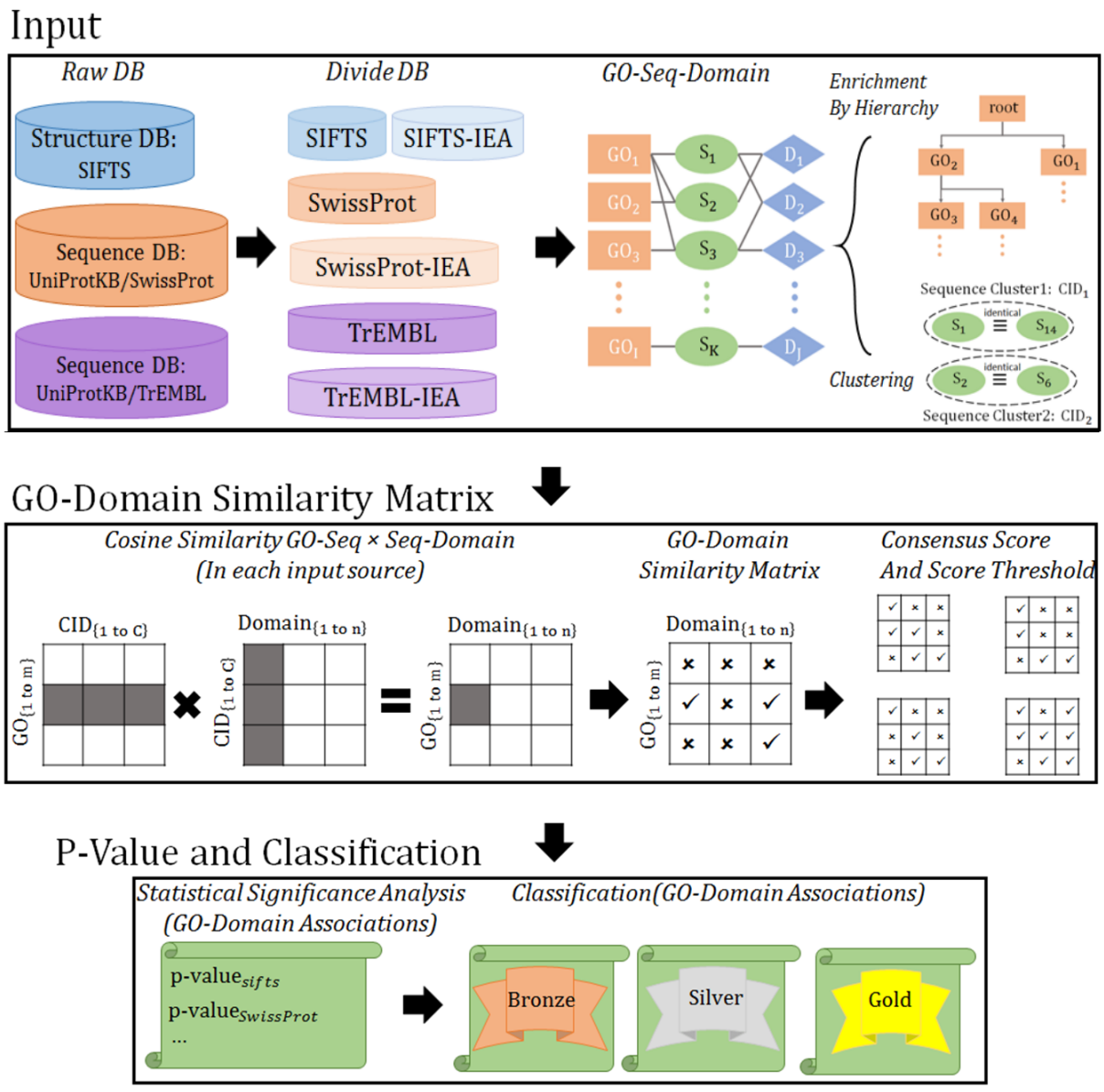


Figure 3.9: GODomainMiner Flowchart. It starts with reading input sources and dividing them based on the GO annotation evidence code. Then, input sources are enriched by the hierarchical information of GO, and sequence clustering. Cosine similarity is used to discover the associations between GO terms and domains in each source. It follows by combining similarity scores of each GO-Domain from different sources into a consensus score. The procedure ends with calculating p-values of GO-Domain associations and their classification.

Chapter 4

Functional Annotation of Protein Sequences and Structures

Contents

There are millions of proteins with known sequences and unknown functions. The most reliable way to assign functions to proteins is by expert curators, but this is an expensive and time-consuming process. The huge gap between the small number of expert curators and the ever increasing number of new unannotated protein sequences has motivated the development of many automatic annotation approaches. These approaches aim for a balance between maximizing the number of annotations while minimizing the number of false assignments. However, achieving this aim in a reliable way remains an open research problem. We present here a novel approach called CARDM (Combinatorial Association Rules Domain Miner) which exploits that fact that many proteins consist of one or more domains. CARDM combines a learning step in which functional annotations are assigned to protein domains, and a combinatorial step in which association rules are generated and filtered using previously validated annotations. The

filtered rules are then aggregated to build predictive models that are used to automatically annotate protein sequences and structures. CARDM has been tested on the entire set of TrEMBL entries and on the dataset provided at the international 2013 CAFA (Critical Assessment of Functional Annotation) challenge. Overall, CARDM predicts 24 million EC numbers and 188 million GO terms for the protein entries in TrEMBL. We find that the performance of CARDM on the CAFA 2013 targets is similar to that of the best predictor groups in that round of CAFA. All predicted associations made by CARDM are available at http://cardm.loria.fr/

4.1 Introduction

The functional annotation of proteins is crucially important for a better understanding of biological processes at the molecular level, and has considerable implications in biomedical and pharmaceutical research. However, the experimental characterization of proteins cannot easily be scaled up because this is a difficult and costly process [Liolios et al., 2009]. Furthermore, the curation and annotation of existing protein sequences by expert curators is almost equally expensive and time-consuming. Thus, the automatic annotation of protein function has become a critical computational problem in bioinformatics [Radivojac et al., 2013]. During the past decade, several protein function prediction approaches have been described [Bork et al., 1998, Rost et al., 2003, Watson et al., 2005, Friedberg, 2006, Sharan et al., 2007, Lee et al., 2007, Punta and Ofran, 2008, Rentzsch and Orengo, 2009, Xin and Radivojac, 2011]. Most approaches use BLAST [Altschul et al., 1997] to compare the sequences of new proteins with proteins whose function have previously been determined experimentally, while some others apply similar principles at the domain level.

In recent years, high-throughput experimental data acquisition techniques for the genomic sequences of many species has opened new possibilities for automatic protein function prediction. For instance, methods using protein-protein interaction networks may assign functional classes to proteins from their physical interaction networks [Vazquez et al., 2003]. Other approaches exploit information from combinations of protein domains and domain interactions [Peng et al., 2014]. Gene expression and molecular interaction data may also be used to create a network of functionally connected genes from which functional annotation may be propagated across the network [Massjouni et al., 2006], and taxonomy information may be used to filter false predictions [Zhu et al., 2007]. Applying machine learning to evolutionary relationships between gene products and genomic contexts is another way to infer protein function annotations. [Enault et al., 2005, Li et al., 2007]. Machine learning techniques are also used to identify and extract functional features from representative proteins, and to propagate functions to unknown proteins. Such methods typically use probabilistic techniques to extract functions from protein interaction networks [Nariai et al., 2007] or phylogenetic information [Engelhardt et al., 2005]. Other approach uses association rule mining techniques to construct rule-based predictive models [Boudellioua et al., 2016].

Protein structural information can also be used to aid function annotation. For example, in [Roy et al., 2012] template proteins having similar folds and functional sites are created, and a target protein is then compared to the closest homologous template. Because the three-dimensional structures of proteins are often more evolutionary conserved than their sequences, using structural templates is an accurate way to find similar functions in different protein sequences [Whisstock and Lesk, 2003]. However, template-based algorithms will fail if no homologous template is available. Hybrid methods can predict protein functions based on learning and finding consensus scores computed from a combination of different protein sources [Hooper et al., 2014] or from a mixture of different methods in order to return a ranked list of annotations

[You et al., 2017].

Many protein function prediction methods use Gene Ontology (GO) [Harris et al., 2004] definitions to describe protein functions. The GO vocabulary is divided into three namespaces that may be used to describe the biological process (BP), molecular function (MF), and cellular component (CC) of a protein. At the molecular level, specific functions are often carried out in highly conserved domains, which may be identified by sequence or structure alignments and which may be classified into domain families. Several functional annotation methods use protein domain families as the basic unit of protein similarity [Peng et al., 2014, Forslund and Sonnhammer, 2008]. Nonetheless, despite the wide variety of existing function annotation techniques, protein function prediction is still an open problem because no universal method exists which clearly provides the best functional annotations. In response to this need, the CAFA (Critical Assessment of protein Function Annotation) experiment [Radivojac et al., 2013] was launched to assess the current state of the art in protein function annotation and to encourage developments in the field.

We previously described a machine learning algorithm called CODAC (Computational discovery of Domain Annotation using Common neighbors) [Alborzi et al., 2018], which we used to assign Enzyme Commission (EC) numbers [Webb et al., 1992] and GO annotations to un-annotated protein domains [Alborzi et al., 2017b, Alborzi et al., 2017c]. It quickly became apparent to us that this approach could also be usefully applied to the automatic functional annotation of protein sequences. This led us to develop an extension of CODAC which we call CARDM (Combinatorial Association Rules Domain Miner). CARDM combines the CODAC learning step, in which function annotations are associated with protein domains, with a combinatorial rule generation and filtering procedure from which aggregated taxon-specific predictive models are constructed and used to annotate protein sequences and structures automatically.

Here, we describe the CARDM approach and its application to EC and GO annotations. The EC annotation models obtained have been applied to the entire TrEMBL database, and our results are compared with those from several existing automatic annotation methods. We also present results from applying CARDM to the three GO namespaces. The generated MF, BP, and CC annotation models have been applied to the target sequences of the 2013 CAFA challenge. We mention here that we also used preliminary GO annotation models in the 2017 CAFA experiment [Alborzi et al., 2017a]. However at the time of writing, the evaluation of this CAFA edition has not yet been published.

4.2 Methods

4.2.1 Method Overview

CARDM aims to create efficient association rules for predicting the functions of protein sequences and structures. The method exploits function-domain associations inferred by our previously developed CODAC method using manually curated information from the UniProtKB and SIFTS databases, and it uses a small set of annotations from the InterPro database [Finn et al., 2016a] as a "Gold Standard". InterPro provides an integrated classification of protein sequences and domains, and links out to many other classification systems. Several InterPro families have been manually annotated with GO terms using expert knowledge and the literature. However, the list of such annotations is incomplete (only around 20% of Pfam domains and families possess MF GO functional annotation).

UniProtKB consists of two disjoint sets of entries. UniProtKB/SwissProt is the high quality, non-redundant, and manually curated section of UniProtKB, while the much larger UniProtKB/TrEMBL

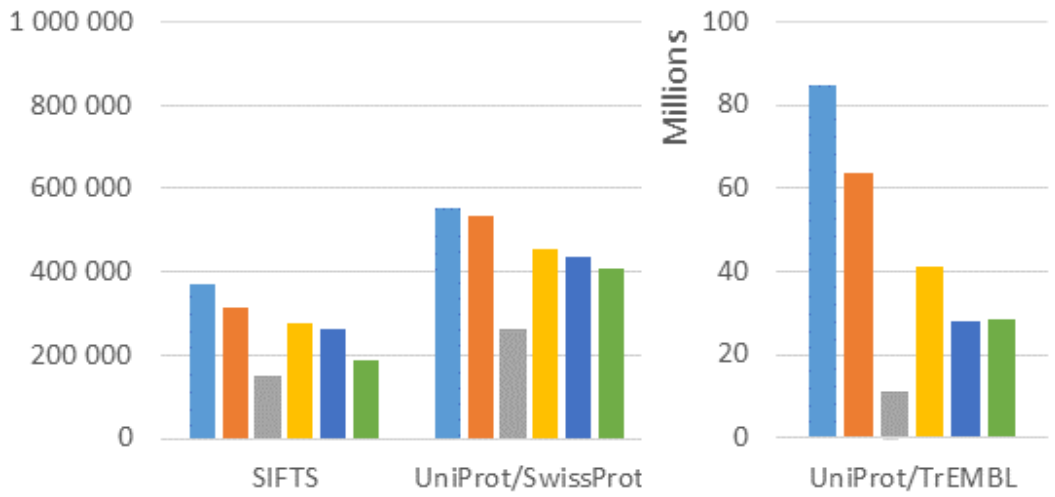


Figure 4.1: Database statistics for SIFTS, SwissProt, and TrEMBL (July 2017 versions). Light blue: total number of entries in SIFTS (369,521), SwissProt (554,241), TrEMBL (84,827,567). Orange: number of entries having at least one domain identified in a reference domain classification (316,265, 534,235, 63,684,389, respectively). Grey: number having at least one EC annotation (150,264, 261,610, 10,933,166). Yellow: number having at least one MF GO annotation (276,340, 454,115, 40,931,904). Dark blue: number having at least one BP GO annotation (261,672, 437,411, 27,930,466). Green: number having at least one CC GO annotation (188,211, 405,636, 28,397,194).

contains automatically annotated and unreviewed protein entries. The SIFTS (Structure Integration with Function, Taxonomy and Sequence) database contains manually curated cross-references between protein chains in the Protein Data Bank (PDB) with functional annotations from biological sequence databases [Gutmanas et al., 2014].

CARDM consists of three main steps, namely learning, modeling, and annotation. The learning step uses CODAC to infer function-domain associations from SwissProt (although any other reliable source of annotated sequences could equally be used instead). The modeling step involves three stages: (i) combinatorially generating association rules involving domains and taxons in each rule antecedent (left-hand-side) and a function (EC number or GO term) in the rule consequent (right-hand-side), (ii) filtering these rules using parameters learnt from SwissProt, and (iii) creating predictive annotation models by rule aggregation. The annotation step assigns EC or GO annotations to those target protein entries that match at least one predictive model. Thus, SIFTS and SwissProt provide appropriate data sources for the learning and modeling steps, whereas TrEMBL contains many targets for the annotation step.

Figure 4.1 shows that the majority of the SIFTS and SwissProt entries are annotated with at least one GO term and one or more domains from our nine selected domain classifications (SIFTS 66%, SwissProt 78%, on average), but that over 50% of these entries lack any EC annotation (59% and 53%, respectively). This figure also shows that 75% of TrEMBL entries have at least one domain assigned from the nine domain classifications, while only around 13% and 38% are annotated with EC numbers and GO terms, respectively.

	Pfam	CATH	SCOP	TIGRFAMs	SMART	Panther	PRINTS	CDD	PROSITE
EC number									
Inferred	22,894	9,888	9,325	8,234	3,935	5,579	3,472	3,596	11,143
InterPro	8,442	1,297	1,144	5,640	814	3,857	1,488	2,058	2,622
MF GO term									
Inferred	132,999	39,096	38,437	34,548	36,741	59,141	35,758	19,110	68,419
InterPro	19,265	2,488	2,893	15,172	3,013	18,556	11,326	4,405	7,149
BP GO term									
Inferred	777,699	207,596	208,602	90,823	275,097	359,920	227,032	91,202	440,068
InterPro	48,128	5,669	7,155	38,666	5,270	55,610	14,770	10,831	10,844
CC GO term									
Inferred	136,917	31,192	31,231	14,651	38,731	67,971	32,132	14,281	62,740
InterPro	9,075	840	998	3,305	833	11,462	3,656	997	1,551

Table 4.1: Number of inferred function-domain associations using the CODAC method on chain and sequence EC and GO annotations extracted from SIFTS and SwissProt for each of the nine classifications studied here. The numbers of reference associations present in InterPro are also indicated. For both inferred and InterPro, associations are extended to the ancestor levels.

4.2.2 Using CODAC to Infer Function-Domain Associations

Our CODAC approach has been described previously [Alborzi et al., 2018]. Briefly, the general principle is to discover candidate function-domain associations by treating the input data as a tripartite graph, $\mathcal{G}(X, Y, Z, E)$, where X and Y are annotations (e.g. EC numbers and domain families), and Z is a common attribute (here, cluster of sequences). A new edge (association), E, may be inferred between X and Y whenever X and Y are found to share a common Z. In the present work, we use nine domain classifications, namely Pfam [Finn et al., 2013], CATH [Orengo et al., 1997], SCOP [Murzin et al., 1995], TIGRFAMs [Haft et al., 2012], SMART [Letunic et al., 2014], PANTHER [Mi et al., 2017], PRINTS [Attwood et al., 2003], CDD [Marchler-Bauer et al., 2016], and PROSITE [Sigrist et al., 200 and we use sequences from SIFTS and SwissProt as sources of common neighbors (Z). The CODAC scores for each function-domain association from each data source are combined using a weighted average. The weights are optimized by calculating the area under the curve (AUC) of receiver-operator-characteristic (ROC) plots, and by maximizing the AUC with respect to a "Gold Standard" set of associations extracted from InterPro. Then, a score threshold is chosen in order to eliminate weak associations. Finally, the statistical significance (p-value) of each score is calculated for each association for each data source using a hypergeometric distribution as the null hypothesis. The CODAC association scores and p-values are then used to classify the inferred associations into one of three categories, namely "Gold", "Silver" or "Bronze" [Alborzi et al., 2018].

Table 4.1 summarises the results obtained by CODAC for the prediction of EC-domain and GO-domain associations from SIFTS and SwissProt for the nine domain classifications used here. Only Gold associations (CODAC score above the threshold and all p-values significant) have been counted. These associations provide the input data for the subsequent rule-based modeling step, as described below.

Rule ID	Antecedent	Consequent
$Rule_1$	$\{\{d_1\}, T_1\}$	EC_1
$Rule_2$	$\{\{d_2\}, T_1\}$	EC_2
$Rule_3$	$\{\{d_2\}, T_2\}$	EC_2
$Rule_4$	$\{\{d_3, d_4, d_5\}, T_3\}$	EC_3
$Rule_5$	$\{\{d_6, d_7\}, T_3\}$	EC_3

Table 4.2: Examples of generated association rules. $Rule_1$ says that a single domain from one taxa is responsible for a particular function (EC number). Taken together, $Rule_2$ and $Rule_3$ say that a particular domain in any one of two taxa is responsible for a particular function. Similarly, $Rule_4$ and $Rule_5$ say that the presence of different combinations of domains in a given taxon can be associated with a particular function.

4.2.3 Combinatorial Generation of Association Rules

The three main stages of the CARDM association rule modeling step are summarised in Algorithm 4 for a generic type of annotation *Func*, associated with domain d and represented by valid annotations from a reference data source SP (here, SwissProt). The procedure is described here for EC annotation but it may equally be applied to the three GO namespaces (MF, BP and CC). The inputs to the rule generation step are the EC-domain association datasets from CODAC. Associations are grouped to give a relation between each EC number, EC_k, and a set of domains from one or more of the nine classifications used here. For each of these grouped associations, all possible subsets containing up to 3 domains are generated ($\{d_1, d_2, ..., d_n\}, n \leq 3$). The subsets of domains are diversified by adding a taxon (T_j, one per subset) from a list of interest. These relations may be represented as a tuple $(\{\{d_1, d_2..., d_n\}, T_j\}, EC_k)$. Each generated tuple is then used to make a candidate association rule having $\{\{d_1, d_2..., d_n\}, T_j\}$ as the antecedent and EC_k as the consequent. Such association rules may be read a follows: *"IF a protein sequence contains the set of domains $\{d_1, d_2..., d_n\}$ AND derives from an organism of taxon T_j, THEN it can be annotated with EC_k."* Table 4.2 illustrates the different kinds of rules that may be generated.

4.2.4 Knowledge-based Filtering of Association Rules

Many of the candidate rules will have little or no support in the actual data, and such rules should be discarded. Hence the next stage is to filter the huge amount of generated association rules using annotations from SwissProt. We achieve this using three common rule mining metrics, namely "Support", "Confidence", and "Lift". The Support of a rule indicates how frequently the antecedent and the consequent appear together in the dataset. Support is calculated as the number of protein entries p, containing both the antecedent (*ante*) and the consequent (*cons*) of a rule divided by the total number of SwissProt entries $|SP|$.

$$Support_{SP}(R_i) = \frac{|p \in SP; ante_i \subseteq p \wedge cons_i \subseteq p|}{|SP|} \tag{4.1}$$

Because $|SP|$ is very large ($> 554,000$) the support ratio can be very low if only one protein entry matches a given rule. Therefore, we replace Support by $SupportCount(R_i)$, which simply counts the number of proteins that match a given rule.

The Confidence of a rule indicates how often the rule is found to be true in a given dataset, and is expressed as the ratio of the number of instances matching both antecedent and consequent with respect

Algorithm 4 CARDM core algorithm

Require: $\{(Func, d)\}_i$: sets of pairwise function-domain associations inferred by CODAC from different domain classifications; SP: a reliable source of functional annotations (e.g. SwissProt); T: a list of taxons present in SP.

Ensure: Annotation models for each function present in the input sets of associations.

AssociationRuleGeneration$(\{(Func, d)\}_i, T)$
AssociationRuleFiltering$(Rules, SP, Thresholds : T_{SC}, T_{Conf}, T_{Lift})$
AnnotationModelConstruction$(FilteredRules)$

procedure $AssociationRuleGeneration(\{(Func, d)\}_i, T)$
    for `each Func` **do**
        $DomainList \leftarrow GroupDomains(\{(Func, d)\}_i)$
        $DomainSubsets \leftarrow GenerateSubset(DomainList, Size \leq 3)$
        for `each S` $\in DomainSubsets$ **do**
            $Ante \leftarrow AddTaxon(S, T)$
            $Cons \leftarrow Func$
            $Rules \leftarrow Rules + Rule(Ante, Cons)$
        end for
    end for
    return$(Rules)$
end procedure

procedure $AssociationRuleFiltering(Rules, SP, Thresholds : T_{SC}, T_{Conf}, T_{Lift})$
    for `each R` $\in Rules$ **do**
        $SC \leftarrow SupportCount(R, SP)$
        $Conf \leftarrow Confidence(R, SP)$
        $Lift \leftarrow Lift(R, SP)$
        if $SC \geq T_{SC} \wedge Conf \geq T_{Conf} \wedge Lift > T_{Lift}$ **then**
            $FilteredRules \leftarrow FilteredRules + R$
        end if
    end for
    return$(FilteredRules)$
end procedure

procedure $AnnotationModelConstruction(FilteredRules)$
    for `each Func` **do**
        for `each R` $\in FilteredRules$ **do**
            if $Cons(R) = Func$ **then**
                $AggregAnte \leftarrow AggregAnte + Ante(R)$
            end if
        end for
        $AnnotationModel = (AggregAnte \Rightarrow Func)$
    end for
    return$(\{AnnotationModel\})$
end procedure

to the number of instances matching the antecedent. Here Confidence is calculated as

$$Conf_{SP}(R_i) = \frac{|p \in SP; ante_i \subseteq p \wedge cons_i \subseteq p|}{|p \in SP; ante_i \subseteq p|}. \tag{4.2}$$

The Lift of a rule measures the dependence of an antecedent on its consequent. The Lift of rule R_i is calculated as the ratio of the support of the rule in a given dataset to the product of the Supports of the antecedent and the consequent. Rules with Lift greater than 1 are considered stronger than random. In our setting, Lift is calculated as

$$Lift_{SP}(R_i) = \frac{|p \in SwissProt; ante_i \subseteq p \wedge cons_i \subseteq p| \times |SP|}{|p \in SP; ante_i \subseteq p| \times |p \in SP; cons_i \subseteq p|} \tag{4.3}$$

These metrics are calculated for each generated rule, and a rule is retained if it (i) is verified in SwissProt, (ii) has high Confidence, and (iii) has a high Lift.

When predicting functional annotation, quality of annotations is an important criterion and only high-confidence rules should be used. In order to eliminate rules that might represent random associations, we set a threshold of 1.0 for the Lift value. Furthermore, in order to be consistent with existing annotation systems in TrEMBL, we set the rule Confidence threshold to 0.95. This means that the filtered annotation rules should provide predictions which agree with existing SwissProt annotations in at least 95% of cases.

Using these fixed parameters, a range of threshold values for the SupportCount (from 1 to 30 in steps of 1) were tested by five-fold cross-validation. First, the SwissProt data is divided into five equal-sized partitions. Then five iterations of training and validation are performed in which at each iteration a different partition is held out for validation and the remaining four are used for the learning and combinatorial rule generation steps. In the validation step, the rules are filtered using the trial threshold values, and the retained rules are applied to the test set. Finally, the predicted annotations are compared to the actual SwissProt annotations.

The different possible hierarchical levels of function annotation in our predictions and in existing annotations are taken into account according to [Radivojac et al., 2013]. For example, if a SwissProt sequence is annotated with an EC number of 1.2.3.4, then the "parent" EC numbers, 1.2.3.-, 1.2.-.-, and 1.-.-.- are also treated as annotations when comparing predicted and known annotations by counting the numbers of matching ("true positive"), non-matching ("false positive"), and missed ("false negative") annotations. GO annotation terms in the GO hierarchy are treated in a similar way. The recall (ratio of predicted SwissProt annotations to existing SwissProt annotations), precision (ratio of predicted SwissProt annotations to all all predicted annotations), and F-measure (harmonic mean of recall and precision) are then calculated. For each set of trial threshold values, the above procedure is repeated with five different SwissProt partitions, and the global result is calculated as the average over the five rounds. This overall procedure is applied separately to each taxonomy kingdom. Because the number of Bacteria entries is much larger than the sum of the other three, the global F-measure depends more strongly on Bacteria than on the other three taxa.

Figure 4.2 shows the recall, precision, and F-measure as a function of SupportCount for each taxonomy kingdom. The numbers of sequences annotated are 19,442, 16,716, 332,976, and 185,107 for Archaea, Viruses, Bacteria, and Eukaryota, respectively. Increasing the SupportCount threshold slightly increases the precision but dramatically decreases the recall and hence reduces the F-measure. These results demonstrate that with Confidence ≥ 0.95, even low support rules often predict correctly with respect to the available SwissProt annotations. Nonetheless, prediction precision plays an important role in the selection of the SupportCount parameter. In all four taxonomy kingdoms, increasing the SupportCount

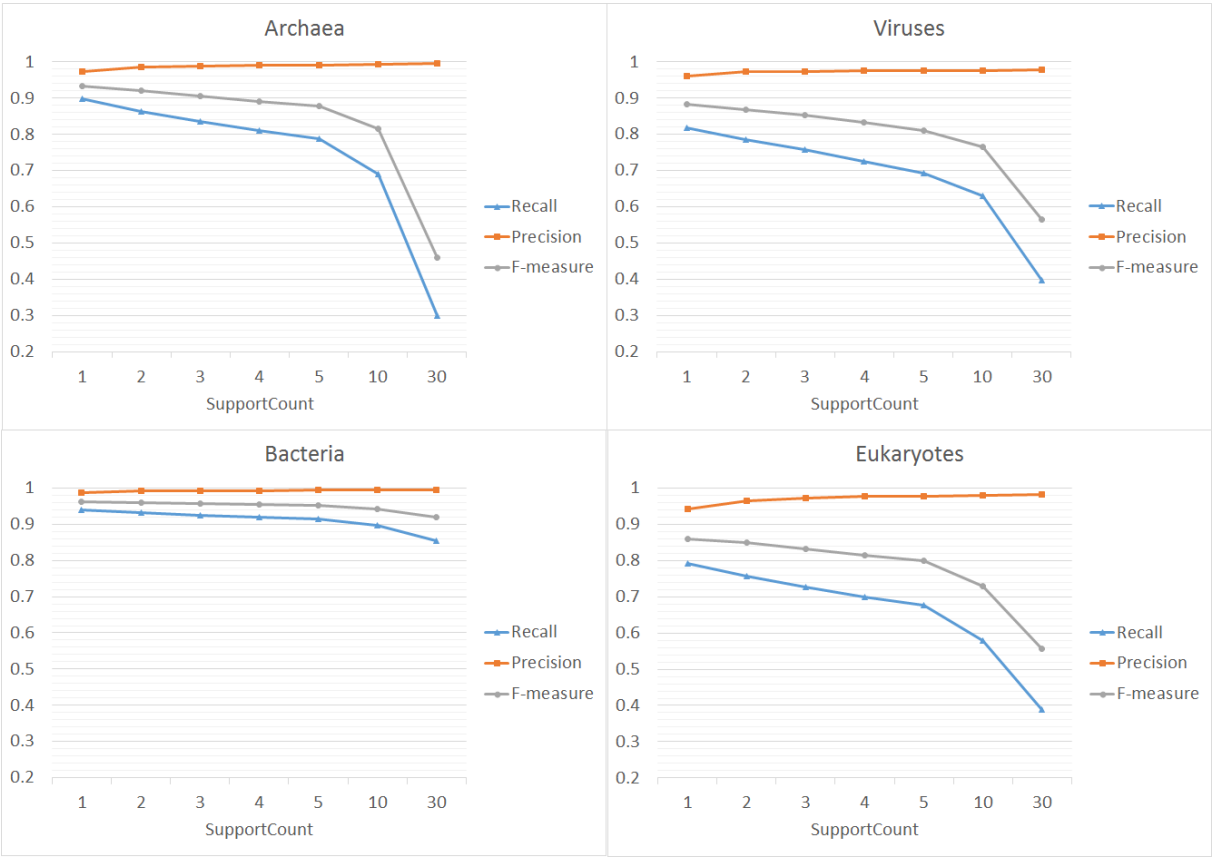


Figure 4.2: Recall, precision, and F-measure curves as a function of SupportCount for annotation rules having Confidence ≥ 0.95 and Lift > 1 in the four taxonomy kingdoms studied.

threshold from 1 to 2 has a greater effect on precision than increasing the SupportCount threshold from 2 to 10. This led us to choose 2 as a good value for annotating TrEMBL entries.

4.2.5 Aggregating and Applying Function Annotation Models

In the final stage of the modeling step, the surviving association rules for a given EC number are aggregated into one "model" for that EC number. (see Algorithm 4 for details). Equation 4.4 shows an example of a model that aggregates the antecedents of several filtered association rules having the same consequent, E_k. In this example, the five antecedents with different combinations of domains and taxa are represented as alternative cases to be matched against the target entry, p. If at least one such case matches p then E_k is assigned to p. Pseudo-code for this procedure is shown in Algorithm 5.

$$M_i : \left\{ \begin{array}{ll} Case_1 : & \{\{d_1\}, T_1\} \\ Case_2 : & \{\{d_2\}, T_1\} \\ Case_3 : & \{\{d_2, d_3\}, T_2\} \\ Case_4 : & \{\{d_4, d_5, d_6\}, T_3\} \\ Case_5 : & \{\{d_7\}, T_3\} \end{array} \right\} \Rightarrow EC_k \tag{4.4}$$

Algorithm 5 CARDM Annotation Algorithm

```
1: procedure Annotation(AnnotationModels, TargetProteins)
2:     for each AnnotModel ∈ AnnotationModels do
3:         {Case} ← Getcases(AnnotModel)
4:         Func ← GetAnnotation(AnnotModel)
5:     end for
6:     for each p ∈ TargetProteins do
7:         {Taxon_p} ← GetTaxon(p)
8:         {Domain_p} ← GetDom(p)
9:         {DomSubset_p} ← Subset({Domain_p}, Size ≤ 3)
10:        for each Case do
11:            if {DomSubset_p, Taxon_p} ∈ Case then
12:                Assign(p, Func)
13:            end if
14:        end for
15:    end for
16:    return(AnnotatedTargetProteins)
17: end procedure
```

4.2.6 Extension to Other Protein Annotations

CARDM was also used to build annotation models involving the GO MF, BP and CC namespaces and the same nine domain classifications. In this case, the SIFTS and SwissProt annotations were split into distinct datasets according to the GO "IEA" (Inferred from Electronic Annnotation) evidence code, leading to four data sources for CODAC learning. This allowed lower weight to be given to the IEA annotations compared to experimentally determined annotations when calculating CODAC's GO-domain association scores. The whole procedure of the method is drawn in Figure 4.3.

4.2.7 Data Preprocessing

Flat data files of SIFTS and SwissProt (July 2017) were downloaded and parsed using in-house Python scripts. Associations between PDB chains and EC numbers, and associations between PDB chains and domains from the nine domain classifications were extracted from the SIFTS data. All CATH and SCOP domain families were transformed into their corresponding superfamilies. Pfam "repeat" and "motif" domain types were discarded. All existing associations between SwissProt sequence accession numbers (ANs) and associations between ANs and EC numbers were collected for each of the nine domain classifications. Target protein entries were parsed to extract their taxonomic lineage information and domain lists from the nine selected domain classifications. The Gold Standard reference set of EC-domain associations required for the learning step was extracted from InterPro.

To avoid bias due to the presence of identical sequences in the data sources, PDB chains and SwissProt sequences were clustered using a sequence similarity threshold of 100% into "Clusters of Identical Sequences" (CIDs) using the Uniref non-redundant cluster annotations [Suzek et al., 2007]. The associations extracted from SIFTS and SwissProt were then converted into domain-CID and function-CID associations.

Function description often involves a hierarchical vocabulary or coding system. This is the case for EC numbers which obey to a four digit hierarchical numbering scheme. In order to exploit this hierarchical information, each extracted function-CID association was expanded to include associations involving the

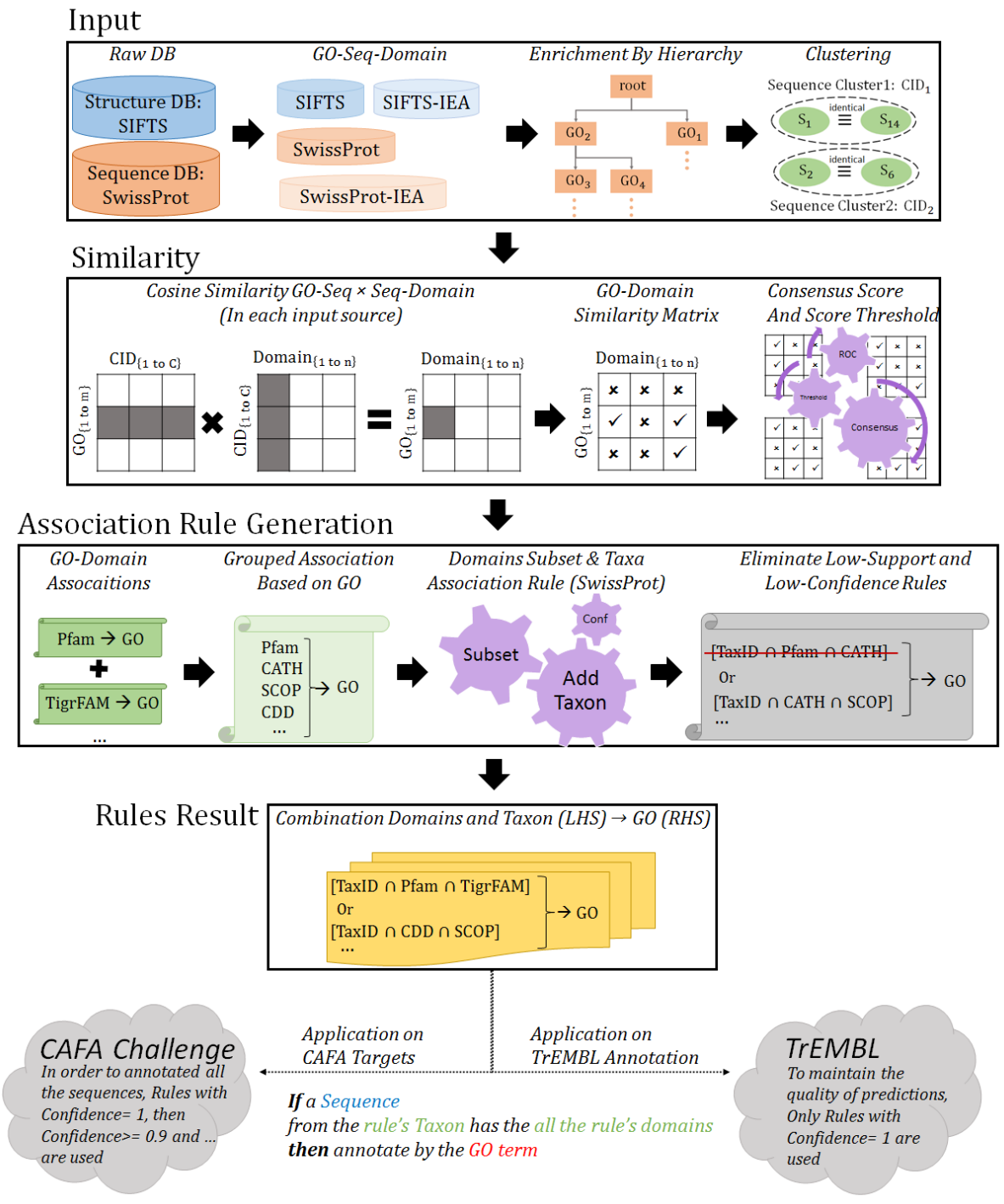


Figure 4.3: CARDM flowchart for generation of functional annotation rules using GO-domain associations.

parent levels of the annotation hierarchy. For an EC number of the form "1.2.3.4", this essentially means inserting associations for "1.2.3.-" and "1.2.-.-".

The SwissProt database was parsed again in the modeling step in order to calculate the Support, Confidence, and Lift of the generated associations rules. Each SwissProt entry was represented by a list of its assigned domains from the nine domain classifications, its taxonomic lineage, and its EC annotation(s). Each taxonomic lineage was split into parts from top to bottom of the hierarchy and assigned to the corresponding entry. For example, a protein sequence annotated with "Thaumarchaeota" is assigned to both "Archaea;Thaumarchaeota" and "Thaumarchaeota." In a similar manner, domain(s), taxonomic lineage, and any EC annotations were extracted for each TrEMBL entry. The annotation models were prepared in JSON and XML formats in order to be readable by other programs.

4.3 Results and Discussion

4.3.1 CARDM Generation of EC Annotation Models

In this work we used two sets of target protein entries: the TrEMBL database and the datasets of the 2013 CAFA challenge. CARDM was applied to EC annotation based on nine domain classifications (Pfam, CATH, SCOP, TIGRFAM, SMART, Panther, PRINTS, CDD, PROSITE). This required nine runs of the CODAC learning step giving nine predicted EC-domain datasets. The CARDM modeling step filtered and merged these associations using SupportCount threshold learnt from SwissProt along with a Confidence threshold of 0.95 and a Lift threshold of 1.0. Using a fixed Confidence threshold of 0.95 is justified by the need for consistency between this study and other annotation systems in TrEMBL.

Table 4.3 shows how the number of filtered EC association rules and the number of distinct EC annotation depends on the SupportCount threshold. The number of available association rules decreases with increase in the SupportCount threshold from 1 to 30. This table also shows the number of taxa and domain subsets involved in these models. A SupportCount threshold of 2 appears to give a good compromise between ensuring good coverage of EC annotation models and avoiding too many false positive annotation inferred from weak rules with support equal to 1.

4.3.2 Annotating TrEMBL Entries

The main purpose of CARDM is to annotate all of the protein entries in TrEMBL. As of July 2017, TrEMBL contains 63,684,389 protein sequences with at least one domain from the nine domain classifications used here. Table 4.4 shows more details about the number of entries in TrEMBL across the four taxonomy kingdoms considered here. This table shows that the number of entries for Bacteria is almost three times the number of entries for Eukaryota. However, the number of distinct taxa in Eukaryota is $\approx$ 22 times more than in Bacteria. The number of Eukaryota domains in TrEMBL is greater than the total number of Bacteria domains, even though the number of Bacteria entries is more than twice the number of Eukaryota entries. On the other hand, the number of distinct Virus domains is less than for the other three kingdoms, which might indicate a lower diversity of virus proteins in TrEMBL.

Table 4.4 also summarises the results obtained after applying the CARDM annotation models (produced using SupportCount ≥ 2, Confidence ≥ 0.95 and Lift > 1) to the protein entries in TrEMBL. The results show that about one-third of the entries in each kingdom can be annotated by CARDM. In total, CARDM annotates over 22.5 million entries using about 2,500 EC numbers and generates more than 24 million annotations. Thus, compared to the 10.9 million entries having at least one EC annotation in

Condition	Association Rules	EC annotation models	Taxa	Domain Subsets
Any confidence value				
$SupportCount \geq 1$	188,434,110	4,810	5,943	839,698
$SupportCount \geq 2$	93,393,615	3,616	3,402	678,467
$SupportCount \geq 5$	36,283,706	2,463	1,809	429,501
$SupportCount \geq 10$	16,971,029	1,822	1,143	280,553
$SupportCount \geq 30$	5,025,448	1,081	534	144,205
confidence ≥ 0.95				
$SupportCount \geq 1$	163,579,783	3,733	5,935	823,107
$SupportCount \geq 2$	77,895,271	2,703	3,372	649,336
$SupportCount \geq 5$	29,225,841	1,855	1,728	399,789
$SupportCount \geq 10$	13,451,737	1,405	1,021	251,942
$SupportCount \geq 30$	4,039,989	930	461	130,123
confidence $= 1.00$				
$SupportCount \geq 1$	163,079,769	3,733	5,935	822,915
$SupportCount \geq 2$	77,395,257	2,703	3,372	649,057
$SupportCount \geq 5$	28,725,827	1,854	1,726	399,326
$SupportCount \geq 10$	12,951,723	1,396	1,012	249,786
$SupportCount \geq 30$	3,666,611	902	435	122,827

Table 4.3: Numbers of association rules and annotation models produced with Confidence score ≥ 0.95 and Lift ≥ 1 and various thresholds for SupportCount.

	UniProtKB/TrEMBL			EC Prediction Results		
Kingdom	Entries	Distinct Domains	Distinct Taxa	EC numbers	Entries	Predictions
Archaea	1,152,973	13,495	346	520	312,045	317,832
Viruses	2,504,372	7,482	1,163	122	732,838	1,673,756
Bacteria	43,155,424	23,683	3,689	1,602	15,941,696	16,582,128
Eukaryota	16,871,620	29,601	82,640	1,610	5,573,911	5,789,926
Total	63,684,389	36,304	87,838	2,564	22,560,488	24,363,642

Table 4.4: **Left**: the numbers of TrEMBL protein entries having at least one domain in Pfam, CATH, SCOP, TIGRFAM, SMART, Panther, PRINTS, CDD or PROSITE along with the corresponding numbers of distinct domains and taxa in the four taxonomy kingdoms (first level of taxonomic lineage). **Right**: EC annotation predictions by CARDM for the TrEMBL entries.

TrEMBL (grey in Figure 4.1), CARDM can provide more than a 2-fold increase in the number of EC annotations. More precisely, our set of EC predictions concerns 12.8 million TrEMBL entries having no EC annotation and 9.7 million entries that had been previously annotated by other automatic systems. There remain 1.2 million previously annotated TrEMBL entries that are not assigned any EC numbers by our prediction models. This likely reflects differences between the CARDM algorithm and those used previously to annotate TrEMBL. Table 4.4 also shows how many distinct EC numbers have been assigned by CARDM to the TrEMBL entries.

It should be noted that CARDM can annotate a protein entry with more than one EC number if the criteria for more than one annotation model are met. Overall, CARDM annotates over one million TrEMBL entries with multiple EC numbers.

4.3.3 Comparison with Existing Annotation Systems in TrEMBL

Here, we compare our EC prediction with existing automatic and semi-automatic annotation systems in TrEMBL such as Rule-base [Morgat et al., 2011], SAAS [Morgat et al., 2011], and HAMAP-Rule [Pedruzzi et al., 2013]. HAMAP-Rule and Rule-base are semi-automatic systems in which bio-curators create annotation rules, but the rule application is automatic. SAAS is a completely automatic annotation system which generates annotation rules using decision trees. It is worth mentioning that both SAAS and Rule-base refine their predictions using automatic and semi-automatic annotation rules respectively, whose confidence score is greater than 0.95.

Figure 4.4 shows some statistics of the $\approx$ 24 million TrEMBL annotations produced by CARDM, compared to existing annotations in the four kingdoms considered in this study. The upper row of this figure represents all CARDM predictions and displays in green new predicted annotations concerning TrEMBL entries that were not previously annotated. These new annotations represent over 50% (reaching around 89% for viruses) of the total predictions. The lower row of this figure compares the results obtained by CARDM for those TrEMBL entries that already have annotations. In these pie-charts, the light blue sectors correspond to the number of identical predictions between CARDM and the existing annotations. In all kingdoms, over 80% of the CARDM predictions (from 82% in Bacteria to 98% in Archae) are in exact agreement with existing annotations. Grey sectors show the proportion of existing annotations that are similar to but more specific than the CARDM predictions (from 0.1% in Viruses to 10.7% in Bacteria) with respect to EC number hierarchy, while red sectors show the proportion that are similar to but less specific than the CARDM predictions (from 0.04% in Viruses to 1.7% in Eukaryota).

Dark blue sectors (from 0.04% in Archaea to 8% in Viruses) correspond to multiple predictions for the same TrEMBL entry for which CARDM not only agrees with existing annotations but also adds additional predictions. Finally, yellow sectors show the proportion of mismatches between CARDM and existing annotations (from 0.14% in Viruses to 4.4% in Bacteria). The very low percentages found here confirm the precision of the CARDM predictions that was indicated in the cross-validation stage. Although CARDM produces many more annotations than exist in TrEMBL and hence the overlap between the CARDM and existing annotations is relatively small, the above analysis strongly indicates that CARDM produces annotations which are highly consistent with those of the annotation systems currently used in TrEMBL.

Overall, from a total of 11,358,629 existing TrEMBL annotations, the CARDM predictions include 8,547,345 (75.3%) identical and 1,078,740 (9.5%) similar EC annotations, where an EC annotation is considered to be similar if it matches on all digits present (i.e. if there are no mismatched digits, excluding hyphens). Only 14% of the existing TrEMBL annotations are missed by the CARDM prediction models.

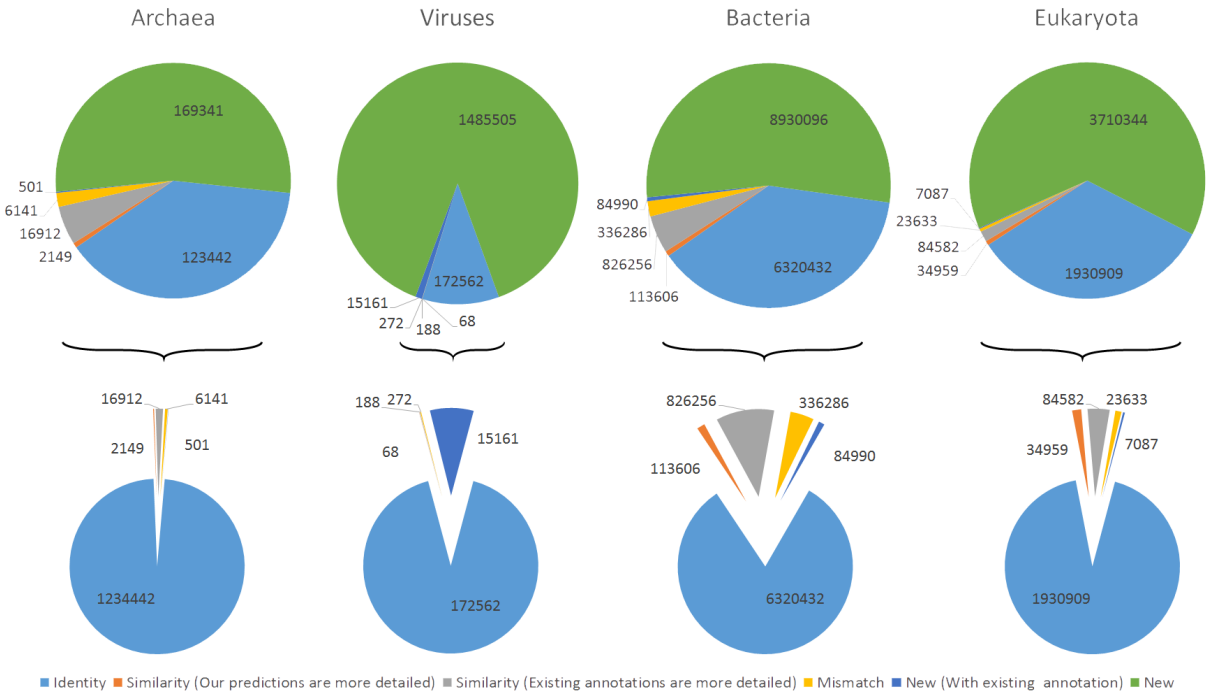


Figure 4.4: Comparison between CARDM predictions and existing annotation systems (Rule-base, SAAS and HAMAP-Rule) for EC annotation of protein entries in TrEMBL. Upper row: complete sets of CARDM predictions for each kingdom. Green: new predictions for TrEMBL entries lacking any annotation. Lower row: comparison of CARDM performance on TrEMBL entries having existing annotations. Light-blue: entries for which CARDM predictions are identical to existing annotations; red: more specific EC number in CARDM prediction; grey: more specific EC number in existing prediction; yellow: mismatch betwen CARDM and existing annotation; dark blue: entries having multiple annotations in the which existing annotation has been confirmed and a new prediction is proposed by CARDM. The actual prediction counts are indicated for each sector.

4.3.4 CARDM Annotation with GO Terms

CARDM builds prediction models for GO terms using largely the same procedure as described for EC annotations. However, in order to give different weights to manually curated GO terms and those inferred automatically, EC-AN associations were split into two groups according to the Inferred from Electronic Annotation (IEA) attribute. These two datasets are subsequently called SwissProt and SwissProt-IEA. The same separation into SIFTS and SIFTS-IEA was performed for SIFTS GO-AN annotations. Hence the consensus score obtained by the CODAC procedure for each GO-domain association is based on a weighted average of the similarity scores obtained in these four datasets. Because GO annotations stem from three namespaces (MF, BP and CC), and because we consider here nine established domain classifications, the CODAC learning procedure was applied separately to each combination of data sources (3×9 times), and separate sets of annotation models were built for each GO namespace.

The GO annotation results for TrEMBL are shown in Table 4.5. When considering all four kingdoms together, the percentages in this table show that the CARDM predictions are distributed rather evenly across the three namespaces, with the highest percentage (42.5%) being for MF annotations. This

Kingdom	GO terms	Entries	Predictions	Ratio
MF GO terms				
Archaea	690	563,539	1,090,101	1.93
Viruses	208	1,830,609	4,829,358	2.63
Bacteria	1,910	28,583,908	51,295,683	1.79
Eukaryota	3,480	11,621,820	22,458,112	1.93
Total	4,278	42,599,876	79,673,254	1.87
(percent)	(38.3%)	(66.7%)	(42.5%)	
BP GO terms				
Archaea	452	409,213	602,182	1.47
Viruses	268	2,081,711	5,518,326	2.65
Bacteria	1,232	21,636,079	34,150,791	1.59
Eukaryota	9,297	9,855,319	19,603,113	1.99
Total	9,740	33,982,322	59,874,412	1.76
(percent)	(87.3%)	(53.4%)	(31.9%)	
CC GO terms				
Archaea	61	241,570	318,201	1.31
Viruses	77	1,709,221	6,003,356	3.5
Bacteria	172	16,069,662	22,031,933	1.37
Eukaryota	1,587	11,591,981	19,555,286	1.69
Total	1,721	29,612,434	47,908,776	1.62
(percent)	(41.5%)	(46.5%)	(25.6%)	

Table 4.5: MF, BP, and CC GO predictions for the TrEMBL entries with annotation rules having Lift > 1, SupportCount ≥ 2 and Confidence ≥ 0.95. "Ratio" is the number of predictions per entity. The percentages in the three columns from left to right are relative to total number of GO terms in the corresponding namespace, total number of target TrEMBL entries and total number of CARDM predictions, respectively.

namespace corresponds to the largest percentage of target TrEMBL sequences (66.7%) but suprisingly to the lowest percentage of GO terms involved (38.3%). The larger involvement of BP terms (87.3%) likely reflects the diversity of BP-domain associations found in the CODAC learning step. Indeed our previous study using GODomainMiner inferred more BP-Pfam associations than MF-associations (75 versus 33 thousand) and these associations involved more BP terms than MF terms (14 versus 6.3 thousand) [Alborzi et al., 2018]. Furthermore, it is easy to see that the major contribution to the number of GO terms involved in each namespace comes from the Eukaryotae. This is consistent with the fact the Gene Ontology was originally developed to annotate eukarotic sequences.

The "Ratio" column of Table 4.5 shows that the numbers of predictions per entry are broadly similar for the three namespaces (from 1.62 to 1.87). However, some variation is observed depending on the kingdom, with a significantly higher prediction rate for Viruses. The high prediction ratios observed in viruses for the three GO namespaces are also associated with a high proportion of Virus entries concerned by these predictions (total number of viruses entries is ≈2,500, see Table 4.4) and with a quite small repertoire of GO terms (from 0.9 to 1.9%, depending on the GO namespace). The fact that function predictions are both more numerous and less diverse for Viruses than for the other kingdoms could deserve further investigation.

	GO terms			Protein Sequences			Predictions		
GO Category	Archaea	Bacteria	Eukaryota	Archaea	Bacteria	Eukaryota	Archaea	Bacteria	Eukaryota
MF	504	1,253	2,632	2,015	8,411	36,258	3,729	14,846	59,286
BP	283	687	6,870	1,484	6,921	26,296	2,029	10,155	75,296
CC	43	85	1,275	882	5,997	34,428	1,085	7,375	67,405
Total	830	2,025	10,777	2,195	11,116	52,040	6,843	32,376	201,987
	Total = 11,623			Total = 65,351			Total= 241,206		

Table 4.6: Summary of the CARDM results for the CAFA 2013 data. Shown are the numbers of assigned GO terms, protein entries, and predictions in the CARDM results for CAFA 2013.

4.3.5 CAFA Results

In order to compare CARDM with other state of the art approaches or those still under development, we applied CARDM to the CAFA 2013 data [Radivojac et al., 2013]. In that round of CAFA, the organisers provided 100,816 target proteins (Bacteria: 15,451, Eucaryotes: 82,074, and Archaea: 3,291 targets), of which predicted annotations for 3,675 proteins were assessed according to recently obtained experimental annotations.

Using the same parameters as described above, CARDM assigned 11,623 GO terms to 65,351 protein targets with a total of 241,206 predictions, i.e. on average 3.7 GO predictions per annotated sequence. Table 4.6 shows the number of protein sequences which are functionally annotated for the MF, BP and CC namespaces and the three taxonomic kingdoms in the CAFA dataset (Archaea, Bacteria, and Eukaryota). According to these results, we calculate that CARDM was able to successfully annotate 65% (65,351) of the CAFA targets. A more detailed analysis of our results showed that out of the 35,465 missed targets only 8,838 targets (8.8%) were not matched by any of our GO annotation models. This suggests that the difference (26,627 targets $\approx$ 26.6%) could have been annotated if the CARDM models had been built with rules filtered at lower SupportCount and Confidence thresholds. Nonetheless, we do not lower the optimal SupportCount or Confidence thresholds because of the risk of producing false positives.

After the CAFA 2013 experiment, the predictors' annotation methods were evaluated using recall-precision curves obtained by varying one parameter of the method and using ground-truth GO-Sequence associations provided by the CAFA organisers (3,675 proteins) [Radivojac et al., 2013]. In order to evaluate CARDM in a similar manner, we varied the Confidence threshold (while keeping the SupportCount threshold fixed at 2) in order to calculate different precision and recall values for the CAFA ground-truth annotations. The curves obtained with MF, BP and CC GO predictions are shown in Figure 4.5. In this Figure, each point (from the right to the left) is obtained by increasing the CARDM annotation rule Confidence threshold 0.0 to 1.0 in steps of 0.1 to make the precision increase and recall decrease. It can be seen that using Confidence threshold values in the range 0.8 to 1.0 yield the best precision values of around 50% and recall values of around 33% for the BP predictions, 40% for CC, and 45%for MF. It is worth noting that these values (and the overall shape of the recall-precision curves) are very similar to those reported for the best prediction methods in the CAFA 2013 assessment. Hence, we believe that the performance of CARDM is at least comparable to the state of the art approaches that participated in CAFA 2013 [Radivojac et al., 2013]. Furthermore, thanks to the generic formalization of our CARDM approach, we believe it could also be applied to other function classification schemes such as UniProt General Annotations.

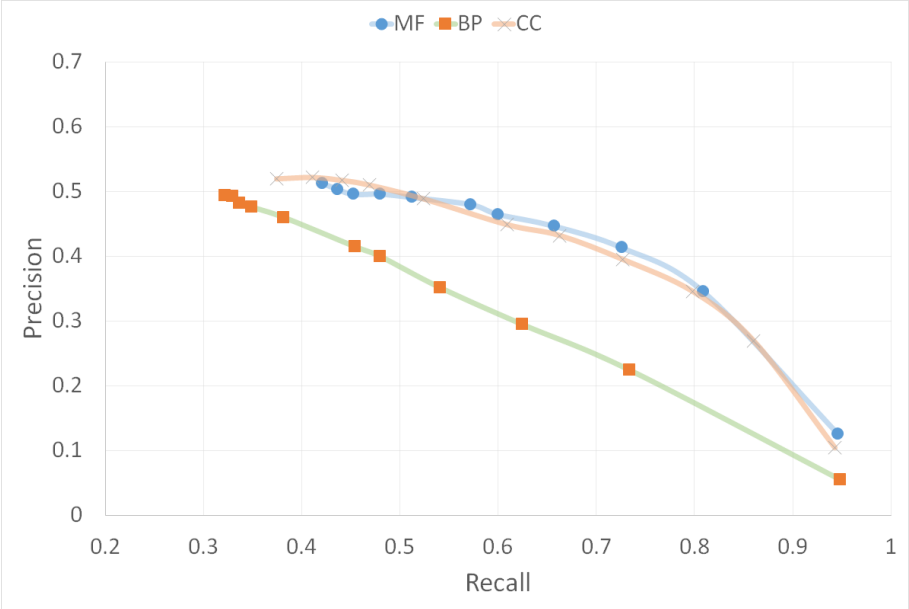


Figure 4.5: Recall-Precision curve for varying confidence threshold of the filtered GO annotation rules. Each circle represents (from the right to the left) a Confidence threshold (from right (0.0) to left (1.0) in steps of 0.1). Maximum F-measures of 53.2%, 43.7%, and 52.6% are obtained for MF, BP, and CC predictions, respectively.

4.4 Conclusion

We have described an automatic approach for functionally annotate protein sequences and structures with EC numbers and GO terms. This was achieved by first inferring function-domain associations using our previously developed CODAC method. A set of candidate association rules were then generated combinatorially using function-domain associations and taxa. The filtered list of association rules were then merged to build annotation models able to predict functions for TrEMBL sequences. CARDM found 24.3 and 187.5 million EC and GO predictions for 22.5 and 50.6 million target TrEMBL entires respectively. Over 60% of these predictions are new. CARDM was also used to annotate the protein sequences in the CAFA 2013 challenge. Our results indicate that the performance of CARDM is comparable to that achieved by the best predictor groups in that round of CAFA. Due to its generic nature, we expect CARDM could equally be applied to many other function prediction problems.

Chapter 5

Discovering Domain-Domain Interaction from Protein-Protein Interaction

Contents

Many biological processes are mediated by protein-protein interactions. However, the experimental determination of such interactions is often difficult and time-consuming. Hence there is much interest in developing computational approaches to predict protein interactions from knowledge of existing interactions. We describe an approach called "PPIDM" (Protein-Protein Interaction Domain Miner) for the computational discovery of protein-protein interactions using knowledge of their constituent domains. The approach is based on our previously described "CODAC" (Computational Discovery of Direct Associations using Common neighbors) method for the prediction of Pfam domain annotations. The approach has been applied to seven widely used protein-protein interactions resources, and it has been validated using a "Gold Standard" of three-dimensional domain-domain interactions extracted from the 3DID and KBDOCK databases. Overall, PPIDM finds a total of 27,363 non-redundant interactions between pairs of individual Pfam domains, and 523,929 interactions between sets of Pfam domains with a F-measure of 97% with respect to our Gold Standard dataset.

The result is publicly available at http://ppidm.loria.fr/.

5.1 Introduction

Many biological processes from metabolic pathways to cellular signaling are mediated by protein-protein interactions. However, the experimental determination of such interactions is often difficult and time-consuming. Furthermore, thanks to recent developments in high-throughput gene sequencing techniques, the gap between the number of known protein sequences and knowledge of their biological interactions is increasing rapidly. There is therefore a pressing need to develop computational approaches to help bridge this gap. There is therefore much interest in developing computational approaches to predict protein interactions from knowledge of existing interactions.

Computational methods for predicting interactions between pairs or groups of proteins often exploit knowledge of the co-evolution of protein pairs, and can be grouped into four main categories. 1) genomic context and structural information, 2) network topology, 3) text and literature mining (or database search), and 4) machine learning using various features from genomic or proteomic data. Gene co-localization is the simplest approach for predicting protein-protein interactions [Dandekar et al., 1998, Tamames et al., 1997]. The main idea is that related genes are located close together in the genome. This method is less appropriate for eukaryote genomes because related genes in eukaryotes are not necessarily co-located. More generally, phylogenic profile based approaches exploit the fact that functionally related genes often remain co-located in distant species. However, this approach is less well adapted when dealing with incomplete genomes or for proteins that are present in almost all organisms. Gene fusion events, in which several interacting proteins are fused into a single multi-functional gene, can be detected from comparative genomics and evolutionary information, and may also be used to infer functional relationships. However, it is less widely applicable.

Protein interaction networks in different organisms have similar topologies. These similarities may be exploited to distinguish predictions as true positives and false positives by assignment of a confidence value to each interaction [Goldberg and Roth, 2003]. Topological analysis of the protein-protein interaction networks is a significant task from the evolution viewpoint and network dynamics that shape the networks. In a given protein-protein interaction network, the properties are compared to the random networks and then confidence values are assigned to the protein interactions for determining the importance of the topological properties. Then, according to the confidence values interactions can be filtered or saved for the network.

Biomedical abstract are proliferating in NCBI PubMed database, with the rate of nearly one paper every thirty seconds [Zahiri et al., 2013]. Thus, protein interaction may also be predicted by text mining methods that exploit the co-occurrence of proteins mentioned in PubMed abstracts. Such literature mining approaches include natural language processing (NLP) approaches that use grammars and parsers to identify protein-protein interaction [Daraselia et al., 2004], rule-based approaches which infer protein-protein interaction from defined linguistic patterns [Huang et al., 2004], and machine learning approaches in which classifiers are trained to identify protein-protein interactions [Donaldson et al., 2003]. Several machine learning approaches have been used to predict protein-protein interaction, including support vector machines (SVM) [Guo et al., 2008, Zhang et al., 2014, Wei et al., 2016], artificial neural networks (ANNs) [Fariselli et al., 2002], naïve Bayes [Hsin Liu et al., 2012, Lin and Chen, 2013], k-nearest neighbors (k-NN) [Browne et al., 2007], decision tree (DT), and random forest (RF) decision [Chen and Liu, 2005, Wei et al., 2016] methods. SVM classifiers are widely used in classifying biological data, by maximizing the margins [Ben-Hur et al., 2008]. The margin for any object depends on the confidence of its classification. Objects for which the assigned labels are correct will have large margins and objects with uncertain classification are likely to have small margins. SVMs can be trained using a train-

ing dataset with certain labels which belong to one class. Then, a prediction model can be constructed to label new samples. SVM is very effective in classifying with arbitrary complexity. However, defining a problem for SVM is intricate and needs large memory. Moreover the selected parameters have strong effect on the results in this classifier [Ben-Hur et al., 2008].

ANNs have also been used to model protein-protein interactionis. One of the most popular ANN approachs is the multilayer perceptron (MLP) [Fariselli et al., 2002]. However, MLP is a black-box classifier because it is difficult to know what the model parameters mean [Yang, 2010].

Bayeseian probability based approaches are mainly applicable to problems having normal distributions, and can be trained efficiently with a small training dataset. However, they may fail in complex classification problems [Witten et al., 2016]. K-Nearest neighbors (K-NN) is one type of classifiers which assigns labels to each item based on the K nearest items in the feature space based on majority vote. K-NN requires no explicit training unlike statistical methods, and it is easy to implement. Nonetheless, memory and computation needs drastically increases if a large dataset or many features are used.

Finally, machine learning classification in protein-protein interactions discovery is Random forest (RF) algorithm. RF consists of many decision trees which are independently constructed according to random feature vectors sampled from a dataset. New items are assigned into one class according to the majority voting of decision trees. RF is useful for large numbers of features in large dataset and recovering missing data. However, it easily overfits databases containing noisy data [Witten et al., 2016].

Other computational methods for predicting domain-domain interactions use techniques such as correlated sequence signatures [Sprinzak and Margalit, 2001, Segura et al., 2015], maximum-likelihood estimation [Deng et al., 2002, Chen et al., 2012], phylogenetic profiling [Pagel et al., 2004, Cheng and Perocchi, 2015], statistical significance analysis [Nye et al., 2004, Bordner and Abagyan, 2005], analysis of domain pair exclusion [Riley et al., 2005], random decision forest [Chen and Liu, 2005, Liu et al., 2016], sequence coevolution [Jothi et al., 2006], parsimony-driven principle [Guimarães et al., 2006], formal concept analysis [Khor, 2014], and GO functional annotations [Lee et al., 2006]. It is worth mentioning that these automatic mining approaches may not produce results as credible as manually curated data, but the growth of manually curated data and combining different techniques and databases may make these methods more reliable.

Using three-dimensional (3D) structures is another way to predict protein interactions. This approach can be very reliable if structural interaction homologues exist, but in comparison to the enormous number of known protein sequences, this approach is potentially limited by the relatively small number of available 3D protein structures. On the other hand, since many proteins consist of well-defined domains, and since the number of different domain families is far smaller than the number of sequences to be considered, for data mining purposes it is natural to consider treating protein domains as fundamental units of function and interaction. However, while a small number of single domain proteins interact with their biological associates directly, a much larger number of proteins have more than one domain [Apic et al., 2001], and interactions between these multi-domain proteins can often involve two or more domains [Bhaskara and Srinivasan, 2011]. Therefore, to predict protein-protein interactions from the compositions of their constituent domains, it is first necessary to deconvolute the constituent domain-domain interactions (DDIs).

3DID [Stein et al., 2005, Stein et al., 2010], iPfam [Finn et al., 2013], INstruct [Meyer et al., 2013], and KBDOCK [Ghoorah et al., 2011, Ghoorah et al., 2013b], are examples of databases containing high quality structural information for experimentally determined DDIs, principally from interactions observed in crystal structures in the Protein Data Bank (PDB). Even though these databases provided thousands of DDIs, the number of inferred PPIs using these DDIs is currently far less than the number of PPIs

in sequence-based interaction databases. For example, it has been estimated that DDIs inferred from structural data in 2010 only cover around 5% of PPIs in Saccharomyces cerevisiae and 19% of PPIs in Homo sapiens [Yellaboina et al., 2010]. These observations encouraged us to develop a new method called "PPIDM" (for Protein-Protein Interaction Domain Miner) for the automatic prediction of DDIs between Pfam protein domains. PPIDM is derived from our previously described CODAC method [Alborzi et al., 2017c, Alborzi et al., 2017b, Alborzi et al., 2018] and is to our knowledge the first method that generates interactions between sets of protein domains.

5.2 Materials and Methods

5.2.1 Algorithm Overview

PPIDM is an extension of our previously developed CODAC method [Alborzi et al., 2018]. CODAC is a graph-based approach to predict new protein domain annotations from knowledge of existing associations between similar pairs of domains, whereas PPIDM treats each protein as a list of one or more domains, and aims to predict protein interactions from inferred relationships between lists of domains, or "itemsets".

Let $\mathcal{G}(X, Y, Z, E)$ be a tripartite graph where X, Y and Z are 3 sets of items and E is the set of all edges connecting X, Y and Z in the input configuration. In PPIDB*, we first assume that each item in Z can be a pair of elements ($Z = (z_l, z_r)$). We take into account that the edges between X and Z have different meaning from edges between Y and Z, namely, edge (x, z_l) means that the item in X belongs to the left element of Z, and edge (y, z_r) means that the item in Y belongs to the right element of Z. Thus, items in X and Y are connected to pairs of items in Z.

We next generates the subsets out of items in both X and Y such the subsets with $size = 1$ have at least one neighbor in Z, and subsets with $size \geq 2$ contain items connected to the same neighbor in Z. This allows us to create a tripartite graph where X, Y includes itemsets and Z items are pairs of elements, and E is the set of all edges connecting X, Y and Z in the input configuration. Figure 5.1 shows how the itemsets in X and Y are connected to a pair of elements in Z.

Let us consider 3 bipartite subgraphs of $\mathcal{G}$, denoted as $\mathcal{G}_l(X, Z, E_l)$, $\mathcal{G}_r(Y, Z, E_r)$, and $\mathcal{G}_g(X, Y, E_g)$. We now presume that the set of edges E_i is incomplete, and that the aim is to compute new edges between itemsets in X and itemsets in Y in order to generate $\mathcal{G}^*_g(X, Y, E^*_g)$ which together with $\mathcal{G}_l$ and $\mathcal{G}_r$ will make the final tripartite graph, $\mathcal{G}^*(X, Y, Z, E^*)$, where E^* denotes an enriched set of edges. New edges may be discovered by exploiting the existing edge distributions in $\mathcal{G}_l$ and $\mathcal{G}_r$. For example, if two itemsets x_i of X and y_j of Y share the same (or almost the same) pairs of elements, $\{(z_{l_k}, z_{r_k})\}$, in Z, then it may be supposed that an edge might exist between two itemsets x_i and y_j. A candidate edge between x_i and y_j is discovered if these itemsets are associated with the same pairs of elements in Z. Candidate edges found in this way are then scored and filtered, as described in more detail in [Alborzi et al., 2018].

It is now possible to instantiate our model with itemsets of Pfam domains (X) and (Y), and a set of protein-protein interactions (Z). E_l is the set of edges representing the Pfam domain content of the left-hand side protein sequence of each protein-protein interaction, E_r is the set of edges representing the Pfam domain content of the right-hand side protein sequence of each protein-protein interaction, and E_g is the set of edges representing observed Pfam-Pfam interactions from the intersection of KBDOCK and 3did dataases. In this case, our aim is to produce E^*_g, which will contain an enriched set of Pfam-Pfam interactions (also considered as associations) weighted by their neighborhood similarity score.

PPIDM infers domain-domain associations from seven existing protein-protein interaction databases. These associations are generated based on an assumption that each domain set is represented as one vector

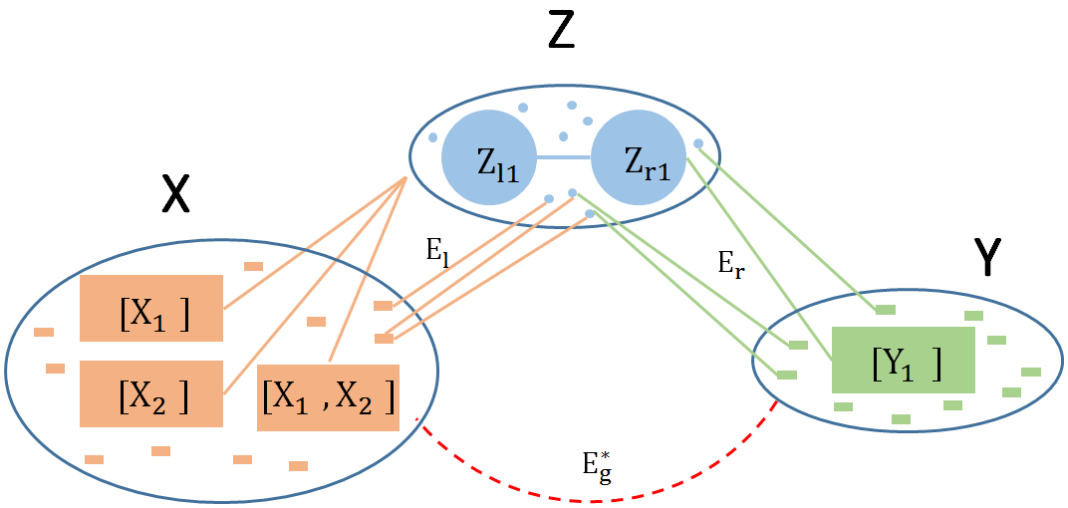


Figure 5.1: Schematic illustration of the extensions of the CODAC. Each item in Z is a pair of elements and itemsets in X and Y are connected to the neighbor items Z_l and Z_r in Z, respectively. CODAC* finds a set of weighted edges (E_g^*) between itemsets in X and Y using neighbors in Z.

with involving protein interactions as its features. Then, the cosine similarity between this domain set can be calculated as their scores for interaction. Assessment of the DDI is then performed by confirming which interaction is statistically significant and then comparing the result with observed domain-domain interactions to demonstrate the reliability of the process.

Protein interactions, which will be treated as the features, are extracted from IntAct, MINT, DIP, HPRD, BioGRID, String, and SIFTS databases. IntAct, MINT, DIP, HPRD, and BioGRID are manually curated physical interaction databases between proteins, while the very extensive STRING database contains both physical and predicted interactions between protein sequences. Protein interactions can also be inferred from PDB chains in the SIFTS database. Therefore, these seven interaction databases together provide a comprehensive combination of protein interactions and are appropriate for our learning procedure. Note that we retrieve all available interactions from these databases and we do not discriminate between stable and transient interactions. The number of protein-protein interactions obtained from the input resources are shown in Table 5.1. This table shows the large number of protein interactions drawn from the STRING database, while SIFTS database provides only a small collection of observed protein-protein interactions.

5.2.2 Input Data Collection

In this section, the CODAC* approach is applied to discover new weighted GO-domain associations. In our $\mathcal{G}(X, Y, Z, E)$ tripartite graph model, both sets X and Y in $\mathcal{G}_l$ and $\mathcal{G}_r$ correspond to Pfam domain classifications. 8 data sources were selected to provide common neighbors (Z) of the itemsets in X and Y, namely: (i) IntAct, (ii) DIP, (iii) MINT, (iv) HPRD, (v) BioGRID, (vi) SIFTS, (vii) STRING-exp, and (x) STRING-rest providing AN-AN associations (AN is the UniProtKB identifier) from IntAct database, DIP database, MINT database, HPRD database, BioGRID database, AN-AN associations inferred from SIFTS database, AN-AN associations with experimental tags from STRING, and AN-AN associations with non-experimental tags from STRING, respectively.

	Number of		
	Interactions	Protein Sequences	Associated Pfams
IntAct	411,624	76,747	9,898
MINT	67,191	24,735	6,438
DIP	53,585	20,489	6,418
HPRD	38,943	9,199	3,985
BioGRID	744,665	36,260	6,476
STRING	24,185,620	324,767	10,320
SIFTS	27,204	23,414	6,968

Table 5.1: Number of interactions, distinct sequences and Pfam domains obtained from the IntAct, MINT, DIP, HPRD, BioGRID, STRING, and SIFTS.

Flat data files of IntAct, DIP, MINT, HPRD, BioGRID, STRING, SIFTS, KBDOCK, 3did and UniProt (February 2017), were downloaded and parsed using in-house Python scripts. Associations between Uniprot sequence accession numbers (ANs) and Pfam domains were then extracted from the UniprotKB/SwissProt and UniprotKB/TrEMBL sections of Uniprot to give a dataset of AN-Pfam. Associations between every two interacting ANs were extracted from the IntAct, DIP, and MINT to give three datasets of protein-protein interactions. In BioGRID, interactions are listed between two interactor IDs. These IDs are associated to the gene names and species-level taxonomic identifier from NCBI. Interactions between ANs were generated using gene names and taxonomy IDs to give a dataset of BioGRID associations. STRING database provides a large list of associations between a pair of proteins using their own identifiers. Interactions between two ANs were extracted by using the mapping of the STRING IDs to the UniProt entries in UniProtKB. This mapping provides a large AN-AN associations database. We categorized the AN-AN associations according to experimental and non-experimental (Text mining, Neighborhood, Fusion-fission events, Occurrence, and Coexpression) labels and stored in STRING-ext and STRING-rest datasets, respectively. From the SIFTS data, associations between PDB chains were extracted and chain associations with high possibility of interaction are highlighted and stored using [Ghoorah et al., 2011]. Then, PDB chains that their representative AN exist, were replaced by the ANs and the AN-AN associations stored in the SIFTS dataset. These AN-AN associations are the pairs of elements, Z, where the left-hand AN is the Z_l and the right-hand AN is the Z_r

We extracted Pfam domains and AN-Pfam associations from the UniProtKB and created itemsets out of the Pfam domains. An itemset is kept in X if all the Pfam domains in the Pfam itemset belong to the Z_l. Similarly, an itemset is kept in Y if all the Pfam domains in the Pfam itemset belong to the right AN in $Z - r$. At the end, the resulting eight datasets are eight input tripartite graphs, Z for the Pfam itemsets in X and Y.

For the positive dataset, we extracted a total of 8,581 and 8,670 Pfam-Pfam interactions from 3did and KBDOCK, respectively. We then obtained 7,254 common Pfam-Pfam interactions between 3did and KBDOCK. These associations were considered to be the incomplete set of edges in $\mathcal{G}_g$ which is called "Gold Standard" and is going to be enriched. Note that the Gold Standard includes interactions with itemsets having only size 1.

5.2.3 Pfam-Pfam Interaction Inference

We use the CODAC algorithm to discover the associations between two sets of Pfam domains and present them as putative interactions between those sets of domains. In CODAC, to determine an edge similarity threshold, we prepare a "Gold Standard" dataset which is the combination of positive and negative examples of Pfam-Pfam associations. Here, we accept all of the associations in the intersection between 3did and KBDOCK as positive examples (E_g). To create negative examples, we use shuffling technique presented in the CODAC [Alborzi et al., 2018]. Next, we randomly split the Gold Standard dataset into two groups with equal numbers of positive and negative examples to give a "Training" and a "Test" subset.

To handle our eight datasets, each input tripartite graph is processed separately to calculate its respective cosine similarity matrix. The cosine similarity scores are then combined as a weighted average to give a consensus similarity matrix. Receiver-operator-characteristic (ROC) analysis provides an objective way to measure the ability of a classifier to distinguish positive and negative examples [Fawcett, 2006]. Therefore, each weight is varied from 0.01 to 1.0 in steps of 0.01, and for each combination of weights a ROC performance curve is calculated using the complete ranked list of consensus scores and our Gold Standard set of positive examples. The combination of weights that gives the largest area under the curve (AUC) is selected and used to calculate the best consensus similarity matrix.

We then rank the scores of all members of the Training subset, and label them "positive" or "negative" according to a score threshold that is varied from 0.0 to 1.0 in steps of 0.01. This allows us to find the true positive, false positive, true negative, and false negative predictions for each threshold. It should be noted that we consider 0.0 as a weight if the database does not have any impact on finding Gold Standard associations. We then calculate the recall, precision, and the F-measure. The similarity threshold T that gives the best F-measure with the Training subset is verified using the Test subset and retained to filter out edges whose similarity score is lower than T.

We systematically predict edges in $\mathcal{G}_g^*$, however, it is important to calculate a probability, or "p-value", for highlighting edges which are simply found by chance. We assume that the probability for finding an edge (x, y) by random chance is given by a hypergeometric distribution of the number of common neighbors (x, z) and (y, z) and apply the formula presented in CODAC [Alborzi et al., 2018]. We consider any p-value less than $0.05/|E_g^*|$ as denoting a statistically significant edge. We finally classify our Pfam-Pfam interactions into "Gold", "Silver", and "Bronze" using the p-values of interactions in different input databases.

5.3 Results and Discussion

5.3.1 Data Source Weights and Similarity Score Threshold

Our merged dataset contains 513,260 IntAct, 75,823 DIP, 97,487 MINT, 69,940 HPRD, 816,807 BioGRID, 4,131,112 STRING-EXP, 4,050,795 STRING-REST, and 30,709 SIFTS candidate Pfam-Pfam interactions with only one Pfam domain at both sides of each interaction, giving a total of 4,592,763 distinct Pfam-Pfam associations (Table 5.2). In our ROC-based training procedure, the best AUC value of 0.9944 was obtained with weights $w_{IntAct} = 0.05$, $w_{DIP} = 0.01$, $w_{MINT} = 0.01$, $w_{BioGRID} = 0.09$, $w_{STRING-Exp} = 0.12$, $w_{STRING-Rest} = 0.06$, $w_{HPRD} = 0.17$, and $w_{SIFTS} = 1.0$. These weights indeed give far greater importance to the candidate interactions in the SIFTS dataset, compared to those from other databases mainly because our positive instances are observed interactions extracted from PDB chains. It also indicates that Pfam-Pfam interactions from HPRD and STRING-EXP are more val-

	Name	Pfam-Pfam Interactions (Setwise)	Pfam-Pfam Interactions (Pairwise)	Pfam entries
Source Datasets	IntAct	2,085,450	513,260	9,898
	DIP	276,465	75,823	6,418
	MINT	385,885	97,487	6,438
	BioGRID	2,708,430	816,807	6,467
	STRING-Exp	17,368,745	4,131,112	10,320
	STRING-Rest	16,947,822	4,050,795	10,313
	HPRD	354,087	69,940	3,985
	SIFTS	1,734,362	60,114	7,449
	Merged	20,505,086	4,592,763	12,622
Reference	3did ∩ KBDOCK	7,254	7,254	5,260
	3did	8,581	8,581	5,545
	KBDOCK	8,670	8,670	5,882
PPIDM	Results	523,929	27,363	7,628
	(Common to Gold Standard)	(6,897)	(6,897)	(5,228)

Table 5.2: Statistics on the source datasets and calculated Pfam-Pfam Interactions.

ued that those from IntAct, DIP, MINT, BioGRID, and STRING-REST. Interestingly, all data sources weights are higher than 0, thus, all these sources have impact, even very low, on the discovering our Pfam-Pfam interactions.

The optimal score threshold was determined according to the F-measure calculated during our procedure using our training dataset. This gave a score threshold of 0.02 for a maximum F-Measure of 0.968. Applying this threshold to the test dataset yielded a comparable F-measure of 0.969, and precision and recall values of 0.98.7 and 0.95, respectively.

5.3.2 Analysis of Inferred Pfam-Pfam Interactions

The overall results of the PPIDM are summarized in Table 5.2. This table shows the number of interactions between sets of Pfam domains (Setwise interactions). It should be noted that the size of the itemsets in both sides of the interactions are limited to two. Table 5.2 also shows the numbers of Pfam-Pfam Interactions along with the numbers of distinct Pfam entries involved in those associations for the eight sources and the merged datasets before filtering. The number of interactions between sets of Pfams are more than three times of the Pfam-Pfam interactions with one Pfam domain at both sides of each interaction (Pairwise interactions). The pairwise interactions are included in the setwise interactions as an interactions between sets of Pfams with size 1.

After applying the 0.02 score threshold, the number of pairwise Pfam-Pfam interactions falls to nearly 0.6% of the merged dataset with an overlap of about 87.2%, 91.9%, and 95.1% of the 3did, KBDOCK, and Gold-Standard (3did ∩ KBDOCK) reference associations, respectively. The results also shows that our PPIDM filtered out around 97.5% of the setwise interactions in the merged dataset and predicted in total 523,929 interactions between two sets of Pfam domains.

Table 5.3 shows the distribution of PPIDM predicted interactions according to our Gold, Silver, and Bronze classification, along with the degree of overlap with the Gold-Standard reference dataset. This table shows that PPIDM provides 5,861 "Gold" domain-domain interactions (present in at least half of the source datasets and having significant p-values in all of the source datasets), 8,954 "Silver" domain-domain interactions (present in less than half of the source datasets and having significant p-values in all the source datasets), and 12,548 "Bronze" domain-domain interactions (having at least one insignificant p-value). It is interesting to see that the 42%, 31%, and 13% of our predicted Pfam-Pfam associations in

Class	PPIDM	Overlap with Gold-Standard	Single Domain PPI
Gold	5,861	2,454	1,960
Silver	8,954	3,322	2,211
Bronze	12,548	1,635	2,880
Total	27,363	6,897	7,051

Table 5.3: The distribution of all pairwise interactions from PPIDM, their overlap with our Gold-Standard, and involving single-domain protein-protein interactions, in the Gold, Silver, and Bronze categories.

the Gold, Silver and Bronze classes are common with the Gold-Standard, respectively.

Table 5.3 also represents the number of our predicted Pfam-Pfam interactions involving at least one protein-protein interaction with a pair of single-domain protein sequences. Thus, these unambiguous associations constitute the most reliable interactions calculated by PPIDM.

5.3.3 Comparison with DOMINE

In order to compare PPIDM with the DOMINE database, we extracted Pfam-Pfam Interaction from the file available from the latest version of the DOMINE database (http://domine.utdallas.edu/). DOMINE file includes 26,219 Pfam-Pfam interactions with 5,410 distinct Pfam domains. This set (shown as purple in Figure 5.2) was compared with the set of all 27,363 calculated Pfam-Pfam interactions found by PPIDM (blue in Figure 5.2). This comparison showed that a total of 7,346 Pfam-Pfam interactions from DOMINE are present in our calculated dataset including 4,779 interactions from the Gold-Standard (Intersection between yellow, blue, and purple in Figure 5.2). The remaining 18,873 DOMINE interactions were then compared with the interactions from the Gold-Standard. This comparison (the intersection of purple and yellow minus blue) showed a total of 155 Pfam-Pfam interactions are common to DOMINE and the Gold-Standard but not PPIDM, indicating that PPIDM misses only 3.1% (155 $\div$ 4,934) of the DOMINE interactions confirmed by observed Pfam-Pfam interactions. Moreover, this comparison also shows that PPIDM result set contains 2,118 $(6,897 - 4,779)$ additional Pfam-Pfam interactions that are in Gold-Standard but not available through DOMINE.

5.3.4 Comparison with INstruct

In order to compare PPIDM with the INstruct database, we extracted the Pfam-Pfam interactions from all the available Pfam interactions in the INstruct database (http://instruct.yulab.org/downloads.html). The INstruct database includes 2,685 interactions with 1,517 distinct Pfam domains (red in Figure 5.3). This set was compared with the set of all 27,363 calculated Pfam-Pfam interactions found by PPIDM. This comparison showed that a total of 1,739 Pfam-Pfam interactions from INstruct are present in our calculated dataset This comparison illustrated a total of 1,499 Pfam-Pfam interactions which are common to INstruct and the Gold-Standard, indicating that PPIDM shared 97.9% of the INstruct interactions existing in our Gold-Standard.

The remaining 946 INstruct interactions were then compared with the interactions from the Gold-Standard. It showed that only 31 of which are common to 7,267 interactions from the Gold-Standard.

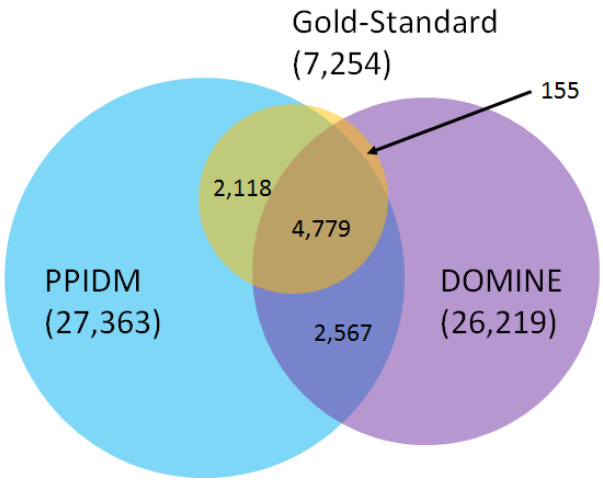


Figure 5.2: Venn diagram for overlapping domain-domain interactions between PPIDM (blue), DOMINE (purple), and our Gold-Standard (KBDOCK ∩ 3did, yellow). PPIDM and DOMINE share 7,346 interactions. The Gold-Standard has 6,897 and 4,934 domain-domain interactions in common to the PPIDM and DOMINE, respectively, while the Gold-Standard, PPIDM, and DOMINE share 4,779 interactions.

Moreover, this comparison also showed that PPIDM results contain 5,013 (6,752 - 1,739) additional Pfam-Pfam interactions that do not exist in INstruct but are available through the Gold-Standard.

5.3.5 Evaluation of PPIDM Predictions

It is very difficult to review individually 20,466 (27,363 - 6,897) Pfam-Pfam interactions predicted by PPIDM and not present in the Gold Standard (KBDOCK ∩ 3did). Thus, we first attempted to estimate our interactions potential value taking into account the KBDOCK and 3did interactions that are not common to both databases (and consequently not in the Gold Standard) are effectively predicted by PPIDM. The results are presented in Table 5.4 reveal that PPIDM finds 91.9% and 87.2% of the KBDOCK and 3did interactions, respectively. This contrasts with the lower percentage observed for the overlap with DOMINE and INSTRUCT databases (28% and 64.8%, respectively). Nonetheless, it did not escape our attention that 75.4% and 44% of the KBDOCK and 3did interactions which are not present in the Gold Standard are indeed predicted by PPIDM.

Moreover, 77% of the predicted Pfam-Pfam interactions overlapping with KBOCK and 3did are interestingly from the gold and silver categories (i.e. all p-values significant). This statistical overview is a good implication that PPIDM predictions likely contain high quality and relevant new DDIs.

We also analyzed a small subset of PPIDM interactions, namely the one-to-one Pfam-Pfam interactions in which the left and right Pfam domains are uniquely associated with only one domain. Such one-to-one interactions can simply be used to evaluate the biological consistency of the discovered knowledge through our method. We obtained a total of 1,606 one-to-one pairwise Pfam-Pfam interactions with consensus scores ranging from 0.2046 to 0.8104, 277 interactions in the gold category, 1,284 and 45 in the silver and bronze categories, respectively. From the 1,606 interactions, 1344 are already observed by our Gold-Standard (250, 1,086, and 8 interactions from the gold, silver, and bronze categories, respectively) and 262 are new (27 from the gold category). We additionally found 67 one-to-one interactions between Pfam itemsets with size of 2. with consensus scores ranging from 0.2311 to 0.6752.

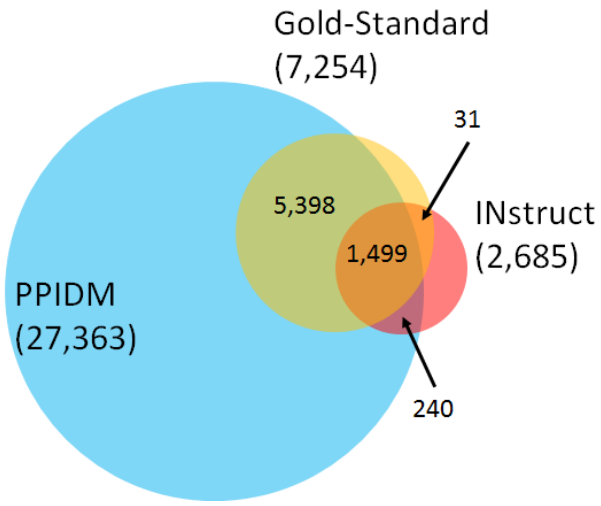


Figure 5.3: Venn diagram for overlapping domain-domain interactions between PPIDM (blue), INstruct (red), and our Gold-Standard (KBDOCK ∩ 3did, yellow). PPIDM and INstruct share 1,739 interactions. The Gold-Standard has 6,897 and 1,530 domain-domain interactions in common to the PPIDM and INstruct, respectively, while the Gold-Standard, PPIDM, and INstruct share 1,499 interactions.

		Overlap with				
Class	PPIDM	Gold Standard	KBDOCK	3did	DOMINE	INstruct
Gold	5,861	2,454	2,752	2,617	3,018	1,255
Silver	8,954	2,808	3,268	2,964	2,238	123
Bronze	12,548	1,635	1,945	1,900	2,090	361
Total	27,363	6,897	7,965	7,481	7,346	1,739
(Percentage of Database)		(95.1%)	(91.9%)	(87.2%)	(28%)	(64.8%)

Table 5.4: The number of overlapping PPIDM interactions with different interactions sources divided into Gold, Silver, and Bronze.

In summary, it appears that the PPIDM predictions constitute an interesting resource for looking at known or unknown DDIs and that in each studied case unknown DDIs reveal to reflect reasonable biological interactions.

5.4 Conclusion

We have presented PPIDM for mining protein-protein interactions at the domain level. This was accomplished by discovering single and subsets of Pfam domains while assuming that each connected neighbors consists of a pair of elements. Our method yields an enrichment of about 4-fold in the number of Pfam-Pfam interactions that currently exist in the intersection of two datasets of observed interactions. PPIDM achieved a F-measure of 0.97, and precision and recall values of 0.99, 0.95, respectively. We believe that the large numbers of inferred Pfam-Pfam interactions can be used to create a novel database of predicted protein-protein interactions as well as annotate UniProt sequences.

Chapter 6

Conclusions and Perspectives

Contents

6.1 Summary of the Main Contributions

In this thesis we contributed to the Knowledge Discovery in Bioinformatics, specifically we have targeted the domains of protein function annotation and protein interaction by developing novel methods and applications called CODAC (ECDomainMiner and GODomainMiner), CARDM, and PPIDM.

In detail, we have presented the development of a new approach to find direct associations between pairs of elements linked indirectly through various common neighbors (CODAC), and then using this approach to directly associate biological functions to protein domains, and to discover domain-domain interactions. Finally, we have extended this approach to generate functional prediction models and comprehensively annotate protein structures and sequences (CARDM).

Concerning the generic formal CODAC approach, it was designed as tripartite graph framework in which one set of sparse edges gets enriched into a new set of weighted edges through the mining of the two other sets of edges. This approach has been implemented as ECDomainMiner and GODomainMiner, for inferring associations between EC numbers and GO terms with protein domains, e.g. Pfam. Our method provides an overall enrichment of more than 13-fold in the number of direct associations between EC and Pfam or GO and Pfam associations that currently exist in the manually curated InterPro database. Our findings had overlap with nearly 99% and 93% of the EC-Pfam and GO-Pfam associations present in InterPro database. Based on our presented analysis, our method has higher coverage of associations in InterPro in comparison to dcGO. Furthermore, a selected subset of one-to-one associations has been analyzed and these all appear to be highly meaningful from a biological point of view and consistent with available knowledge. We believe that these high quality function-domain associations simplify our understanding and investigating of protein structure-function relationships at the domain level.

Nevertheless, our method also infers a large amount of new function-domain associations that cannot be validated in a simple manner. We are aware that our method cannot be considered as a learning method as we lack any independent function-domain dataset to test the prediction. We prefer to consider it as a score-based inference method which is reminiscent of information retrieval methods, especially for InterPro-derived associations which are internal positive controls used to calibrate the system (weight optimization and optimal threshold finding).

Concerning functional annotation of proteins, we decided to explore these large numbers of associations between protein functions and domains inferred by our approach, leading us to introduce a new approach called CARDM. This new systemic approach is designed and developed to functionally annotate protein sequences and structures in a completely automatic way. We thus applied our annotation rules for functional prediction of the sequences in the UniProtKB/TrEMBL and target sequences provided by CAFA 2013. The automatic functional annotation protein sequences was done in collaboration with UniProt team at European Bioinformatics Institute (EBI). According to the latest detailed analysis from the UniProt curators, our large amount of predicted annotations are very useful with low disagreement with their existing annotation systems, therefore it can be integrated in the UniProtKB/TrEMBL.

Finally, the CODAC approach was extended to deal with more general tripartite graphs including itemsets rather than single items, and sets of edges with different semantics. This was implemented as "PPIDM" to computationally discover interactions between single or subsets of Pfam protein domains. All automated methods to predicting domain-domain interactions return interactions between two single protein domains, however, our PPIDM is the first method that can predict interactions between both single and subsets of protein domains on each side of the interactions.

During the course of this thesis, two peer-reviewed journal articles: "ECDomainMiner: discovering hidden associations between enzyme commission numbers and Pfam domains" (BMC Bioinformatics 2017) and "Computational Discovery of Direct Associations between GO terms and Protein Domains" (BMC Bioinformatics 2018-Accepted), in addition to two peer-reviewed conference papers of "EC-PSI: Associating Enzyme Commission Numbers with Pfam Domains" (bioRxiv 2015) and "Associating Gene Ontology Terms with Pfam Protein Domains" (Lecture Notes in Computer Science 2017), have been published. One manuscript about "Combinatorial Association Rules for Protein Functional Annotation Using Inferred Function-Domain Associations" is in preparation with the UniProt team and one more manuscript about "PPIDM" is in the final stage of preparation.

The ECDomainMiner, GODomainMiner, CARDM, and PPIDM result databases are publicly available at http://ecdm.loria.fr/, http://godm.loria.fr/, http://cardm.loria.fr/, http://ppidm.loria.fr/, respectively.

6.2 Future Directions

6.2.1 Short-Term Perspectives

- **Algorithmic Improvements**: GODomainMiner and ECDomainMiner provide highly reliable associations between protein functions and domains. In the current version of the ECDomainMiner, UniProtKB input dataset is divided into three datasets based on the manual, automatic and UniRule annotation types. Similarly, GODomainMiner divides UniProt entires into four datasets according to SwissProt IEA and manual, and TrEMBL IEA and manual evidence codes. However, further separation of such annotations in the input datasets according to the annotation source might improve the results. UniRule annotations can break into Hamap-Rule, RuleBase and SAAS for

ECDomainMiner, and further separation of GODomainMiner input datasets based on the manual evidence codes is one short-time future direction for improving our two main inferred datasets.

The CODAC algorithm finds weighted associations between two items using a tripartite graph. Edges are binary in the current version of the CODAC. One extension to the CODAC is that the edges between X, Y, and Z could be weighted between zero and one based on the quality of the associations.

CARDM produces prediction models that contain annotation rules with less than four protein domains. Increasing the number of protein domains in the annotation rules will increase the execution time, however, it could improve the quality of the annotation rules and find more annotation rules with higher confidence. Moreover, the annotation rules can be defined as a combination of protein domains, taxonomic information, and other experimentally discovered functions as the left-hand side of the rule. Moreover, statistical analysis of generated rules (e.g. p-value) would increase the reliability of accepted rules.

- **More Biological Applications**: We proposed a method to computationally discover interactions between Pfam domains, however, we can design a system to create models for Protein-Protein Interaction using association rule mining techniques. Similar to CARDM, our method may consist of three main steps, namely the learning, modeling, and annotation steps in which the learning step consists of inferring and filtering domain-domain interactions using the CODAC approach and modeling step comprises the task of generating and filtering association rules involving domains and taxons and creating annotation models by rule aggregation. Last but not least, the annotation step includes using the selected prediction models to discover protein-protein interactions.

  Protein domain structure classification systems such as CATH and SCOP provide a useful way to describe evolutionary structure-function relationships. Similarly, the Pfam sequence-based classification identifies sequence-function relationships. Nonetheless, there is no complete direct mapping from one classification to another. This means that functional annotations that have been assigned to one classification cannot always be assigned to another. We can use our CODAC approach to systematically analyze multiple protein domain relationships in the SIFTS and UniProtKB databases in order to infer direct mappings between CATH superfamilies, Pfam clans or families, and SCOP superfamilies. Our preliminary results show that we provide 3-fold increase in the number of available CATH-SCOP mappings in the Genome3D whilst our result covers nearly 99% of the existing mappings.

  Another interesting application of the CODAC is to find the mappings between different categories of Gene Ontologies. Such a list of mappings between molecular function, biological process and cellular components helps locating the lack of integrity in the annotations of the UniProtKB/TrEMBL and also better understanding of the relations between ontologies terms.

  In addition, using CODAC and CARDM methods, existing general annotations in the UniProtKB can also be predicted for the protein sequences and structures that currently lack any annotations. Such predictions could be carried out by creating annotation rules using either direct associations between general comments and protein domains, or direct associations between general comments and GO terms.

  PPIDM provided a large number of domain interactions between Pfam domains. Pfam is one of the most widely spread domain classification over protein sequences and structures. However, interactions for other protein domain classifications such as CDD or TIGRFAMs could be interesting for

functional annotation of interacting proteins. Moreover, such interactions could assist in generating better prediction models for protein-protein interaction.

- **Execution Time Improvement**: Inferring associations between two items in X and in Y is processed separately from other items. Moreover, there are often a large number of combinations between items in X and items in Y in biological data. This indicates that from the execution point of view, the CODAC algorithm could be improved and highly parallelized in GPU using CUDA [8] Python (https://developer.nvidia.com/how-to-cuda-python) or CPU using symmetric multiprocessing libraries (https://wiki.python.org/moin/ParallelProcessing). Similarly, generating annotation rules in CARDM algorithm and their application on the large number of protein sequences can also be parallelized because the generation and application of each prediction model is separate from the other models.

6.2.2 Wider Perspectives

- **RNA-Protein and DNA-Protein Interactions**: RNA-protein and DNA-protein interactions play essential roles in many cellular processes. For instance, there are interactions between RNA and proteins within the ribosome. RNA-protein interactions mediate RNA metabolic processes such as poly-adenylation, splicing, stability of messenger RNA, translation, and localization [Tuschl, 2003]. Moreover, several RNA-binding proteins are involved in human diseases [Cooper et al., 2009]. DNA-protein interactions also have an impact on gene expression for example through recognition of DNA short sequences and transcription factors or other regulatory proteins. There are some computational methods developed to predict the interactions between proteins and DNA [Nagarajan et al., 2013] or RNA [Mann et al., 2017, Puton et al., 2012] using structure and sequence data.

  We believe that we can use the CODAC approach to find interactions between DNA and protein domains on one hand and RNA and protein domains on the other hand. Such findings could be used to create prediction models for DNA-protein and RNA-protein prediction using the CARDM approach.

- **Domains Architectures**: In the characterization of the protein with functions, Bashton and his colleagues claim that functions of an individual protein are not only due to the combination of the functions of its constituent domains, but also originate from a unique way that the building blocks of the protein are interactively contributing [Bashton and Chothia, 2007]. This leads us to the notion of domain architecture as an unique feature of proteins based on the arrangement of domains. These features consist of the domain content and the linear order of the domains in the protein sequences. Domain architecture concept has already been used by the SMART protein domain classification [Letunic et al., 2014] and recently been considered as an important feature in protein function prediction [Doğan et al., 2016]. In a larger scale, we believe such a concept of domain arrangements may upgrade our annotation prediction models produced by the CARDM system.

- **Function Similarity**: GODomainMiner finds direct associations between GO terms and protein domains. We used hierarchy of the Gene Ontology by adding the GO terms ancestors to improve the discovery of GO-domain associations. However, GO terms similarity can also be computed between

[8] Compute Unified Device Architecture (CUDA) is an application programming interface model for parallel computing created by Nvidia. It allows software application developers to use a CUDA-enabled GPU for general purpose processing. Today, hundreds of applications such as [Alborzi et al., 2014] are GPU-accelerated.

terms which are not in one branch of the hierarchy. Such similarities between two GO terms can be calculated using different methods such as IntelliGO [Benabderrahmane et al., 2010]. Using GO terms similarity in addition to the GO hierarchy could further improve our findings inferred by GODomainMiner.

- **Negative Taxonomic Information**: Feuermann and his colleagues infer function according to an approach called "GO Phylogenetic Annotation" [Feuermann et al., 2016]. This approach integrates GO annotations from genes across different organisms which are evolutionarily related. Then, they construct a model of the evolution of gene functions. Therefore, a function could be active in two proteins of two distant organisms but could be inactive in a closer related organisms. This lead us to the idea that due to evolution a function could exist for a given protein through all organisms of a taxonomy classification except a subbranch. Therefore, absence of a subbranch of taxonomic lineage could also be integrated in our prediction models. CARDM generates annotation rules and then prediction models based only on the presence of taxonomic lineage. In short, if an annotation rule has the same taxon as the target sequence, we go to the verification if the domains of annotation rule are a subset of domains in the target sequence. However, we could upgrade our prediction models during learning phase by including the absence of certain taxonomic information. In this situation, we could create smarter annotation rules that assign a function to a given protein through organisms belonging to the taxon branch excluding one or more subbranches.

- **Beyond Bioinformatics**: As a generic approach, CODAC can be applied on any tripartite setting in which we have indirect connections between two sets (please refer to section 3.1). One example is that we can use CODAC to predict the best applications for jobs. In this problem, language patterns of experiences and required skills are extracted from both resumes of applicants and the job descriptions. Thus, resumes, jobs, and patterns are the items in the X, Y, and Z of our approach. We next run CODAC to find the similarities between resumes and jobs based on associated patterns. Then, the best similarity shows the best applicant for the job. The schematic illustration of the selection adapted to our tripartite graph is shown in Figure 6.1. This problem could be solved simply by core CODAC algorithm, however, we could improve results by dividing the language patterns into time ranges (each time range is considered as one input source), clustering the similar patterns based on a thesaurus (Python programming and Python development are identical), and enriching skills hierarchically (a person who knows Python programming, knows programming in general).

  Other examples are suggesting vacation spots or weekend getaways to customers, finding possible foods with new ingredients which could be added to the menu of a chained restaurant and accepted by frequent customers, recommendation of items to the loyalty card holder customers, highlighting cosmetic materials usable for a certain group of customers, and many other classical recommendation problems.

6.2.3 Further Verification of Inferred Functions

CODAC and CARDM provide a large set of function-domain and function-sequences/structure associations which are above certain thresholds. Such thresholds reduce the number of false-positive predictions remarkably. However, further confirmation of the prediction would add more credibility to our findings and allow us to design knowledge-base filters. There are different ways to verify the findings such as wet-lab experiments. At the time of writing this thesis, we are collaborating with a microbiology lab in

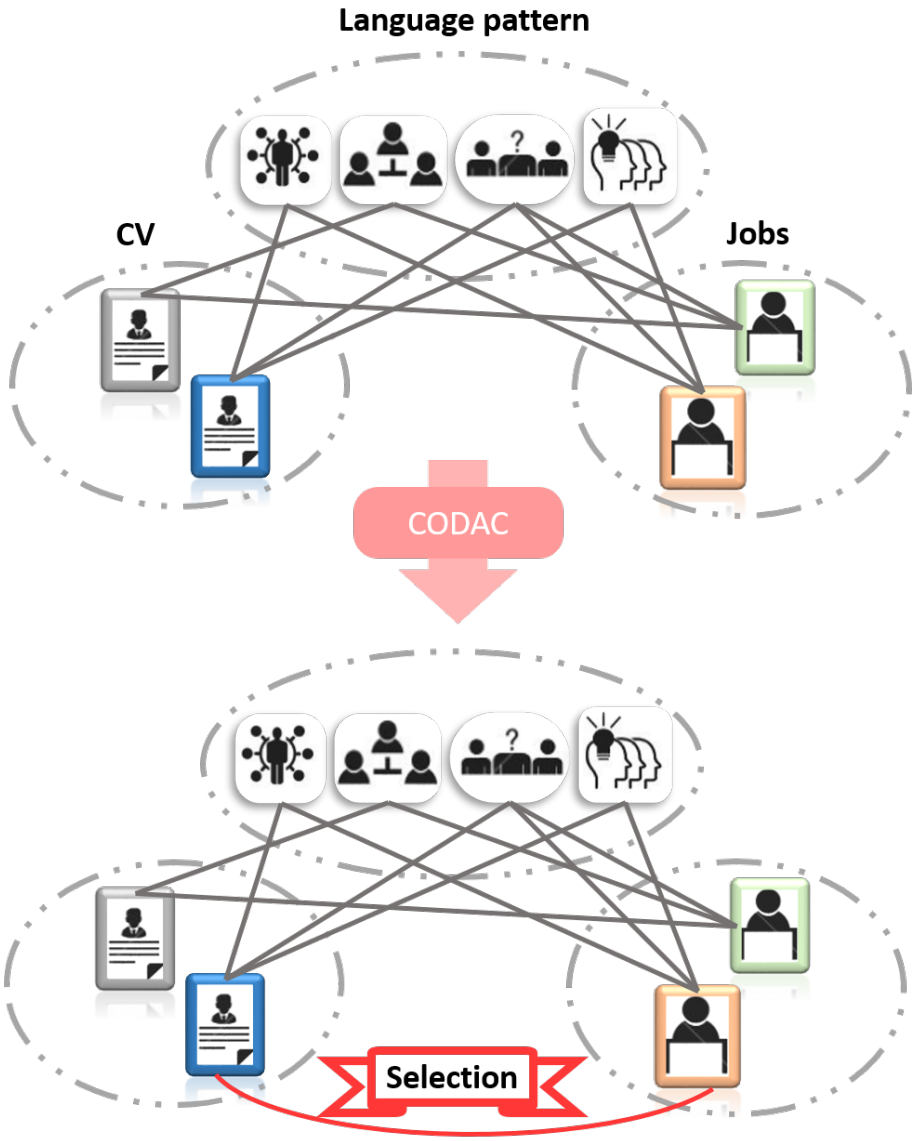


Figure 6.1: Schematic illustration of finding best resumes for job advertisements. In an instantiation of CODAC, X is a set of resumes, Y a set of job advertisement, and Z a set of language patterns extracted from the text of resumes and job descriptions. Selection dataset contains all newly discovered resume-job associations which are sorted and represented as the list of best candidates.

Nancy to confirm our predictions on a small number of proteins (636 proteins) encoded by particular genetic elements.

Another way to verify our predictions is to use observed domain-domain interaction databases such as KBDOCK, and check if the interacting domains have same the functions. This can be extended to the observed protein-protein interactions and verify functions of interacting partners in whole protein level.

Appendix A

ECDomainMiner/GODomainMiner Web-Servers

A.1 Introducing the ECDomainMiner/GODomainMiner Web Server

The ECDomainMiner and GODomainMiner are developed as a web-server, which we believe will of remarkable interest to the functional annotation community.

The ECDomainMiner web server may be queried by EC number or Pfam domain (Figure A.1). Thus, if one wishes to search for associations for a protein chain that currently lacks any EC annotation in the PDB (e.g. chain 2q7xA), one first needs to retrieve from the PDB the Pfam domain(s) that it contains (in this example, PF01933). Then, querying the ECDomainMiner server with each Pfam domain identifier will show the associated EC numbers (in this example, 2.7.8.28), along with the associated filtering scores and quality classes. In this example, ECDomainMiner finds a Gold quality association between PF01933, present in PDB chain 2q7xA, and EC number 2.7.8.28 (2-phospho-L-lactate transferase) which consequently can be associated with PDB entry 2q7x. Interestingly, PDB entry 2q7x is described as a putative phospho transferase from *streptococcus pneumoniae* tigr4, which is consistent with the enzymatic activity found by ECDomainMiner, and which could not be deduced from the Pfam domain name (UPF0052).

The GODomainMiner web server works in a very similar way and can be queried by GO term or Pfam domain (Figure A.2).

A.2 Implementation Details

ECDomainMiner and GODomainMiner web servers were written principally in the PHP scripting language, and JavaScript. PHP was used for creating all the transactions between client and servers. jQuery is one of the most interesting libraries in JavaScript, and it was used for handling the web page events. DataTables Table is a plug-in for jQuery and has been used to show the result. Data are store in MySQL and queries are processed using the PHP MySQL interface. Data in the MySQL database are prepared with Python scripts. The web interface has been tested using several popular browsers for the Windows, Linux, and Mac OS X operating systems.

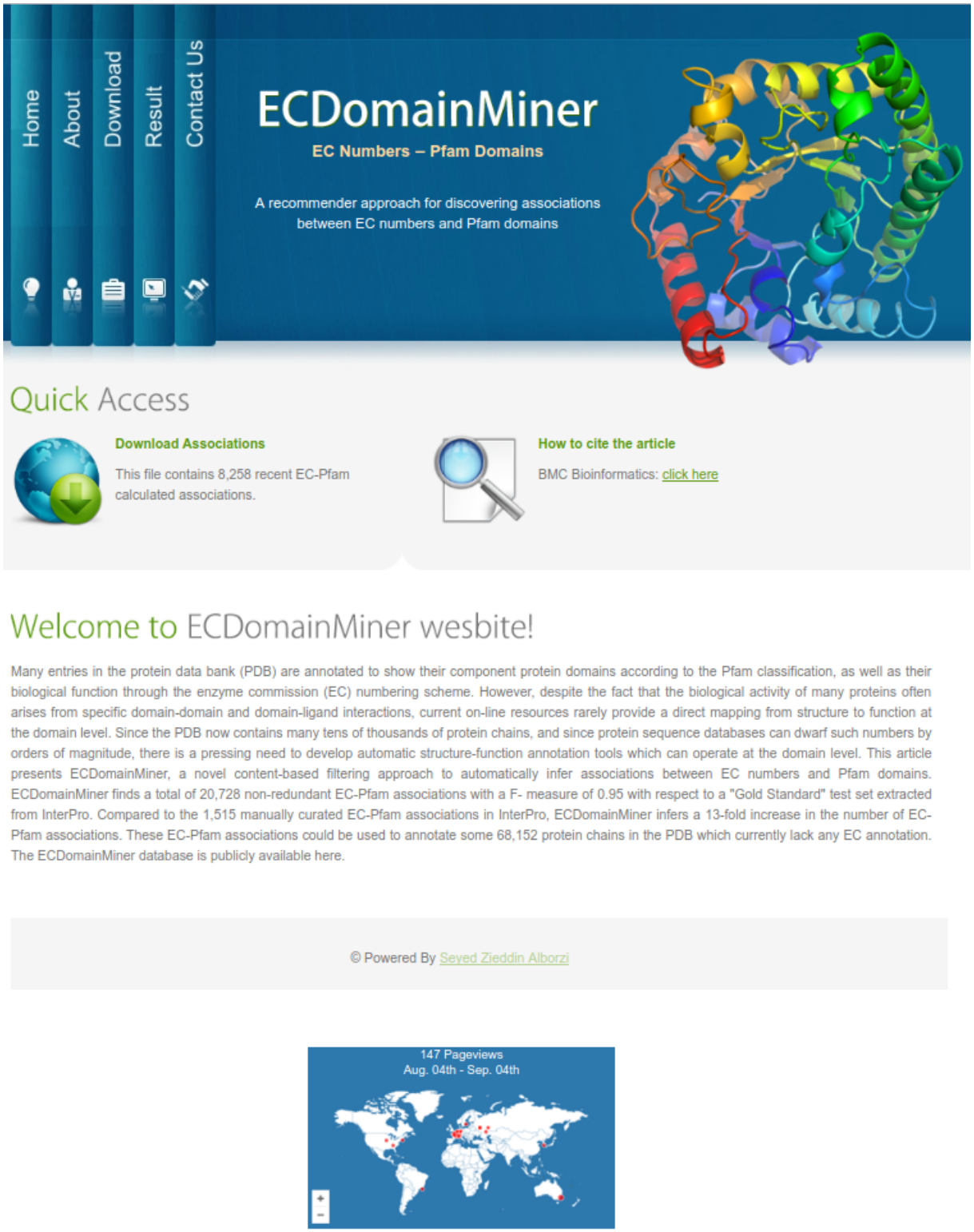


Figure A.1: A screenshot of the ECDomainMiner Home page

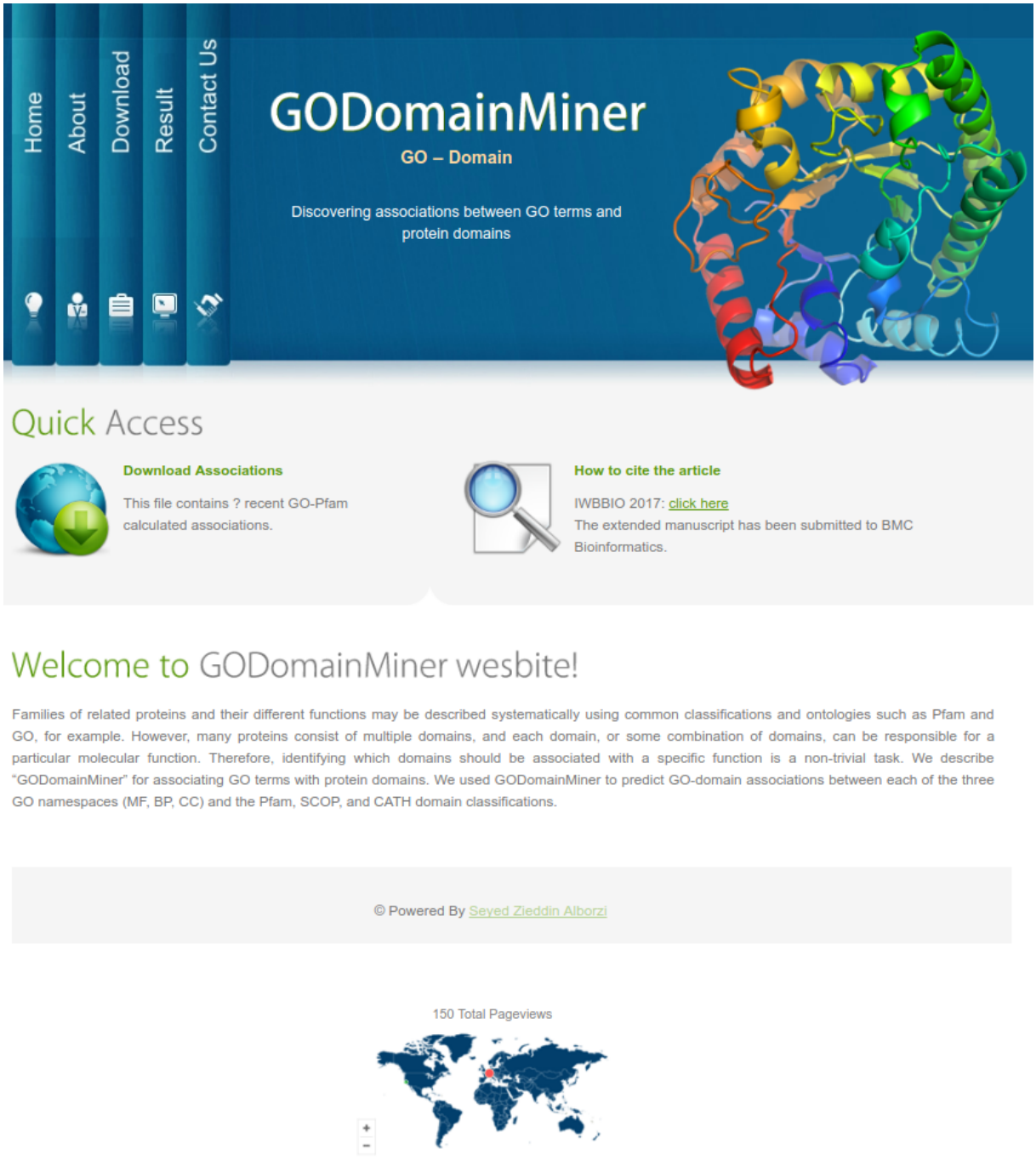


Figure A.2: A screenshot of the GODomainMiner Home page

Appendix B

Integrating inferred EC-domain and GO-domain in KBDOCK 2 Server

B.1 Introduction to KBDOCK 2

Three years ago KBDOCK web server is introduced for analyzing 3D protein domain interactions according to the Pfam protein domain family classification [Ghoorah et al., 2013a]. The original KBDOCK web server (http://kbdock.loria.fr/) has since received over 22,000 distinct visits, thus demonstration that the server provides a useful resource for the community.

We have recently, updated and extended the KBDOCK server and database, which we believe will of considerable interest to the structural bioinformatics community. For comparison and evaluation purposes, the new server is currently available with a new URL (http://kbdock2.loria.fr/). Notable features of the new server include:

- The server's database has been re-built using the September 15 2016 snapshot of the PDB and version 30.0 (September 2016) of the Pfam domain classification, giving an increase of over 33% in the number of Pfam domains having 3D structures (385,686 Pfam domain structures compared to 288,309 in 2013), and a 40% increase in the number of 3D domain-domain interactions (334,748 compared to 239,494), as well as similar increases in the coverage of domain-peptide interactions.

- 3D visualisation of structural interactions by the old Jmol and Jsmol tools has been replaced by "PV" (https://biasmv.github.io/pv/), a modern HTML5 graphical interface which allows for higher quality graphics and fast hardware rendering on the user's desktop.

- A new tooltip has been added to the results pages which provides quick access to the Enzyme Classification (EC) number and GO gene ontology entries for each Pfam domain through links to the Brenda (http://www.brenda-enzymes.org/), Amigo (http://amigo.geneontology.org/amigo/), and QuickGO (https://www.ebi.ac.uk/QuickGO/) web services, respectively. This tooltip also proposes EC numbers and GO terms that have been predicted by our new software tools "ECDomainMiner" [Alborzi et al., 2017c], and "GODomainMiner" [Alborzi et al., 2017b].

Overall, we believe the new KBDOCK server provides a convenient way for users to analyze the latest available 3D protein domain interactions and to consider the known and predicted functional annotations of those interactions from a structural point of view.

The web site contains an easy-to-use Help page (http://kbdock2.loria.fr/help.php) which explains through worked examples how to use the server (this section is currently being updated to show screen-shots with PV instead of the old Jmol).

B.2 Functions Associated with Pfam Domains in KBDOCK 2

Figure B.1 depicts the EC and GO functions associated to Trypsin domain, PF00089, in the KBDOCK2 server. EC and GO Functions are divided into two groups; The existing functions in the InterPro database, and predicted functions by ECDomainMiner and GODomainMiner. Clicking on the EC numbers or GO terms opens the function information webpage in the Brenda, QuickGO or Amigo websites.

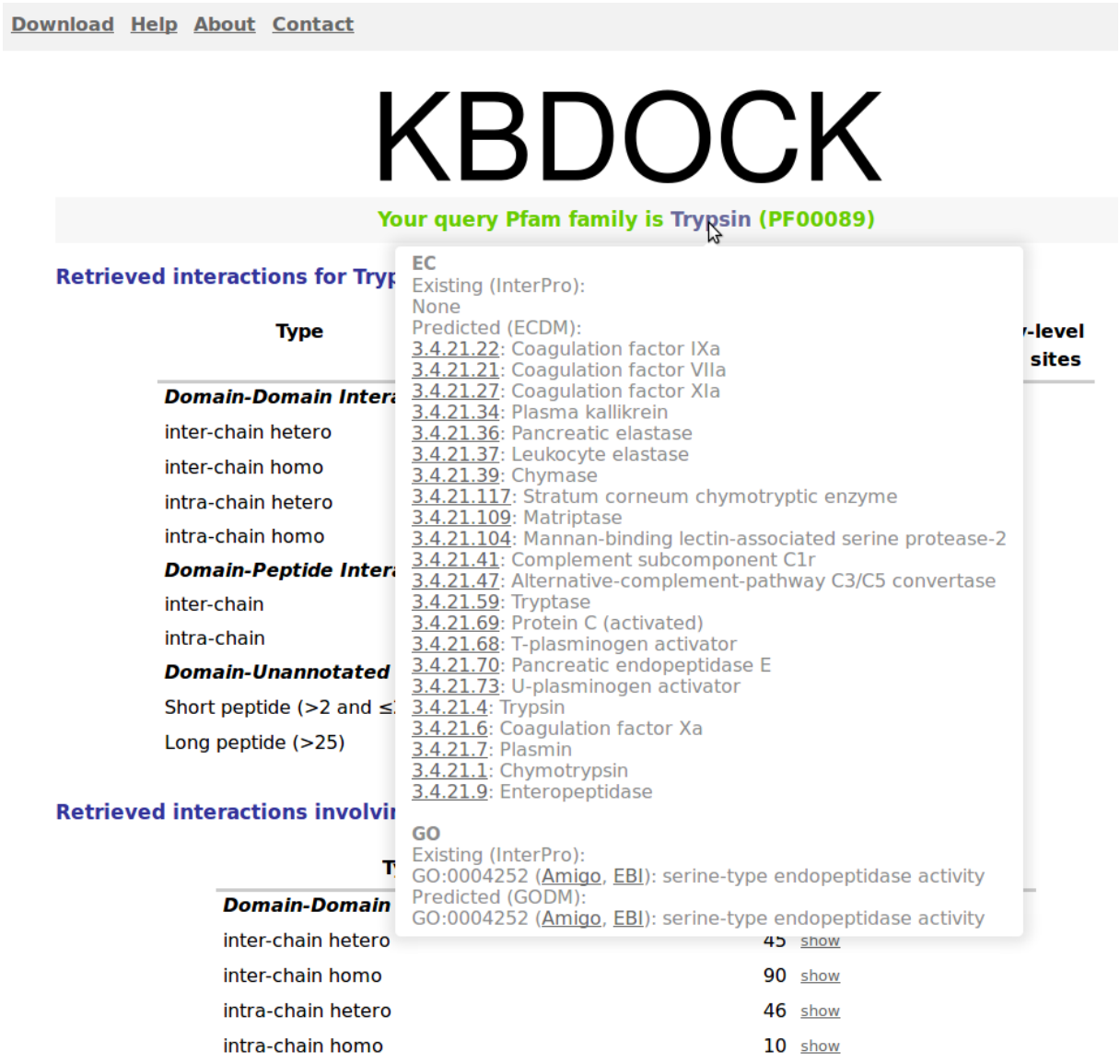


Figure B.1: 22 EC numbers and 2 GO terms which are associated to the Trypsin domain (PF00089).

Appendix C

Scientific Articles and Posters

C.1 Published Journal and Conference Papers

C.1.1 EC-PSI: Associating Enzyme Commission Numbers with Pfam Domains

With the growing number of protein structures in the protein data bank (PDB), there is a need to annotate these structures at the domain level in order to relate protein structure to protein function. Thanks to the SIFTS database, many PDB chains are now cross-referenced with Pfam domains and enzyme commission (EC) numbers. However, these annotations do not include any explicit relationship between individual Pfam domains and EC numbers. This article presents a novel statistical training-based method called EC-PSI that can automatically infer high confidence associations between EC numbers and Pfam domains directly from EC-chain associations from SIFTS and from EC-sequence associations from the SwissProt, and TrEMBL databases. By collecting and integrating these existing EC-chain/sequence annotations, our approach is able to infer a total of 8,329 direct EC-Pfam associations with an overall F-measure of 0.819 with respect to the manually curated InterPro database, which we treat here as a "gold standard" reference dataset. Thus, compared to the 1,493 EC-Pfam associations in InterPro, our approach provides a way to find over six times as many high quality EC-Pfam associations completely automatically.

EC-PSI: Associating Enzyme Commission Numbers with Pfam Domains

Seyed Ziaeddin ALBORZI[1,2], Marie-Dominique DEVIGNES[3] and David W. RITCHIE[2]

[1] Université de Lorraine, LORIA, UMR 7503, Vandœuvre-lès-Nancy, F-54506, France
[2] INRIA, Villers-lès-Nancy, F-54600, France
[3] CNRS, LORIA, UMR 7503, Vandœuvre-lès-Nancy, F-54506, France

Corresponding Author: dave.ritchie@inria.fr

Abstract *With the growing number of protein structures in the protein data bank (PDB), there is a need to annotate these structures at the domain level in order to relate protein structure to protein function. Thanks to the SIFTS database, many PDB chains are now cross-referenced with Pfam domains and enzyme commission (EC) numbers. However, these annotations do not include any explicit relationship between individual Pfam domains and EC numbers. This article presents a novel statistical training-based method called EC-PSI that can automatically infer high confidence associations between EC numbers and Pfam domains directly from EC-chain associations from SIFTS and from EC-sequence associations from the SwissProt, and TrEMBL databases. By collecting and integrating these existing EC-chain/sequence annotations, our approach is able to infer a total of 8,329 direct EC-Pfam associations with an overall F-measure of 0.819 with respect to the manually curated InterPro database, which we treat here as a "gold standard" reference dataset. Thus, compared to the 1,493 EC-Pfam associations in InterPro, our approach provides a way to find over six times as many high quality EC-Pfam associations completely automatically.*

Keywords Enzyme Commission Number (EC Number), Pfam Domains, Protein Structure Annotation, Machine Learning.

1. Introduction

Proteins are macromolecules comprising one or more chains of amino acid residues. Protein molecules carry out many essential biological functions such as catalysing metabolic reactions and mediating signals between cells, for example. These functions are often carried out by distinct "domains", which may often be identified as highly conserved regions within a multiple alignment of the sequences of a group of similar proteins, as in the Pfam database [1], for example. It is widely accepted that such protein domains often correspond to distinct and stable three-dimensional (3D) structures, and that there is often a close relationship between protein structure and protein function [2]. Indeed, it is well known that protein structures are often more highly conserved than protein sequences [3], and this suggests that proteins with similar structures will have similar biological functions [4]. The Protein Data Bank (PDB) [5, 6] now contains over 107,000 3D structures, most of which have been solved by X-ray crystallography or NMR spectroscopy. Structure-based classifications of protein domains such as SCOP [7] and CATH [8] have revealed many conserved structure-function relationships at the molecular level, and these classifications are now widely used in the community. However, because there does not exist a standard way to define a protein domain precisely, there is not always a one-to-one correspondence between domains defined by SCOP and those defined by CATH, for example, or between such structural domains and the domains defined by Pfam.

As well as sequence-based and structure-based classifications, proteins may also be classified according to their function. For example, the Enzyme Commission [9] uses a hierarchical four-digit numbering system to classify enzymatic function of many proteins. The first digit, or top-level "branch" of the hierarchy, selects one of six principal enzyme classes (oxidoreductase, transferase, hydrolase, lyase, isomerase, and ligase). The second digit defines a general enzyme class (chemical substrate type). The third digit defines a more specific enzyme-substrate class (e.g. to distinguish methyl transferase from formyl transferase), while the fourth digit, if present, defines a particular enzyme substrate. However, it should be noted that because EC numbers are assigned according to the reaction catalyzed, it is possible for distinct proteins to be assigned the same EC number even if they have no sequence similarity or if they belong to different structural families.

While the above classification schemes are very useful, they do not generally provide a direct relationship between enzymatic function and a 3D domain structure or a (sequence-based) Pfam domain. Thus, except for single-domain proteins where the mapping is obvious, unless a 3D structure has been very carefully annotated at the time it was deposited in the PDB (which is often not the case), it is generally not possible to compare and classify structure-function relationships at the domain level. Nonetheless, several groups have described approaches or resources that can associate PDB protein chains with enzyme EC numbers. For example, both the IMB Jena library [10] and the latest version of the PDBsum web site [11, 12] map each chain from a PDB file to its component CATH and Pfam domains, and each provides a link to the Enzyme database [13] for each PDB chain that has an EC number. PDBSprotEC [14] maps PDB chains to SwissProt and then uses the Enzyme database to obtain a mapping between SwissProt codes and EC numbers. Additional partial EC assignments are also retrieved directly from SwissProt. Columba [15] integrates annotation data from 12 different databases including PDB, SwissProt, CATH, SCOP, and Enzyme. For each PDB entry that has an EC number, Columba annotates the biological unit with the enzyme name and biochemical reaction, and it links SCOP and CATH domain information to each protein chain. PDB-UF [16] aims to assign EC numbers to unannotated protein structures which have no detectable sequence similarity to other proteins of known function. This approach first clusters existing protein structures using the 3D-hit structure alignment program [17]. It then assigns an unknown query structure to the most similar cluster, and it assigns a complete or partial EC number to the query using the EC numbers found in the cluster. Probably the most up-to-date and exhaustive association between PDB chains and EC numbers is currently provided by SIFTS [18], which is a collaboration between the Protein Data Bank in Europe and UniProt [19]. SIFTS incorporates a semi-automated procedure which links PDB chain entries to external biological resources such as Pfam, IntEnz [13], CATH and SCOP.

While all of the above approaches can provide associations between PDB protein chains and enzyme EC numbers, to our knowledge, SCOPEC [20] is the only published approach for automatically assigning EC numbers to structural domains. SCOPEC uses sequence information from SwissProt and PDB entries that have been previously annotated with EC numbers in order to assign EC numbers to SCOP domains. The SCOPEC approach first looks for PDB chains that fully map to SwissProt entries (to within up to 70 residues) and that match on at least the first three EC number digits. It then extracts the single domain structures which can thus be associated unambiguously with an EC number. It then uses these annotated domains as queries against the multi-domain structures to annotate homologous domains. It also uses the Catalytic Site Atlas [21] to locate catalytic domains in multi-domain structures. However, a limitation of the SCOPEC approach is that it normally associates EC numbers only with single domain proteins. Although SCOPEC can also propagate a known EC-domain association to a matching domain in a multi-domain protein, it is not designed to deconvolute EC-chain associations into individual EC-domain associations. Furthermore, it appears that the SCOPEC database is no longer available on-line. There is therefore a fresh need to develop a way of associating EC numbers with individual domains in order to study the large number of structural domains that now exist in the PDB.

Here, we present a novel statistical training-based approach for finding associations between EC numbers and Pfam domains directly from existing EC-chain associations from SIFTS and EC-sequence associations from SwissProt and TrEMBL. We call our approach "EC-PSI" (being short for "EC-Pfam statistical inferencing"). While SwissProt and TrEMBL were originally developed separately, both databases have since been incorporated in the UniProt resource. SwissProt is now a high quality, non-redundant, and manually curated part of UniProt Knowledge Base (UniProtKB). In contrast, TrEMBL is an automatically annotated and unreviewed section of UniProtKB, and contains around 40 times more entries than SwissProt. In order to parameterise and evaluate EC-PSI, we use the InterPro database [22] which contains a large number of manually curated Pfam-EC associations. Thus it may be used as a "gold standard" reference dataset against which our predicted Pfam-EC associations may be compared.

2. Methods

a. Data Preparation

Flat data files of SIFTS (October 2014), SwissProt and TrEMBL (November 2014), and InterPro (version 48.0) were downloaded and parsed using in-house Python scripts. From the SIFTS data, we extracted associa-

tions between Pfam domains and PDB chains, and associations between PDB chains and EC numbers. These associations were imported into two tables of our relational database. In the first table, each PDB chain is related to one or more Pfam domains (FIG. 1 A). In the second table, each PDB chain is related to one or more EC numbers (FIG. 1 B). Thus, these two tables together define a many-to-many relation between EC numbers and Pfam domains (FIG. 1 C). UniProtKB provides another source of relationships between EC numbers and Pfam domains. However, these relationships are mediated by UniProt accession numbers (ANs) instead of PDB chains. Since UniProtKB is divided into SwissProt and TrEMBL, we parsed and extracted the corresponding AN-Pfam and AN-EC associations from the SwissProt and TrEMBL databases, and we stored the resulting many-to-many relations in two further pairs of tables, similar to the two SIFTS tables.

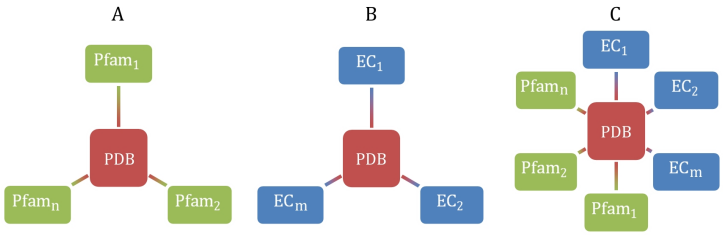


Figure 1. Illustration of the relationships extracted from the SIFTS database between (A) PDB chains and Pfam domains, (B) PDB chains and EC numbers, and (C) the many-to-many relationship between EC numbers and Pam domains.

As mentioned above, we used the InterPro manually curated EC-Pfam associations as a "gold standard" reference dataset. When considering only full four-digit EC numbers, we extracted a total of 1,493 EC-Pfam associations from InterPro, which we stored in our MySQL relational database. However, because we assume that all of the InterPro relations are "true" (i.e. correct) EC-Pfam associations, we needed to generate some plausible examples of false relations in order to train the EC-PSI algorithm. We therefore used our confidence score (see Section 2.b) to calculate and rank all possible EC-Pfam associations from SIFTS, SwissProt, and TrEMBL, and we extracted and stored 1,493 low-scoring EC-Pfam associations which could be calculated using data from at least two of the three databases. Because these associations have very little support in the data, we consider them to be "false" associations for the purpose of training our algorithm. In the rest of this paper, we will refer to the combined set of 1,493 "true" EC-Pfam associations from InterPro and our 1,493 calculated "false" associations as our "GoldStandard" dataset.

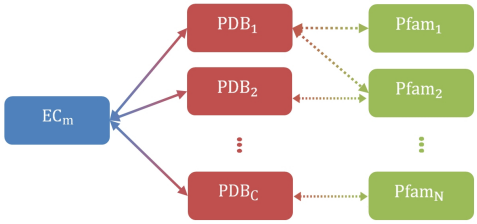


Figure 2. Graphical representation of the relationships between an EC number, m, and N Pfam domains via C PDB chains.

b. Inferring Associations Between EC Numbers and Pfam Domains

In order to infer direct Pfam-EC relations from each of the above data sources, we collected all tuples of SIFTS data in the form (EC,PDB,Pfam), and we sorted these tuples by four-digit EC number and then by PDB chain in order to extract a tree-like set of relations for each EC number, as illustrated in FIG. 2. A similar sorting procedure was applied to the corresponding tuples extracted from the SwissProt and TrEMBL datasets. Then,

for each EC number, we analyse its tree of associations by counting the numbers of occurrences of PDB chains (or ANs for SwissProt and TrEMBL) and Pfam domains. More specifically, for each Pfam domain within an EC tree, we calculate an EC-Pfam frequency score as the ratio between the number of chains in the tree that possess the given Pfam domain and the total number of PDB chains in the tree. In particular, letting m denote an EC number, i denote a PDB chain identifier, and supposing that the m^{th} EC tree contains C^m PDB chains denoted by P_i^m (for $i = 1, ...C^m$) and that D_n denote the n^{th} Pfam domain, we define the PPFEC ("Pfam-PDB Frequency for a given EC-Pfam association") score as

$$PPFEC_n^m = \frac{|\{P_i^m; D_n \in P_i^m, i = 1, ..., C^m\}|}{C^m}, \tag{1}$$

where $|\{P^m\}|$ denotes the cardinality of a set of PDB chains. The notation $D_n \in P_i^m$ is understood to mean that chain P_i^m possesses domain D_n. Equation (1) may be understood more graphically by considering FIG. 2. For a given EC number, m, and a given Pfam domain, n, the PPFEC is calculated as the degree of the Pfam node (number of connecting dashed lines) divided by the degree of the EC node (number of solid lines).

The corresponding frequencies for an inferred association between a Pfam domain and an EC number derived from the SwissProt and TrEMBL sequence annotations may be calculated in a similar way to give a "PSFEC" score (Pfam-SwissProt Frequency for a given EC-Pfam relation), and a "PTFEC" score (Pfam-TrEMBL Frequency for a given EC-Pfam relation), respectively. Thus, we obtain a frequency-based association score for each of the three data sources. However, because we wish to draw upon the relations from all three datasets, we combine the three frequency scores to give a single normalised "confidence score",

$$ConfidenceScore_{m,n} = \frac{a \times PPFEC_{m,n} + b \times PSFEC_{m,n} + c \times PTFEC_{m,n}}{(a + b + c)}, \tag{2}$$

where a, b, and c are weight factors, to be determined, and where an individual frequency score is set to zero whenever there is missing data for a given m and n.

In order to find the best values for the above three weight factors, we varied their values from 0.0 to 1.0 in steps of 0.1, and for each combination we scored and ranked each of the 2,986 GoldStandard associations. Next, using the ranked list of true and false associations, we labeled true associations found in the top half of the ranked list as true positives (TPs), and we labeled true associations found in the bottom half of the list as false negatives (FNs). Similarly, we labeled false associations found in the top half of the list as false positives (FPs), and false associations in the bottom half as true negatives (TNs). We then calculated a Receiver-Operator (ROC) curve [23] of the TP rate against the FP rate, and we used the area under the curve (AUC) of the ROC plot as the overall quality measure of the scoring function.

c. Defining a Confidence Score Threshold

Given that the best weights for each data source have been determined, we next wished to determine an overall threshold for our EC-Pfam association confidence score. In order to do this in an objective way, we randomly split the GoldStandard dataset into two equal groups with equal numbers of true and false instances to give a "Training" dataset and a "Test" dataset. Next, we scored and ranked the members of the Training dataset, and we divided the ranked list into two subsets according to a threshold value that ranged from 0.0 to 1.0 in steps of 0.01. For each threshold value, we counted the number of TPs (true associations above the threshold), FPs (false associations above the threshold), TNs (false associations below the threshold), and FNs (true associations below the threshold). We then calculated the recall, R, precision, P, and their harmonic mean in order to obtain the "F-measure" according to

$$R = \frac{TP}{TP + FN}, \qquad P = \frac{TP}{TP + FP}, \quad \text{and} \quad F = \frac{2RP}{P + R}. \tag{3}$$

The score threshold that gave the best F-measure was selected as the best threshold to use for accepting predicted associations.

3. Results and Discussion

a. Parameters of Our EC-PSI Procedure

Our EC-PSI procedure takes as input three large datasets of EC-chain associations from SIFTS, and EC-sequence associations from SwissProt and TrEMBL. These individual source datasets, which contain $6,204$, $9,879$, and $28,572$ associations respectively, were merged to give a global dataset of 32,018 non-redundant EC-Pfam associations. Our scoring function was trained using our GoldStandard dataset consisting of 1,493 "true" associations taken from InterPro and 1,493 "false" associations taken from low-scoring associations from SIFTS, SwissProt, and TrEMBL. We found that the best ROC-plot AUC is obtained with weights $a = 0.1$, $b = 1.0$, and $c = 0.1$ (Section 2.a), for a maximum AUC value of 0.888. These weights clearly give a 10-fold greater importance to the associations derived from SwissProt than to those derived from SIFTS and TrEMBL.

Using these weights, various threshold values of the confidence score were tested on the "Training" subset of our GoldStandard dataset, using the F-measure to quantify the results objectively (Section 2.b). The optimal score threshold was found to be 0.08 for a maximum F-Measure of 0.828. Applying this threshold to the GoldStandard Test subset yielded a comparable F-measure value of 0.810, and precision and recall values of 0.948 and 0.707, respectively. This threshold was then used to infer new EC-Pfam relations from the merged dataset, with a confidence score for each association being calculated by our scoring function.

b. Global Analysis of Calculated EC-Pfam Associations

The results of the filtering process are summarized in Table 1. This table shows the numbers of EC-Pfam associations along with the numbers of distinct EC numbers and Pfam entries involved in those associations for the three source datasets, our merged global dataset before and after filtering (the latter corresponding to our "calculated" EC-Pfam associations), and for the InterPro dataset of true associations. The overlap between these two last datasets is shown in the last line of the table.

Dataset	EC-Pfam associations	4-digit EC numbers	Pfam entries
SIFTS	6,204	2,575	2,606
SwissProt	9,879	3,959	3,147
TrEMBL	28,572	3,538	5,839
Merged	32,018	4,588	6,290
InterPro	1,493	676	1,273
EC-PSI (calculated)	$\mathbf{8,329}$	$\mathbf{4,436}$	$\mathbf{2,462}$
Common to EC-PSI and InterPro	1,089	592	944

Table 1. Statistics on the given and calculated EC-Pfam associations.

Overall, Table 1 shows that our EC-PSI procedure yielded a total of $8,329$ calculated EC-Pfam associations that include $1,089$ associations already present in InterPro. While this shows that EC-PSI finds 73% ($100 * 1,089/1,493$) of the "correct" EC-Pfam associations in InterPro, it also shows that 27% ($404/1,493$) of correct InterPro associations have EC-PSI confidence scores below our chosen score threshold of 0.08. This relatively high proportion of "missed" associations reflects the fact that our EC-PSI method is designed to discover EC-Pfam associations with strong factual support, whereas InterPro contains a large number of low frequency expert-annotated associations. More specifically, the score threshold of 0.08 was chosen to give a good trade-off between precision and recall through the F-score. If, for example, the score threshold is reduced from 0.08 to 0.01, the recovery of correct InterPro associations increases to 90% (1,354/1,493), but the number of "false" InterPro associations rises from just 75 to 822.

Given that InterPro may be considered to represent the largest manually curated source of Pfam-EC associations currently available, it is interesting to consider the relative increase in the number of associations that our EC-PSI approach can provide. We therefore calculated as ratios (or "scale-up factors") the differences between the associations calculated by EC-PSI and those of InterPro in terms of the total number of associations and the numbers of distinct EC numbers and Pfam entries involved in those associations. In FIG. 3, the scale-up factors are displayed across the 6 top-level branches of the EC classification (1-6) and for the entire datasets (All). It can be seen that the scale-up factors for EC-Pfam associations and for EC entries reach their maximum

levels in branch 1 (oxydoreductases), and that they fluctuate around their average values (All) rather evenly in all other branches, with branch 6 (ligases) having the lowest values. The same is true for the scale-up factor for the number of Pfam entries, but the difference is less marked in branch 1. In fact, the average increase in Pfam entries is only about 2-fold compared to about 6-fold for Pfam-EC associations and EC numbers. This is consistent with the fact that not all Pfam entries can be assigned an EC number because not all Pfam domains are associated with an enzymatic activity.

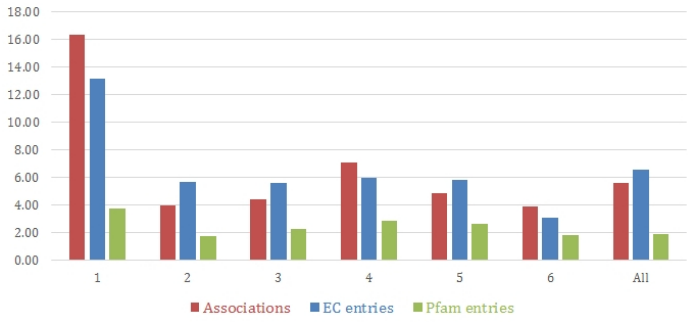


Figure 3. Scale-up factors for the EC-PSI and InterPro associations according to the EC branch. 1 : oxydoreductases ; 2 : transferases ; 3 : hydrolases ; 4 : lyases ; 5 : isomerases ; 6 : ligases ; All : all EC numbers.

c. Comparison Between Calculated and InterPro EC-Pfam Associations

In FIG. 4 A, the average number of EC-Pfam associations is plotted per EC number (1) and per Pfam entry (2) for both InterPro and our calculated dataset. The ratios are very close for the EC numbers (2.2 and 1.9, respectively), suggesting that our method follows the quality of annotation of InterPro and does not propose an excess of possibly incorrect EC-Pfam associations. On the other hand, the ratio is much higher for Pfam entries (3.38 versus 1.17), which reflects a significant enrichment in the annotation of Pfam domains. The rest of the figure shows the distribution of EC numbers (B) and Pfam entries (C) with respect to the number of associations they are involved in. Clearly the proportion of EC numbers (respectively, Pfam entries) that are involved in only one EC-Pfam association is reduced in our calculated dataset.

Overall, FIG. 4 shows that our collection of EC-Pfam associations rather favours multiple associations, thereby reflecting the complex many-to-many relationships that exists within the original datasets. Furthermore, many of the multiple associations calculated by EC-PSI seem to be quite reasonable from a biological point of view. For example, EC-PSI finds the unique InterPro association between EC 6.1.1.9 (valine-tRNA ligase) and the Pfam domain PF10458 (Valyl tRNA synthetase, tRNA binding arm) with a confidence score of 0.781, but it also finds two further associations with the same EC number that are not in InterPro, namely with PF08264 (tRNA anticodon-binding domain ; EC-PSI score 0.976) and PF00133 (tRNA synthetases class I ; EC-PSI score 0.997). These two additional associations complete the biological picture of a tRNA ligase because they comprise a second constitutive domain, in addition to PF10458, of this complex enzyme. On the Pfam side, another interesting example is the unique InterPro association between PF04715 (Anthranilate synthase component, N terminal region) and EC 4.1.3.27 (anthranilate synthase). EC-PSI finds this association with a score of 0.522, but in addition it finds two further associations for the same Pfam domain, namely with EC 2.6.1.85 (aminodeoxychorismate synthase, EC-PSI score 0.675) and EC 2.6.1.86 (2-amino-4-deoxychorismate synthase ; EC-PSI score 0.833). In this case, the multiple association found by EC-PSI for PF04715 may be explained by the fact that all three enzymes share a common substrate (i.e. chorismate).

d. Future Work

The approach presented here calculates associations using four-digit EC numbers. However, because EC numbers have an embedded hierarchy, and because it seems reasonable to suppose that enzymes that act on

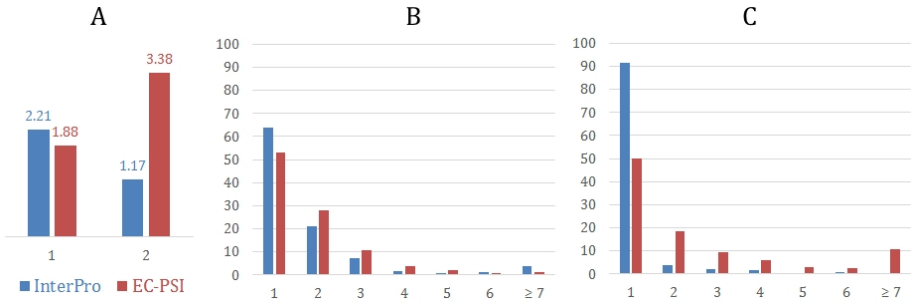


Figure 4. A : average number of EC-Pfam associations per EC number (1) and per Pfam entry (2) for the InterPro (blue) and calculated EC-PSI (red) datasets. B : distribution of EC numbers according to their numbers of associations with Pfam entries. C : distribution of Pfam entries according to their numbers of associations with EC numbers.

similar substrates are likely to be evolutionarily related, it could be interesting to consider making additional associations by collecting and analysing less specific three-digit associations. This could provide a way to infer additional associations that have weak direct support (low four-digit confidence scores), but which have good support at the three-digit level. We plan to analyse the support of EC numbers associated with more than one Pfam entry in order to detect those EC numbers that correspond to combinations of domains (in other words to detect cases where two or more domains are physically necessary to support a given enzyme function). We also want to improve the way that candidate associations from different sources are combined. Even though our current scoring function gives 10 times more weight to SwissProt than SIFTS and TrEMBL, it is still useful use all three data sources because our algorithm finds 312 EC-Pfam associations from SIFTS and 797 from TreMBL which are not present in the SwissProt data. However, it would be desirable to use a more statistically sound measure of the reliability of each data source, perhaps based on a comparision of the associations found after random shuffling of the data, for example.

4. Conclusions

Given the extensive protein chain/sequence annotations that now exist in the SIFTS, SwissProt, and TrEMBL databases, there is a need to be able to exploit this rich knowledge at the protein domain level. We achieved this aim by first collecting existing associations between EC numbers and protein chains or sequences, and then by using a statistical training-based scoring method to analyse the many-to-many relations embedded in these data. Using the above data sources, our approach is able to infer a total of 8,329 direct EC-Pfam associations. Thus, compared to the 1,493 manually curated InterPro EC-Pfam associations, our approach provides a way to find over six times as many associations completely automatically. We have also proposed some possible ways to extend and further analyse the coverage of the EC-PSI approach. We believe that the large numbers of EC-Pfam associations calculated using our approach can contribute considerably to enriching the annotations of PDB protein chains, and that this will facilitate a better understanding and exploitation of structure-function relationships at the protein domain level.

Acknowledgements

This project is funded by the Agence Nationale de la Recherche (grant reference ANR-11-MONU-006-02), the Institut National de Recherche en Informatique et Automatique (Inria), and the Lorraine Region.

Références

[1] R. D. Finn, A. Bateman, J. Clements, P. Coggill, R. Y. Eberhardt, S. R. Eddy, A. Heger, K. Hetherington, L. Holm, J. Mistry, Sonnhammer E. L. L., John Tate, and M. Punta. Pfam : the protein families database. *Nucleic Acids Research*, 42(D1) :D222–D230, 2014.

[2] J. M. Berg, J. L. Tymoczko, and L. Stryer. *Protein structure and function*. W.H. Freeman, 2002.

[3] C. Chothia and A. M. Lesk. The relation between the divergence of sequence and structure in proteins. *EMBO Journal*, 5(4) :823, 1986.

[4] A. C. R. Martin, C. A. Orengo, E. G. Hutchinson, S. Jones, M. Karmirantzou, R. A. Laskowski, J. B. O. Mitchell, C. Taroni, and J. M. Thornton. Protein folds and functions. *Structure*, 6(7) :875–884, 1998.

[5] F. C. Bernstein, T. F. Koetzle, G. J.B. Williams, E. F. Meyer, M. D. Brice, J. R. Rodgers, O. Kennard, T. Shimanouchi, and M. Tasumi. The protein data bank. *European Journal of Biochemistry*, 80(2) :319–324, 1977.

[6] A. Gutmanas, Y. Alhroub, G. M. Battle, J. M. Berrisford, E. Bochet, M. J. Conroy, J. M. Dana, M. A. F. Montecelo, G. van Ginkel, S. P. Gore, P. Haslam, R. Hatherley, P. M. S. Hendrickx, M. Hirshberg, I. Lagerstedt, S. Mir, A. Mukhopadhyay, T. J. Oldfield, A. Patwardhan, L. Rinaldi, G. Sahni, E. Sanz-García, S. Sen, R. A. Slowley, S. Velankar, Wainwright, and M. E. Kleywegt G. J. PDBe : protein data bank in europe. *Nucleic Acids Research*, 42(D1) :D285–D291, 2014.

[7] A. G. Murzin, S. E. Brenner, T. Hubbard, and C. Chothia. SCOP : a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4) :536–540, 1995.

[8] C. A. Orengo, A. D. Michie, S. Jones, D. T. Jones, M. B. Swindells, and J. M. Thornton. CATH – a hierarchic classification of protein domain structures. *Structure*, 5(8) :1093–1109, 1997.

[9] E. C. Webb. *Enzyme nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes.* Academic Press, 1992.

[10] J. Reichert, A. Jabs, P. Slickers, and J. Sühnel. The IMB Jena image library of biological macromolecules. *Nucleic Acids Research*, 28(1) :246–249, 2000.

[11] T. A. P. de Beer, K. Berka, J. M. Thornton, and R. A. Laskowski. PDBsum additions. *Nucleic Acids Research*, 42(D1) :D292–D296, 2014.

[12] R. A. Laskowski. PDBsum : summaries and analyses of PDB structures. *Nucleic Acids Research*, 29(1) :221–222, 2001.

[13] A. Fleischmann, M. Darsow, K. Degtyarenko, W. Fleischmann, S. Boyce, K. B. Axelsen, A. Bairoch, D. Schomburg, K. F. Tipton, and R. Apweiler. IntEnz, the integrated relational enzyme database. *Nucleic Acids Research*, 32(suppl 1) :D434–D437, 2004.

[14] A. C. R. Martin. PDBSprotEC : a web-accessible database linking PDB chains to EC numbers via SwissProt. *Bioinformatics*, 20(6) :986–988, 2004.

[15] S. Trißl, K. Rother, H. Müller, T. Steinke, I. Koch, R. Preissner, C. Frömmel, and U. Leser. Columba : an integrated database of proteins, structures, and annotations. *BMC Bioinformatics*, 6(1) :81, 2005.

[16] M. von Grotthuss, D. Plewczynski, K. Ginalski, L. Rychlewski, and E. I. Shakhnovich. PDB-UF : database of predicted enzymatic functions for unannotated protein structures from structural genomics. *BMC Bioinformatics*, 7(1) :53, 2006.

[17] D. Plewczyński, J. Paś, M. von Grotthuss, and L. Rychlewski. 3D-Hit : fast structural comparison of proteins. *Applied Bioinformatics*, 1(4) :223–225, 2001.

[18] S. Velankar, J. M. Dana, J. Jacobsen, G. van Ginkel, P. J. Gane, J. Luo, T. J. Oldfield, C. O'Donovan, M. J. Martin, and G. J. Kleywegt. SIFTS : structure integration with function, taxonomy and sequences resource. *Nucleic Acids Research*, 41(D1) :D483–D489, 2013.

[19] The UniProt Consortium. The universal protein resource (UniProt) in 2010. *Nucleic Acids Research*, 38(suppl 1) :D142–D148, 2010.

[20] R. A. George, R. V. Spriggs, J. M. Thornton, B. Al-Lazikani, and M. B. Swindells. SCOPEC : a database of protein catalytic domains. *Bioinformatics*, 20(suppl 1) :i130–i136, 2004.

[21] C. T. Porter, G. J. Bartlett, and J. M. Thornton. The Catalytic Site Atlas : a resource of catalytic sites and residues identified in enzymes using structural data. *Nucleic Acids Research*, 32(suppl 1) :D129–D133, 2004.

[22] A. Mitchell, H. Y. Chang, L. Daugherty, M. Fraser, S. Hunter, R. Lopez, C. McAnulla, C. McMenamin, G. Nuka, S. Pesseat, A. Sangrador-Vegas, M. Scheremetjew, C. Rato, S.-Y. Yong, A. Bateman, M. Punta, T. K. Attwood, C. J. A. Sigrist, N. Redaschi, C. Rivoire, I. Xenarios, D. Kahn, D. Guyot, P. Bork, I. Letunic, J. Gough, M. Oates, D. Haft, H. Huang, D. A. Natale, C. H. Wu, C. Orengo, I. Sillitoe, H. Mi, P. D. Paul D. Thomas, and Finn R. D. The InterPro protein families database : the classification resource after 15 years. *Nucleic Acids Research*, 43(D1) :D213–D221, 2015.

[23] T. Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 7 :861–874, 2006.

C.1.2 ECDomainMiner: discovering hidden associations between enzyme commission numbers and Pfam domains

Many entries in the protein data bank (PDB) are annotated to show their component protein domains according to the Pfam classification, as well as their biological function through the enzyme commission (EC) numbering scheme. However, despite the fact that the biological activity of many proteins often arises from specific domain-domain and domain-ligand interactions, current on-line resources rarely provide a direct mapping from structure to function at the domain level. Since the PDB now contains many tens of thousands of protein chains, and since protein sequence databases can dwarf such numbers by orders of magnitude, there is a pressing need to develop automatic structure-function annotation tools which can operate at the domain level.

This article presents ECDomainMiner, a novel content-based filtering approach to automatically infer associations between EC numbers and Pfam domains. ECDomainMiner finds a total of 20,728 non-redundant EC-Pfam associations with a F-measure of 0.95 with respect to a "Gold Standard" test set extracted from InterPro. Compared to the 1515 manually curated EC-Pfam associations in InterPro, ECDomainMiner infers a 13-fold increase in the number of EC-Pfam associations.

Alborzi *et al. BMC Bioinformatics* (2017) 18:107
DOI 10.1186/s12859-017-1519-x

BMC Bioinformatics

RESEARCH ARTICLE Open Access

ECDomainMiner: discovering hidden associations between enzyme commission numbers and Pfam domains

Seyed Ziaeddin Alborzi[1,3], Marie-Dominique Devignes[1,2] and David W. Ritchie[3*]

Abstract

Background: Many entries in the protein data bank (PDB) are annotated to show their component protein domains according to the Pfam classification, as well as their biological function through the enzyme commission (EC) numbering scheme. However, despite the fact that the biological activity of many proteins often arises from specific domain-domain and domain-ligand interactions, current on-line resources rarely provide a direct mapping from structure to function at the domain level. Since the PDB now contains many tens of thousands of protein chains, and since protein sequence databases can dwarf such numbers by orders of magnitude, there is a pressing need to develop automatic structure-function annotation tools which can operate at the domain level.

Results: This article presents ECDomainMiner, a novel content-based filtering approach to automatically infer associations between EC numbers and Pfam domains. ECDomainMiner finds a total of 20,728 non-redundant EC-Pfam associations with a F-measure of 0.95 with respect to a "Gold Standard" test set extracted from InterPro. Compared to the 1515 manually curated EC-Pfam associations in InterPro, ECDomainMiner infers a 13-fold increase in the number of EC-Pfam associations.

Conclusion: These EC-Pfam associations could be used to annotate some 58,722 protein chains in the PDB which currently lack any EC annotation. The ECDomainMiner database is publicly available at http://ecdm.loria.fr/.

Keywords: Content-based filtering, Protein domain, Protein function, Enzyme commission number, Pfam domain

Background

Proteins perform many essential biological functions such as catalysing metabolic reactions and mediating signals between cells. These functions are often carried out by distinct "domains", which may be identified as highly conserved regions within a multiple alignment of a group of similar protein sequences, as in the Pfam classification [1]. It is widely accepted that such protein domains often correspond to distinct and stable three-dimensional (3D) structures, and that there is often a close relationship between protein structure and protein function [2]. Indeed, it is well known that protein structures are often more highly conserved than protein sequences [3], and this suggests that proteins with similar structures will have similar biological functions [4]. The Protein Data Bank (PDB) [5, 6] now contains over 107,000 3D structures, most of which have been solved by X-ray crystallography or NMR spectroscopy.

As well as sequence-based and structure-based classifications, proteins may also be classified according to their function. For example, the Enzyme Commission [7] uses a hierarchical four-digit numbering system to classify the enzymatic function of many proteins. The first digit, or top-level "branch" of the hierarchy, selects one of six principal enzyme classes (oxidoreductase, transferase, hydrolase, lyase, isomerase, and ligase). The second digit defines a general enzyme class (chemical substrate type). The third digit defines a more specific enzyme-substrate class (e.g. to distinguish methyl transferase from formyl transferase), while the fourth digit, if present, defines a particular enzyme substrate. However, it should be noted that because EC numbers are assigned according to the reaction catalyzed, it is possible for different

*Correspondence: dave.ritchie@inria.fr
[3]Inria Nancy Grand-Est, 54600 Villers-lès-Nancy, France
Full list of author information is available at the end of the article

proteins to be assigned the same EC number even if they have no sequence similarity or if they belong to different structural families.

Furthermore, there are several ways in which a protein might provide one or more enzymatic functions, as illustrated in Fig. 1. In the simplest case (Fig. 1a), a protein contains just one domain, and there is is a one-to-one association between that domain and a particular enzymatic function. In this case, it is reasonable to suppose that the catalytic site is located entirely on that domain. Similarly, a protein may have two or more distinct domains, each of which provides a distinct enzymatic (or non-enzymatic) function (Fig. 1b). On the other hand, a protein domain could be involved in more than one catalytic activity, as illustrated in Fig. 1c. Finally, a catalytic site may be at the interface between two domains, or one domain serves as a necessary co-factor for the other (Fig. 1d). Clearly, it is biologically relevant to be able to distinguish all such cases. However, except for the simplest case (Fig. 1a), it can be seen that finding domain-EC associations automatically is a non-trivial task. Several groups have described approaches or resources that can associate entire PDB protein chains with enzyme EC numbers [8–11]. Probably the most up-to-date and exhaustive association between PDB chains and EC numbers is provided by SIFTS [12], which is a collaboration between the Protein Data Bank in Europe and UniProt [13]. SIFTS incorporates a semi-automated procedure which links PDB chain entries to external biological resources such as Pfam, and IntEnz [14].

While all of the above mentioned approaches can provide associations between PDB protein chains and enzyme EC numbers, to our knowledge, very few approaches have been published for automatically assigning EC numbers to structural domains. SCOPEC [15] uses sequence information from SwissProt and PDB entries that have been previously annotated with EC numbers in order to assign EC numbers to SCOP domains [16]. It first looks for PDB chains that fully map to SwissProt entries (to within up to 70 residues) and that match on at least the first three EC number digits. In this way, SCOPEC identifies single domain structures that can be associated unambiguously with an EC number. Although SCOPEC can subsequently propagate a known EC-domain association to a matching domain in a multi-domain protein, it is generally not able to resolve cases where multiple ECs are associated with multi-domain chains (parts B, C, and D in Fig. 1. Furthermore, it appears that the SCOPEC database is no longer available on-line.

In contrast, the dcGO ontology database for protein domains produced in 2012 is still available online and provides several ontological annotations (Gene Ontology: GO, EC, pathways, phenotype, anatomy and disease ontologies) for more than 2000 SCOP domain families [17].

The dcGO approach follows the principle that if a GO term tends to be attached to proteins in UniProtKB that contain a certain domain, then that term should be associated with that domain. The statistical significance of an association is assessed against a random chance association using a hypergeometric distribution followed by multiple hypotheses testing in terms of false discovery rate. The dcGO approach addresses the issues of hierarchical structure of most biological ontologies and the nature of domain composition for multi-domain proteins. However, a mapping onto Pfam

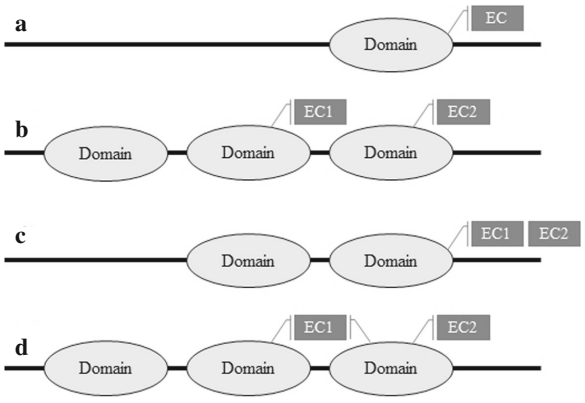


Fig. 1 a) One domain provides one enzyme function; (**b**) two domains on the same chain each provide a different enzyme function; (**c**) one domain provides two different enzyme functions; (**d**) one domain provides one enzyme function, while a second domain acts as a co-factor with the first domain to provide an additional enzyme function

domains is proposed only for GO terms and not for EC numbers.

Here, we describe a recommender-based approach call "ECDomainMiner" for associating Pfam domains with EC numbers, which builds on our previously described statistical approach [18]. Recommender systems are a class of information filtering system [19, 20] which aim to present a list of items that might be of interest to an on-line customer. There are two main kinds of recommender systems. Collaborative filtering approaches make associations by calculating the similarity between activities of users [21, 22]. Content-based filtering aims to predict associations between user profiles and description of items by identifying common attributes [20, 23]. Such an approach has recently been applied to a quite different problem of discovering novel cancer drug combinations [24].

Here, we use content-based filtering to associate EC numbers with Pfam domains from existing EC-chain and Pfam-chain associations from SIFTS, and from EC-sequence and Pfam-sequence associations from SwissProt and TrEMBL, where protein chains and sequences serve as the common attributes through which EC-Pfam associations are made. Note that our approach *does not* attempt to identify catalytic sites or catalytic residues. Rather, we aim to detect frequent co-occurrences of Pfam domains and EC numbers in order to deconvolute the often complex EC-Pfam relationships within multi-domain and multi-function protein chains. We assess the performance of our approach against a "Gold Standard" dataset derived from InterPro [25], and we compare our results with the Pfam-EC associations derived from the dcGO database. We also show how our database of more than 20,000 EC-Pfam associations can be exploited for automatic annotation purposes.

Methods

Data preparation

Our data sources are SIFTS for EC number and Pfam domain annotations of PDB chains, and Uniprot for EC number and Pfam domain annotations of protein sequences. UniProt is divided into three parts: (i) the non-redundant, high quality, manually curated SwissProt part, (ii) the TrEMBL data that are annotated using Unified Rules [26], called here UniRule, and (iii) the rest called here TrEMBL.

In addition, in order to parameterise and evaluate ECDomainMiner, we use the InterPro database [25] which contains a large number of manually curated EC-Pfam associations. Flat data files of SIFTS (July 2015), Uniprot (July 2015), and InterPro (version 53.0) were downloaded and parsed using in-house Python scripts. From the SIFTS data, associations between EC numbers and PDB chains, and associations between PDB chains and Pfam domains were extracted. Associations between Uniprot sequence accession numbers (ANs) and EC numbers, and AN-Pfam associations were then extracted from the SwissProt section of Uniprot to give a dataset of Swissprot associations. For the TrEMBL entries, we collected and stored the corresponding AN-EC and AN-Pfam associations which had been annotated by UniRule, and those associations lacking UniRule annotations to give two further sequence-based datasets of associations, which we call the UniRule and TrEMBL association datasets.

To avoid bias due to duplicate structures or sequences in the four source datasets, all PDB chains and Uniprot sequences were grouped into clusters having 100% sequence identity using the Uniref non-redundant cluster annotations [27], and each cluster was assigned a cluster unique identifier (CID). Note that since just a few point mutations can dramatically change an enzyme's substrate specificity, making clusters of identical rather than highly similar sequences avoids the risk of falsely clustering proteins that share highly similar folds but which have quite different substrates. The source EC-chain and EC-AN associations were then mapped to the corresponding CID in order to make four sets of EC-CID associations. A similar mapping was applied to the source Pfam-chain and Pfam-AN associations to give four sets of Pfam-CID associations.

For the reference data, we extracted from InterPro a total of 1515 EC-Pfam associations in which each EC number had all four digits and each Pfam accession number referred either to a Pfam domain or a Pfam family (i.e. Pfam motifs and repeats were excluded). These associations were considered to be "positive examples", and were randomly divided into two equal "training" and "test" subsets. However, for training purposes, we also needed some "negative examples". We therefore created a set of "false" EC-Pfam associations by first shuffling the CID-EC and CID-Pfam associations from SIFTS dataset, and by then randomly collecting 1515 wrong EC-Pfam associations from the shuffled datasets. In the rest of this article, we will refer to the combined set of 758 randomly chosen positive examples from InterPro and 758 randomly chosen negative examples as our "training dataset" and the remaining 1513 positive and negative examples as our "test dataset".

Inferring EC-Pfam domain associations

The main idea underlying the discovery of hidden EC-Pfam associations is to assign a feature vector to each EC number and each Pfam domain, where the length of the vector is given by the total number of PDB and UniProt CIDs, and where each vector element marks the existence (1) or absence (0) of an EC number or Pfam domain annotation for a particular CID. Each possible EC-Pfam association is then scored using the cosine similarity

between the corresponding pair of EC and Pfam feature vectors.

The various steps of our content-based filter approach for finding associations between 4-digit EC numbers and Pfam domains are illustrated in Fig. 2 for the SIFTS dataset. First, all relations between PDB CIDs and EC numbers, and between PDB CIDs and Pfam domains are extracted from SIFTS, as described above. Joining these two lists of relations then yields a complex many-to-many graph that contains relations between EC numbers, PDB CIDs, and Pfam domains.

After this join operation, all EC-CID relations are encoded in a binary matrix, where a 1 represents the presence of an association and a 0 represents no association. This matrix is then row-normalised such that each row has unit magnitude when considered as a vector. Similarly, all PDB CID-Pfam relations are encoded in a second binary matrix which is column-normalised. Consequently, the product of the two normalised matrices corresponds to a matrix of cosine similarity scores between the rows of the first matrix and the columns of the second matrix. Thus, each element, $S(ec, d)$, of the product matrix represents a raw association score between an EC number, *ec*, and a Pfam domain, *d*.

Similarly, raw EC-Pfam association scores are calculated from EC-CID and Pfam-CID relations extracted from SwissProt, TrEMBL and Unirule. Then, because we wish to draw upon the relations from all four datasets, we combine the four raw scores as a weighted average to give a single normalized confidence score, $CS_{ec,d}$:

$$CS_{ec,d} = \frac{\sum_i w_i S_i(ec, d)}{\sum_i w_i} \tag{1}$$

where $i \in \{SIFTS, Swissprot, TrEMBL, UniRule\}$ enumerates the four datasets, w_i are weight factors, to be determined, and where an individual association score, $S_i(ec, d)$, is set to zero whenever there is no data for the (ec, d) pair in dataset *i*.

In order to find the best values for the four weight factors, receiver-operator-characteristic (ROC) curves [28] were calculated using the positive examples of our Interpro-based training dataset, against the remaining associations (background associations).

Each weight was varied from 0.0 to 1.0 in steps of 0.1, and for each combination of weights a ROC curve of the ranked association scores was calculated. The combination of weights that gave the largest area under the curve (AUC) of the ROC curve was selected.

Defining a confidence score threshold

Having determined the best weight for each data source, we next wished to determine an overall threshold for the confidence score. To do this in an objective way, we scored and ranked the members of the training dataset, and labeled them true or false according to a threshold value that was varied from 0.0 to 1.0 in steps of 0.01. For each threshold value, we counted the number of positive examples above the threshold (TPs), negative examples above the threshold (FPs), negative examples below the threshold (TNs), and positive examples below the threshold (FNs). We then calculated the recall, *R*, precision, *P*, and their harmonic mean in order to obtain a "F-measure" using:

$$R = \frac{TP}{TP + FN}, \qquad P = \frac{TP}{TP + FP}, \qquad F = \frac{2RP}{P + R}. \tag{2}$$

The score threshold that gave the best F-measure was checked on the test subset and selected as the best threshold to use for accepting inferred associations.

Exploiting the EC number hierarchy

The above approach has focused on finding explicit co-occurrences between Pfam domains and 4-digit EC numbers. However, it is possible to find more associations by relaxing the criteria for co-occurrences of EC-Pfam annotations by looking for matches only at the 3-digit EC level. Indeed, we have observed several cases where true associations according to the InterPro training dataset were assigned confidence scores below the threshold value because they had too few (4-digit EC number) instances to provide sufficient support. Therefore, the above procedure was repeated using 3-digit EC numbers to give a 3-digit scoring scheme (with different weight factors

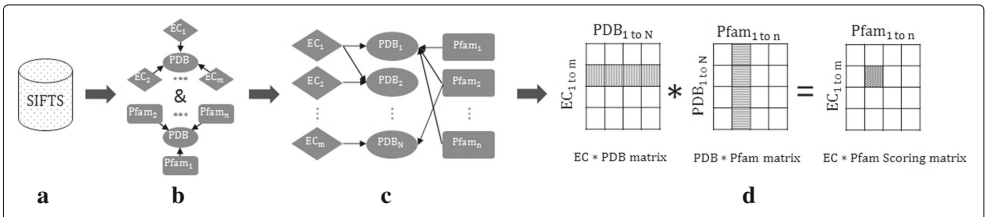


Fig. 2 A graphical illustration of calculating raw EC-Pfam association scores from existing SIFTS EC-CID and Pfam-CID associations

and a different score threshold). Then, any 4-digit EC-Pfam association below the 4-digit threshold, but consistent with a 3-digit EC-Pfam association above the 3-digit threshold, was added to the final list of accepted 4-digit EC-Pfam associations. It should be clarified that "consistent" means here that the 4-digit EC number is a descendant of the 3-digit EC number and that the Pfam domains are the same.

Hypergeometric distribution *p*-value analysis

While the above procedure provides a systematic way to infer EC-Pfam associations, we wished to estimate the statistical significance, and thus the degree of confidence, that might be attached to those predictions. More specifically, we wished to calculate the probability, or "*p*-value", that an EC number and a Pfam domain might be found to be associated simply by chance. For example, it is natural to suppose such associations can be predicted at random if *ec* or *d* are highly represented in the structure/sequence CIDs. In principle, in order to estimate the probability of getting our EC-Pfam associations by chance, one could generate random datasets by shuffling the relations between EC numbers and CIDs on the one hand, and between Pfam domains and CIDs on the other hand. However, this is quite impractical given the very large numbers of CIDs, EC numbers, and Pfam domains, and the complexity of the filtering procedure that would have to be repeated for each shuffled version of the dataset. Therefore, as in [17], we assume that a random association of CIDs to pairs of *ec* and *d* follows a hypergeometric distribution.

Letting N denote the total number of CIDs, N_d the number of CIDs related to the Pfam domain d, and N_{ec} the number of CIDs related to the EC number ec, the hypergeometric probability distribution is given by

$$p(X_{ec,d} \geqslant K_{ec,d}) = \frac{\sum_{i=K_{ec,d}}^{\min(N_d,N_{ec})} \binom{N_{ec}}{i}\binom{N-N_{ec}}{N_d-i}}{\binom{N}{N_d}}, \qquad (3)$$

where $p(X_{ec,d} \geqslant K_{ec,d})$ represents the probability of having a number $X_{ec,d}$ equal to or greater than the observed number $K_{ec,d}$ of CIDs associated with both d and ec. Traditionally, a p-value of less than 0.05 is taken to be statistically significant. However, because this test is applied to a large number of EC-Pfam associations, we apply a Bonferoni correction which takes into account the so-called family-wise error rate (FWER) [29]. We therefore consider any p-value less than $0.05/T$ as denoting a statistically significant inferred EC-Pfam association in a dataset, with T the total number of tested EC-Pfam associations for this dataset, In order to distinguish EC-Pfam associations using both confidence scores and p-values, we classify them into three classes, "Gold", "Silver", and "Bronze". An association is assigned to the Gold class if both its EC-Pfam score is greater than the determined threshold and all its p-values (in all datasets) are statistically significant. An association is labeled Silver if its score is above the threshold but one or more of its p-values is not statistically significant, or if its score is below the threshold (due to the 3-digit procedure, see "Exploiting the EC number hierarchy" section) but all its p-values are statistically significant. All other associations are labeled Bronze.

Results and discussion

Data source weights and score threshold

After clustering identical structures and sequences, and calculating raw association scores (Fig. 2), our merged dataset contains 6306 SIFTS, 18,917 SwissProt, 124,699 TrEMBL, and 141,990 UniRule candidate EC-Pfam associations, giving a total of 262,571 distinct EC-Pfam associations to draw from Table 1. In our ROC-based training procedure, the best AUC value of 0.985 was obtained with weights $w_{SIFTS} = 0.1$, $w_{SwissProt} = 1.0$, $w_{TrEMBL} = 0.1$, and $w_{UniRule} = 0.6$. These weights clearly give greater importance to the candidate associations in SwissProt and UniRule, respectively, compared to those in SIFTS and TrEMBL.

The optimal score threshold was determined according to the F-measure training procedure using our training dataset ("Defining a confidence score threshold" section). This gave a score threshold of 0.04 for a maximum F-Measure of 0.9476. Applying this threshold to the test dataset yielded a comparable F-measure of 0.935, and precision and recall values of 0.99 and 0.893, respectively.

Global analysis of inferred EC-Pfam associations

The results of the ECDomainMiner approach are summarized in Table 1. This table shows the numbers of 4-digit EC-Pfam associations along with the numbers of distinct EC numbers and Pfam entries involved in those associations for the four sources and the merged datasets before filtering.

After applying the 0.04 score threshold, the number of EC-Pfam associations falls to 8,256 with an overlap of about 96% of InterPro reference associations. Using the relaxed 3-digit association approach ("Exploiting the EC number hierarchy" section), the final ECDomainMiner dataset contains 20,728 EC-Pfam associations that overlap by 99.3% the InterPro reference dataset. These numbers show that our approach efficiently retrieves the InterPro reference EC-Pfam associations, including a small percentage (about 3.3%) that have a low confidence score.

Table 1 also shows that our ECDomainMiner set of EC-Pfam associations represents a 13.7 fold-increase (20,728/1515) in EC-Pfam associations with respect to InterPro. Moreover, the list of EC-Pfam associations produced by ECDomainMiner contains 6.4 times more EC

Table 1 Statistics on the source datasets and calculated EC-Pfam associations

	Dataset	EC-Pfam associations	Distinct 4-digit EC numbers	Distinct Pfam entries
Source	SIFTS	6306	2648	2611
Datasets	SwissProt	18,917	4013	3101
	TrEMBL	124,699	3751	5703
	UniRule	141,990	1020	2907
	Merged	262,571	4648	6639
Reference	InterPro	1515	688	1284
ECDomainMiner	With CS above threshold	*8256*	*3701*	*3022*
Results	(Overlap with InterPro)	(*1461*)	(*688*)	(*1245*)
	Including low CS	*20,728*	*4455*	*3613*
	(Overlap with InterPro)	(*1498*)	(*688*)	(*1273*)

CS is the Confidence Score
All italicized entries are calculated by ECDomainMiner

numbers and 2.8 times more Pfam domains than InterPro. Figure 3 shows how this increase in EC-Pfam associations distributes across the 6 top-level branches (i.e. 1-digit codes) of the EC classification.

The greatest ECDomainMiner scale-up factor occurs for associations involving the oxydoreductases (EC branch 1). The smaller scale-up factor observed for Pfam domains (2.8 versus 6.4 for EC numbers) can be explained by the fact that not all Pfam domains display an enzymatic activity. Thus there is a natural limit in the coverage of Pfam database by our EC-Pfam associations, whereas there is no such limit for the coverage of EC numbers. Combining the confidence scores with the calculated p-values as described in "Hypergeometric distribution p-value analysis" section gave 4552 Gold associations (having scores above the threshold and significant p-values in all source datasets), 11,426 Silver associations (with either scores above the threshold and one or more non-significant p-values, or with a score below the threshold but with significant p-values in all source datasets), and 4201 Bronze associations.

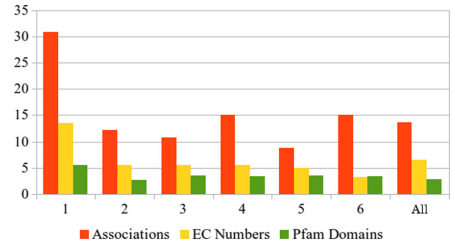


Fig. 3 Scale-up factors for ECDomainMiner compared with InterPro. Ratios between the numbers in ECDomainMiner and in Interpro have been calculated for associations (*red*), EC numbers (*yellow*), and Pfam domains (*green*) after dividing the dataset according to each EC branch represented in the associations (1 to 6) and for all the dataset (All). 1: oxydoreductases; 2: transferases; 3: hydrolases; 4: lyases; 5: isomerases; 6: ligases

Comparison with dcGO

In order to compare ECDomainMiner with the dcGO approach [17], we extracted SCOP2EC associations from the Domain2EC file available from the dcGO database (http://supfam.org/SUPERFAMILY/dcGO). The Domain2EC file includes 7249 associations with 4-digit EC numbers, of which 3774 are related to SCOP "Families" and 3475 to SCOP "SuperFamilies". Because InterPro only tabulates SCOP family domains, we limited our comparison to the set of 3774 SCOP2EC family associations. The SCOP families were mapped to Pfam families according to InterPro mapping files in order to generate a set of 2500 "Pfam2EC" associations (i.e. EC-Pfam associations which may be deduced directly from the SCOP2EC data). This set (shown as set a in Fig. 4) was compared with the set of all 262,571 merged EC-Pfam associations found by ECDomainMiner (set b in Fig. 4).

This comparison showed that a total of 480 Pfam2EC associations from SCOP2EC are not present in our merged dataset. The remaining 2020 Pfam2EC associations were then compared with the 20,728 associations calculated by ECDomainMiner (set c in Fig. 4). This comparison (the intersection of sets a and c) produced a total of 1892 EC-Pfam associations which are common to Pfam2EC and ECDomainMiner, indicating that ECDomainMiner agrees with 75.7% of the Pfam2EC associations from dcGO. Furthemore, this comparison also shows that ECDomainMiner result set contains 18,836 ($20,728 - 1,892$) additional EC-Pfam associations that are not available through dcGO.

Selecting plausible associations in multi-domain proteins

Because ECDomainMiner finds many new EC-Pfam associations, it is important to ask to what extent it also

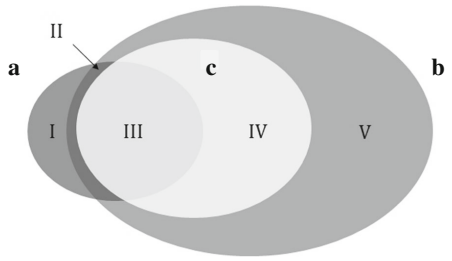


Fig. 4 Venn diagram showing the intersection between **a** Pfam2EC (2500 associations) from dcGO, **b** All-Merged (262,571 associations), and **c** ECDomainMiner (20,728 associations). Region I (480 associations) is the portion of (**a**) for which there is no data in any of our four source datasets. Region II (128 associations) is the portion of (**a**) that exists in (**b**) but is not retained in ECDomainMiner (**c**). Region III (1892 associations) is the overlap between (**a**) and (**c**). Region IV (18,836 associations) is the portion of ECDomainMiner associations that are not available from SCOP2EC. Region V (241,363 associations) is the rest of the merged set of EC-Pfam source associations that are absent from (**a**) and not retained as Gold, Silver, or Bronze associations by ECDomainMiner

might produce false associations. Firstly, we recall that ECDomainMiner eliminated more than 92% (241,843 out of 262,571) of low-scoring associations from the merged source dataset. This suggests that most of the eliminated associations involve Pfam domains that are not catalytically active. Indeed, if a Pfam domain is not regularly associated with protein chains or sequences having an enzymatic activity, the ECDomainMiner score for that domain is very low, and hence no EC number is assigned to that domain. This applies in particular to accessory domains that can co-occur with various catalytic domains in multi-domain proteins. A good example of such an accessory domain is PF00188 (the CAP protein family) which is a part of 216 different architectures. Among these architectures, there are 3 and 5 different architectures, which additionally contain PF00112 (Peptidase C1 domain) and PF00069 (Protein kinase domain), respectively. According to Pfam website, PF00188 is catalytically inactive but PF00112 and PF00069 are active. In fact, ECDomainMiner assigns PF00112 to 26 different EC numbers with a majority of EC 3.4.22 (Cysteine endopeptidases), and PF00069 to 28 different EC numbers that all start with 2.7 (Transferring phosphorus-containing groups). However, ECDomainMiner does not assign PF00188 to any EC number. This is because a large number of protein chains and sequences containing either PF00112 or PF00069 and associated with the above-mentioned EC activities, do not contain PF00188. In other words the catalytic activities of PF00112 and PF00069 are not strictly dependent on the presence of PF00188. Moreover, the SIFTS and UniProt databases indicate that PF00188 is associated with 43 different PDB chains and 5197 different protein sequences. However, none of those PDB chains are associated with an EC number in SIFTS and only 31 protein sequences (24 in TrEMBL and 7 in UniRule) are associated with at least one 4-digit EC number. Consequently, the association score of PF00188 with any EC number is zero for both the SIFTS and SwissProt datasets and is quite low (less than 0.02) for both the TrEMBL and UniRule datasets. Thus, the confidence scores of all of the associations involving PF00188 in ECDomainMiner are lower than our threshold of 0.04, and so these candidate associations are filtered out. This mechanism explains how an accessory domain is not assigned to an EC number by ECDomainMiner, and suggests that most of the retained associations are proper candidates for domain functional annotation.

Single and multiple EC-Pfam associations

Exploring the ECDomainMiner results readily reveals that a given EC number or Pfam domain can be involved in one or more distinct EC-Pfam associations. Figure 5 shows the relative distribution of EC numbers and Pfam domains according to the number of EC-Pfam associations they are involved in. This figure shows that 1576 out of 4393 EC numbers and 1280 out of 3542 Pfam domains are involved in a single EC-Pfam association.

Although this represents rather high proportions of the total number of EC numbers and Pfam domains in ECDomainMiner (35.9 and 36.1%, respectively), the intersection of the concerned EC-Pfam single associations yields a list of only 97 one-to-one EC-Pfam associations, of which 62, 34, and 1 are Gold, Silver, and Bronze associations, respectively. Comparison with the InterPro reference dataset reveals that two thirds (65) of these one-to-one associations are novel compared to InterPro. Interestingly, we confirmed in our source datasets that all of these associations involve single-domain proteins. Thus, these unambiguous associations constitute the most reliable novel associations calculated by ECDomainMiner.

The complete list of one-to-one EC-Pfam associations found by ECDomainMiner may be downloaded from the ECDomainMiner web site. Interestingly 14 of these associations (8 Gold, of which 2 match InterPro reference associations, and 6 Silver) concern "DUF" (domain of unknown function) or "UPF" (uncharacterised protein family) Pfam entries. These are listed in part (A) of Table 2 in order of decreasing confidence score.

These examples demonstrate that ECDomainMiner can be used to enrich domain annotation. Visual inspection of the one-to-one EC-Pfam associations indicates that about one quarter of them (23) could have been retrieved simply by comparing the names associated with the EC number and the Pfam identifier, which are nearly identical (see example in Table 2b). However, only 10 of

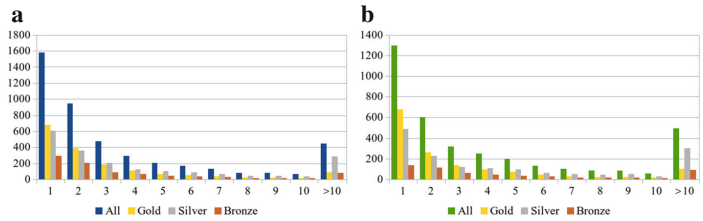


Fig. 5 Distribution of EC numbers (**a**) and Pfam domains (**b**) in multiple associations. Numbers (1 to 10 and >10) represent the arity of the association in which a given EC number, respectively Pfam domain, is involved. In addition, for each arity, the normalized number of Gold, Silver, and Bronze associations is plotted. It can be observed that for arities equal to or greater than 4, the proportion of Silver associations is always the highest but significant numbers of Gold associations remain present even for high arity numbers

these associations were in fact already known in InterPro. Clearly, minor and unpredictable spelling differences impair the automatic retrieval of such similar but non-identical EC and Pfam names. Nonetheless, while these associations could be found by clever text matching, we emphasise that ECDomainMiner's confidence scores and *p*-values provide a level of support for each association that would be very difficult to obtain from text mining alone.

The multi-partner associations calculated by ECDomainMiner provide many more complex EC-Pfam associations. As a first analysis of such multiple associations, we looked for obligate pairs or tuples of Pfam domains that are always associated with a given EC number. Briefly, for any pair of Pfam domains, (d_1, d_2), associated with the same EC number, *ec*, (i) we reject those pairs for which at least one *ec*-annotated CID (in any source dataset) occurs in relation with d_1 and not d_2 or with d_2 and not d_1, (ii) for all other pairs we calculate for each source dataset the ratio of the number of *ec*-annotated CIDs related to d_1 and d_2, to the total number of *ec*-annotated CIDs. A support ratio of 1 means that all CIDs annotated with *ec* in a dataset are also related to d_1 and d_2. A similar algorithm was used for triplets and quadruples of Pfam domains. For a support ratio of 1 in at least one source dataset, we found 907, 191 and 47 obligate associations between an EC number and a pair, a triplet or a quadruplet of Pfam domains. These associations are available from the ECDomainMiner website. Two examples are given in part (C) of Table 2.

Interestingly, filtering the names of the Pfam domains with the expressions "N-terminal" and "C-terminal" yielded 58 obligate pairs containing both a N-terminal and a C-terminal domain of the same function. This indicates that our approach is finding enzymes in which the catalytic function is provided by the interface between two consecutive Pfam domains. Only 4 of these obligate pair associations are currently documented in InterPro.

Annotating PDB chains with EC numbers

Our analysis of the December 2015 release of the SIFTS database reveals that about 45% of PDB entries lack an EC number annotation. Of course, such an annotation is not expected to be present in all PDB entries because not all proteins have enzymatic activity. Nonetheless, it is interesting to use ECDomainMiner to analyse the number of PDB entries that contain Pfam domains which are present in EC-Pfam associations. Table 3 shows that a total of 58,722 PDB chains lacking EC annotations in SIFTS include at least one of the 3542 Pfam domains present in ECDomainMiner.

Overall, we calculated that these chains map to a total of 24,995 PDB entries that could benefit from the additional annotations inferred by ECDomainMiner. For those chains lacking EC annotations, ECDomainMiner finds Gold, Silver, and Bronze EC-Pfam associations for 41,246, 44,406 and 34,820 PDB chains, respectively. In particular, 1334 PDB chains could benefit from our dataset of 97 non ambiguous one-to-one EC-Pfam associations.

The ECDomainMiner web server

The ECDomainMiner web server may be queried by EC number or Pfam domain. Thus, if one wishes to search for associations for a protein chain that currently lacks any EC annotation in the PDB (e.g. chain 2q7xA), one first needs to retrieve from the PDB the Pfam domain(s) that it contains (in this example, PF01933). Then, querying the ECDomainMiner server with each Pfam domain identifier will show the associated EC numbers (in this example, 2.7.8.28), along with the associated filtering scores and quality classes. In this example, ECDomainMiner finds a Gold quality association between PF01933, present in PDB chain 2q7xA, and EC number 2.7.8.28 (2-phospho-L-lactate transferase) which consequently can be associated with PDB entry 2q7x. Interestingly, PDB entry 2q7x is described as a putative phospho transferase from *streptococcus pneumoniae* tigr4, which is consistent with

Table 2 (A) Fourteen one-to-one EC-Pfam associations found by ECDomainMiner and involving domains of unknown function, (B) an example of one-to-one EC-Pfam association with very similar EC and Pfam descriptions, and (C) two examples of obligate Pfam pairs associated with an EC number

	EC	Pfam	Score	EC name	Pfam name	Quality	PDBs (SIFTS)
A	2.7.8.28	PF01933	0.972	2-phospho-L-lactate transferase	Uncharacterised protein family UPF0052	Gold	9/0/11
	4.1.99.5	PF11266	0.944	Aldehyde oxygenase (deformylating)	Protein of unknown function DUF3066	Gold	18/0/0
	2.1.1.286	PF11968	0.889	25S rRNA (adenine(2142)-N(1))-methyltransferase	Putative methyltransferase DUF3321	Gold	0/0/0
	1.13.99.1	PF05153	0.667	Inositol oxygenase	Family of unknown function DUF706	Gold	4/0/0
	2.4.1.155	PF15027	0.611	Alpha-1,6-mannosyl-glycoprotein 6-beta-N-acetylglucosaminyltransferase	Domain of unknown function DUF4525	Gold	0/0/0
	4.2.3.130	PF10776	0.611	Tetraprenyl-beta-curcumene synthase	Protein of unknown function DUF2600	Gold	0/0/0
	2.3.1.78	PF07786	0.609	Heparan-alpha-glucosaminide N-acetyltransferase	Protein of unknown function DUF1624	Gold	0/0/0
	3.1.4.45	PF09992	0.584	N-acetylglucosamine-1-phosphodiester alpha-N-acetylglucosaminidase	Predicted periplasmic protein DUF2233	Gold	0/0/1
	1.13.12.20	PF08592	0.556	Noranthrone monooxygenase	Domain of unknown function DUF1772	Gold	0/0/0
	2.1.1.312	PF11312	0.556	25S rRNA (uracil(2843)-N(3))-methyltransferase.	Protein of unknown function DUF3115	Gold	0/0/0
	2.1.1.313	PF10354	0.556	25S rRNA (uracil(2634)-N(3))-methyltransferase	Domain of unknown function DUF2431	Gold	0/0/0
	2.5.1.128	PF01861	0.556	N4-bis(aminopropyl) spermidine synthase	Protein of unknown function DUF43	Gold	0/0/1
	5.2.1.14	PF13225	0.556	Beta-carotene isomerase	Domain of unknown function DUF4033	Gold	0/0/0
	1.14.99.29	PF04248	0.333	Deoxyhypusine monooxygenase	Domain of unknown function DUF427	Silver	0/0/5
B	6.3.2.25	PF03133	0.610	Tubulin–tyrosine ligase	Tubulin-tyrosine ligase family	Gold	0/2/21
C	2.7.1.30	PF00370	0.847	Glycerol kinase	FGGY family of carbohydrate kinases, N-terminal domain	Gold	85/32/9
		PF02782	0.828		FGGY family of carbohydrate kinases, C-terminal domain	Gold	85/32/7
	6.3.4.23	PF06973	0.997	Formate-phosphoribosyl-amino-imidazol	DUF1297	Gold	16/3/0
		PF06849	0.997	carboxamide ligase	DUF1246	Gold	16/3/0

The 'PDBs (SIFTS)' column contains 3 counts of PDB chains containing the mentioned Pfam domain and having either the same EC annotation in SIFTS as calculated by ECDomainMiner (first position), or different EC annotations between SIFTS and ECDomainMiner (second position), or no EC annotations in SIFTS (third position). Complete lists of PDB identifiers may be retrieved from the ECDomainMiner web server

Table 3 The numbers of PDB protein chains that could be annotated by ECDomainMiner associations

Association type	ECDM associations concerned	PDB chains concerned
Any	14,573	58,722
Gold	3591	41,246
Silver	7796	44,406
Bronze	3186	34,820
One-to-One	44	1334

the enzymatic activity found by ECDomainMiner, and which could not be deduced from the Pfam domain name (UPF0052).

Conclusion

We have presented a content-based filtering approach for associating EC numbers with Pfam domains. This approach has been shown to be able to infer a total of 20,728 non-redundant EC-Pfam associations, which corresponds to over 13 times as many EC-Pfam associations as currently exist in InterPro. Furthermore, thanks

to our calculated *p*-values, we have assigned an intuitive quality rating (Gold, Silver, or Bronze) to each EC-Pfam association found. These calculated associations are publicly available on the ECDomainMiner web site. We anticipate that our content-based filtering approach may be applied to other annotation vocabularies or ontologies, and we are currently working to extend our approach to discover new GO-Pfam annotations.

We believe that enriching protein chain annotations will facilitate a better understanding and exploitation of structure-function relationships at the domain level. While many of the associations calculated by ECDomainMiner are consistent with those recently made available by the domain-centric dcGO approach for finding EC-SCOP associations, the ECDomainMiner results set contains many more associations than dcGO. Indeed, the ECDomainMiner result set contains 18,836 EC-Pfam which are not available in dcGO. Our analysis of the simple one-to-one associations found by ECDomainMiner shows that several DUF or UPF entries in Pfam may be assigned functions from the EC classification, and that obvious inconsistencies in the annotation texts may easily be corrected or unified. However, only a relatively small number (less than 0.5%) of EC-Pfam associations in our result set are simple one-to-one associations, indicating that there exist a large number of many-to-many relations between EC numbers and Pfam domains. Further analyses of these complex associations using graph database and machine-learning techniques could reveal many more hidden protein structure-function relationships.

Abbreviations

AN: Accession numbers; AUC: Area under the curve; CID: Cluster unique IDentifier; CS: Confidence score; dcGO: Domain centric Gene Ontology database; DUF: Domain of unknown function; EC: Enzyme commission; FN: False negative; FP: False positive; GO: Gene ontology; NMR: Nuclear magnetic resonnance; P: Precision; PDB: Protein data bank; Pfam: Protein family database; R: Recall; ROC: Receiver operator characteristics; SCOP: Structural classification of proteins; SCOPEC: a database of catalytic domains; SIFTS: Structure integration with function, taxonomy and sequence; TN: True negative; TP: True positive; TrEMBL: Translated sequences from the European molecular biology laboratory bank; UniProtKB: UniProt knowledge base; UPF: Uncharacterized protein family

Acknowledgements

Not applicable.

Funding

This project is funded by the Agence Nationale de la Recherche (grant reference ANR-11-MONU-006-02), Inria and the Region Lorraine.

Availability of data and materials

The ECDomainMiner results can be accessed with a web browser at http://ecdm.loria.fr/. The ECDomainMiner database will be be updated bi-annually.

Authors' contributions

SZA designed the study and was involved in data processing and management, analysis and testing. MDD was involved in biological interpretation. All authors discussed the results and drafted together the manuscript. All authors read and approved the final manuscript.

Competing interests

The authors declare that they have no competing interests.

Consent for publication

Not applicable.

Ethics approval and consent to participate

Not applicable.

Author details

[1]Université de Lorraine, LORIA, UMR 7503, 54506 Vandœuvre-lès-Nancy, France. [2]CNRS, LORIA, UMR 7503, 54506 Vandœuvre-lès-Nancy, France. [3]Inria Nancy Grand-Est, 54600 Villers-lès-Nancy, France.

Received: 28 August 2016 Accepted: 1 February 2017
Published online: 13 February 2017

References

1. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer ELL, Tate J, Punta M. Pfam: the protein families database. Nucleic Acids Res. 2014;42(D1):222–30.
2. Berg JM, Tymoczko JL, Stryer L. Protein structure and function. New York: WH Freeman; 2002.
3. Chothia C, Lesk AM. The relation between the divergence of sequence and structure in proteins. EMBO J. 1986;5(4):823.
4. Martin ACR, Orengo CA, Hutchinson EG, Jones S, Karmirantzou M, Laskowski RA, Mitchell JBO, Taroni C, Thornton JM. Protein folds and functions. Structure. 1998;6(7):875–84.
5. Bernstein FC, Koetzle TF, Williams GJB, Meyer EF, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The protein data bank. Eur J Biochem. 1977;80(2):319–24.
6. Gutmanas A, Alhroub Y, Battle GM, Berrisford JM, Bochet E, Conroy MJ, Dana JM, Montecelo MAF, van Ginkel G, Gore SP, Haslam P, Hatherley R, Hendrickx PMS, Hirshberg M, Lagerstedt I, Mir S, Mukhopadhyay A, Oldfield TJ, Patwardhan A, Rinaldi L, Sahni G, Sanz-García E, Sen S, Slowley RA, Velankar S, Wainwright ME, Kleywegt GJ. PDBe: protein data bank in europe. Nucleic Acids Res. 2014;42(D1):285–91.
7. Webb EC, et al. Enzyme nomenclature 1992. recommendations of the nomenclature committee of the international union of biochemistry and molecular biology on the nomenclature and classification of enzymes, Ed. 6. San Diego: Academic Press; 1992.
8. Reichert J, Jabs A, Slickers P, Sühnel J. The IMB Jena image library of biological macromolecules. Nucleic Acids Res. 2000;28(1):246–9.
9. de Beer TAP, Berka K, Thornton JM, Laskowski RA. PDBsum additions. Nucleic Acids Res. 2014;42(D1):292–6.
10. Laskowski RA. PDBsum: summaries and analyses of PDB structures. Nucleic Acids Res. 2001;29(1):221–2.
11. Martin ACR. PDBSprotEC: a web-accessible database linking PDB chains to EC numbers via SwissProt. Bioinformatics. 2004;20(6):986–8.
12. Velankar S, Dana JM, Jacobsen J, van Ginkel G, Gane PJ, Luo J, Oldfield TJ, O'Donovan C, Martin MJ, Kleywegt GJ. SIFTS: structure integration with function, taxonomy and sequences resource. Nucleic Acids Res. 2013;41(D1):483–9.
13. The UniProt Consortium. The universal protein resource (UniProt) in 2010. Nucleic Acids Res. 2010;38(suppl 1):142–8.
14. Fleischmann A, Darsow M, Degtyarenko K, Fleischmann W, Boyce S, Axelsen KB, Bairoch A, Schomburg D, Tipton KF, Apweiler R. IntEnz, the integrated relational enzyme database. Nucleic Acids Res. 2004;32 (suppl 1):434–7.
15. George RA, Spriggs RV, Thornton JM, Al-Lazikani B, Swindells MB. SCOPEC: a database of protein catalytic domains. Bioinformatics. 2004;20(suppl 1):130–6.
16. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol. 1995;247(4):536–40.
17. Fang H, Gough J. dcGO: database of domain-centric ontologies on functions, phenotypes, diseases and more. Nucleic Acids Res. 2013;41(D1):536–44.

18. Alborzi SZ, Devignes MD, Ritchie DW. EC-PSI: associating enzyme commission numbers with Pfam domains. In: Proceedings of JOBIM; 2015. doi:10.1101/022343.
19. Hanani U, Shapira B, Shoval P. Information filtering: Overview of issues, research and systems. User Model User-Adap Inter. 2001;11(3):203–59.
20. Ricci F, Rokach L, Shapira B. Introduction to recommender systems handbook. NewYork: Springer; 2011.
21. Breese JS, Heckerman D, Kadie C. Empirical analysis of predictive algorithms for collaborative filtering. In: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence. San Francisco: Morgan Kaufmann; 1998. p. 43–52.
22. Koren Y, Bell R. Advances in collaborative filtering on recommender systems handbook. New York: Springer; 2015. p. 77–118.
23. Basu C, Hirsh H, Cohen W, et al. Recommendation as classification: Using social and content-based information in recommendation. In: Proceedings of IAAI. Palo Alto: AAAI Press; 1998. p. 714–20.
24. Huang L, Li F, Sheng J, Xia X, Ma J, Zhan M, Wong ST. Drugcomboranker: drug combination discovery based on target network analysis. Bioinformatics. 2014;30(12):228–36.
25. Mitchell A, Chang HY, Daugherty L, Fraser M, Hunter S, Lopez R, McAnulla C, McMenamin C, Nuka G, Pesseat S, Sangrador-Vegas A, Scheremetjew M, Rato C, Yong SY, Bateman A, Punta M, Attwood TK, Sigrist CJA, Redaschi N, Rivoire C, Xenarios I, Kahn D, Guyot D, Bork P, Letunic I, Gough J, Oates M, Haft D, Huang H, Natale DA, Wu CH, Orengo C, Sillitoe I, Mi H, Thomas PD, Finn RD. The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Res. 2015;43(D1):213–21.
26. Pedruzzi I, Rivoire C, Auchincloss AH, Coudert E, Keller G, De Castro E, Baratin D, Cuche BA, Bougueleret L, Poux S, et al. Hamap in 2013, new developments in the protein family classification and annotation system. Nucleic Acids Res. 2013;41(D1):584–9.
27. Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics. 2007;23(10):1282–8.
28. Fawcett T. An introduction to ROC analysis. Pattern Recogn Lett. 2006;27(8):861–74.
29. Cui X, Churchill GA, et al. Statistical tests for differential expression in cDNA microarray experiments. Genome Biol. 2003;4(4):210.

C.1.3 Associating Gene Ontology Terms with Pfam Protein Domains

With the growing number of three-dimensional protein structures in the protein data bank (PDB), there is a need to annotate these structures at the domain level in order to relate protein structure to protein function. Thanks to the SIFTS database, many PDB chains are now cross-referenced with Pfam domains and Gene ontology (GO) terms. However, these annotations do not include any explicit relationship between individual Pfam domains and GO terms. Therefore, creating a direct mapping between GO terms and Pfam domains will provide a new and more detailed level of protein structure annotation. This article presents a novel content-based filtering method called GODM that can automatically infer associations between GO terms and Pfam domains directly from existing GO-chain/Pfam-chain associations from the SIFTS database and GO-sequence/Pfam-sequence associations from the UniProt databases. Overall, GODM finds a total of 20,318 non-redundant GO-Pfam associations with a F-measure of 0.98 with respect to the InterPro database, which is treated here as a "Gold Standard". These associations could be used to annotate thousands of PDB chains or protein sequences for which their domain composition is known but which currently lack any GO annotation.

Associating Gene Ontology Terms with Pfam Protein Domains

Seyed Ziaeddin Alborzi[1,3], Marie-Dominique Devignes[2], and David W. Ritchie[3( )]

[1] Université de Lorraine, LORIA, UMR 7503, 54506 Vandœuvre-lès-Nancy, France
[2] CNRS, LORIA, UMR 7503, 54506 Vandœuvre-lès-Nancy, France
marie-domonique.devignes@loria.fr
[3] Inria Nancy Grand-Est, 54600 Villers-lès-Nancy, France
{seyed-ziaeddin.alborzi,dave.ritchie}@inria.fr

Abstract. With the growing number of three-dimensional protein structures in the protein data bank (PDB), there is a need to annotate these structures at the domain level in order to relate protein structure to protein function. Thanks to the SIFTS database, many PDB chains are now cross-referenced with Pfam domains and Gene ontology (GO) terms. However, these annotations do not include any explicit relationship between individual Pfam domains and GO terms. Therefore, creating a direct mapping between GO terms and Pfam domains will provide a new and more detailed level of protein structure annotation. This article presents a novel content-based filtering method called GODM that can automatically infer associations between GO terms and Pfam domains directly from existing GO-chain/Pfam-chain associations from the SIFTS database and GO-sequence/Pfam-sequence associations from the UniProt databases. Overall, GODM finds a total of 20,318 non-redundant GO-Pfam associations with a F-measure of 0.98 with respect to the InterPro database, which is treated here as a "Gold Standard". These associations could be used to annotate thousands of PDB chains or protein sequences for which their domain composition is known but which currently lack any GO annotation. The GODM database is publicly available at http://godm.loria.fr/.

Keywords: Protein structure · Protein function · Gene Ontology · Content-based filtering

1 Introduction

Proteins carry out many important biological functions. At the molecular level, these functions are often performed by highly conserved regions called "domains". Currently, the Pfam database is one of the most widely used sequence-based classifications of protein domains and domain families [1]. Protein domains may also be considered as building blocks which are combined in different ways in order to endow different proteins with different functions.

I. Rojas and F. Ortuño (Eds.): IWBBIO 2017, Part II, LNBI 10209, pp. 127–138, 2017.
DOI: 10.1007/978-3-319-56154-7_13

A given Pfam domain might exist in several different proteins. It is widely accepted that protein domains often correspond to distinct and stable three-dimensional (3D) structures, and that there is often a close relationship between protein structure and protein function [2]. The Protein Data Bank (PDB) [3,4] contains more than 107,000 3D structures, that have been determined by X-ray crystallography or NMR spectroscopy. As well as sequence-based and structure-based classifications, proteins may also be classified according to their function. For example, the Gene Ontology (GO) [5] organizes a controlled vocabulary describing the biological process (BP), molecular function (MF), and cellular component (CC) aspects of gene annotation. It provides an ontology of defined terms to unify the representation of the gene and protein roles in cells. The GO vocabulary is structured as a rooted Directed Acyclic Graph (rDAG) in which GO terms are nodes connected by different hierarchical relations. Each GO term within the gene ontology has a term name, a distinct alphanumeric identifier, and a namespace indicating to which ontology it belongs.

Although the GO is very useful, it does not generally provide a direct relationship between biological function and a (sequence-based) Pfam domain. Figure 1 illustrates the different kinds of relationships that can occur when considering GO-protein annotations at the domain level. Except for simple single-domain proteins where the mapping is obvious, it is generally not possible to compare and classify structure-function relationships at the domain level. An interesting exception is the dcGO database which provides multiple ontological annotations (Gene Ontology: GO, EC, pathways, phenotype, anatomy and disease ontologies) for protein domains [6]. In dcGO, an association between an ontology term and a domain is inferred from the principle that if a term tends to be attached to proteins in UniProtKB that contain a certain domain, then the term should be associated with that domain. For each Pfam domain, dcGO compares the number of Uniprot sequences containing that domain and annotated with a certain GO term to what could be obtained if association was random. The statistical significance of the association is then assessed using a hypergeometric distribution, followed by multiple hypotheses testing in terms of false discovery rate. Only significant associations are retained in the dcGO database.

Nonetheless, we found that there are several GO-Pfam associations from manually curated data sources (e.g. InterPro) which are not present in dcGO. Moreover, based on our previous ECDomainMiner approach [7,8] to discover associations between EC numbers and protein domains, we found that there are many reliable EC-Pfam associations which are not covered by dcGO. Furthermore, there are thousands of protein structures in the PDB which lack GO annotations. If there is a direct association between protein domains and GO terms, these structures can be annotated through their associated domains. Based on our analysis, we estimated that dcGO associations can only annotate 43% of the unannotated PDB structures. Therefore, we were motivated to develop a more systematic approach, which we call "GODM" ("GO Domain Miner"), with the aim of discovering a much larger set of GO-domain associations than dcGO.

GODM uses a "recommender-based" approach for finding direct associations between GO terms and Pfam domains. We recently developed a similar recommender-based approach called "ECDomainMiner" for assigning enzyme classification (EC) numbers to Pfam domains [8]. Thus, the GODM approach described here represents a natural extension of our previously developed ECDomainMiner approach. Recommender systems are a subclass of information filtering system [9,10] which seek to predict a list of items that might be of interest to an on-line customer, and are divided into two main types. Collaborative filtering approaches make associations by calculating the similarity between activities of users [11,12]. In contrast, content-based filters predict associations between user profiles and description of items by identifying common attributes [10,13]. Here, we use content-based filtering to associate GO terms with Pfam domains from existing GO-chain and Pfam-chain associations from SIFTS [14], and GO-sequence and Pfam-sequence associations from SwissProt and TrEMBL. As well has handling simple one-to-one associations as in dcGO (Fig. 1 part A), GODM can also resolve cases where multiple GO terms are associated with multi-domain chains (Fig. 1 parts B, C, and D).

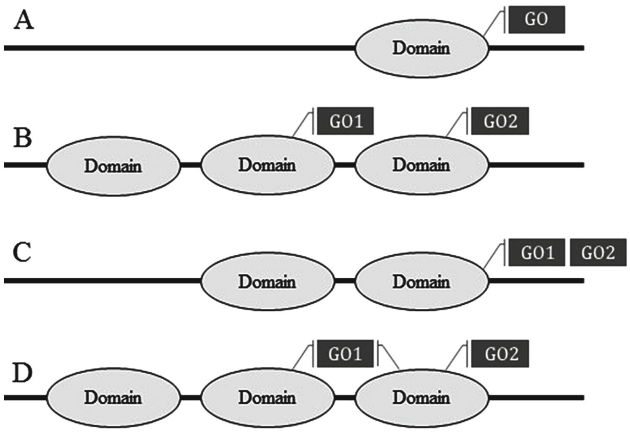


Fig. 1. A graphical representation of different situations of GO-Domain association in a protein sequence or structure.

While SwissProt and TrEMBL were originally developed separately, both databases have since been incorporated in the UniProt resource. SwissProt now represents a non-redundant, high quality, manually curated part of UniProt Knowledge Base (UniProtKB). In contrast, TrEMBL is an automatically annotated and unreviewed part of UniProtKB, and contains around 40 times more entries than SwissProt. In order to parameterise and evaluate our method, we use the InterPro database [15] which contains a large number of manually curated GO-Pfam associations. We assess the performance of our approach against a "Gold Standard" dataset derived from InterPro, and we compare our results with

the GO-Pfam associations available from the dcGO database. We also show how our database of more than 20,000 GO-Pfam associations for molecular function ontology can be exploited for automatic annotation purposes.

2 Methods

2.1 Data Preparation

Flat data files of SIFTS (July 2015), Uniprot (July 2015), and InterPro (version 53.0) were downloaded and parsed using in-house Python scripts. From the SIFTS data, associations between PDB chains and GO terms, and associations between PDB chains and Pfam domains were extracted in which each GO term is a leaf in the hierarchy of the Molecular Function ontology (GO-MF) and each Pfam refers either to a Pfam domain or a Pfam family (i.e. Pfam motifs and repeats were excluded). Associations between Uniprot sequence accession numbers (ANs) and GO terms from GO-MF, and AN-Pfam associations were then extracted from the SwissProt and TrEMBL sections of Uniprot to give two datasets of Swissprot associations and TrEMBL associations, respectively. Then, based on the evidence code of the GO term, associations in SwissProt and TrEMBL datasets were divided into two groups namely, associations for which GO terms were assigned in UniProtKB by manual curation, and Inferred from Electronic Annotation (IEA). These four datasets are subsequently called Swissprot, Swissprot-IEA, TrEMBL, and TrEMBL-IEA. Note that there were no evidence codes in the SIFTS.

To reduce bias due to the various numbers of identical sequences and sequences of chains in the five source datasets, all PDB chains and Uniprot sequences were grouped into clusters having identical sequences using the Uniref non-redundant cluster annotations [16]. Each cluster was assigned a unique identifier (CID), and the source GO-chain and GO-AN associations were then mapped to the corresponding cluster in order to make five sets of GO-CID associations. A similar mapping was applied to the source Pfam-chain and Pfam-AN associations to make five sets of Pfam-CID associations.

For the InterPro reference data, we extracted a total of 1,561 GO-Pfam associations in which each GO term is a leaf node of the molecular function ontology and each Pfam refers to either a Pfam domain or a Pfam family. These associations were considered to be “true” associations. However, for training and filtering purposes, we also needed some examples of “false” associations. We therefore selected a set of the lowest-scoring GO-Pfam associations with the same size as InterPro dataset from the other datasets. These associations have to belong to at least two out of five datasets with no intersection with InterPro dataset. Because these associations have very little support in the data, we consider them to be “false” associations. Then, we randomly divided the InterPro dataset and our calculated “false” associations into two “Training” and “Test” subsets of the same size (each having half of the “true” and “false” associations). These two subsets were used for training and evaluation purposes respectively.

In the rest of this article, we will refer to the InterPro dataset as our "Gold Standard" dataset.

2.2 Finding GO-Pfam Associations by Content-Based Filtering

For each of the five datasets, all GO-CID relations are encoded in a binary (GO × CID) matrix, where a 1 represents the presence of a GO annotation and a 0 represents no annotation. This matrix is then row-normalised such that each row has unit magnitude when considered as a vector. Similarly, all CID-Pfam relations are encoded in a second binary (CID × Pfam) matrix which is column-normalised. Consequently, calculating the product of the two normalised matrices corresponds to calculating a matrix of cosine similarity scores between the rows of the first matrix and the columns of the second matrix. Thus, the product matrix represents an array of raw GO-Pfam association scores. Because we wish to draw upon the relations from all five input datasets, we combine the five scores to give a single normalized confidence score (CS):

$$CS_{go,d} = \frac{\sum_i w_i S_i(go,d)}{\sum_i w_i} \tag{1}$$

where $i \in \{SIFTS,\ Swissprot,\ Swissprot\text{-}IEA,\ TrEMBL,\ TrEMBL\text{-}IEA\}$ enumerates the five datasets, w_i are weight factors, to be determined, and where an individual association score, $S_i(go,d)$ is set to zero whenever there is no data for a given go and d. In order to calculate the weight factors, we calculated Receiver-Operator-Characteristic (ROC) curves [17] using the true associations from the Interpro Training set and all other associations as background associations. The weights were varied from 0.0 to 1.0 in steps of 0.1, and for each combination, associations were scored and ranked, and area under the curve (AUC) was calculated. Finally, we selected the combination of weights that gave the best area under the curve (AUC) of the ROC curve.

2.3 Defining a Confidence Score Threshold

Having determined the best weight for each data source, we next wished to determine a threshold for the confidence score. We scored and ranked the members of the Training set of InterPro, and divided the ranked list into two subsets according to a threshold value that was varied from 0.0 to 1.0 in steps of 0.01. For each threshold value, we counted the number of true associations above the threshold, here called true positives (TPs), false associations above the threshold, false positives (FPs), false associations below the threshold, true negatives (TNs), and true associations below the threshold, false negatives (FNs). We then calculated the "F-measure" which is a harmonic mean of recall and precision using:

$$F = \frac{2 \times TP}{2 \times TP + FP + FN} \tag{2}$$

The score threshold that gave the best F-measure was confirmed by verifying that the F-measure calculated on the Test dataset is also very high. This threshold was thus selected as the best threshold to use for accepting predicted associations.

2.4 Hypergeometric Statistical Analysis

While the above procedure provides a systematic way to infer GO-Pfam associations, we wished to estimate the statistical significance, and thus the degree of confidence, that might be attached to those predictions. More specifically, we wished to calculate the probability, or "p-value", that a GO term, *go*, and a Pfam domain, *d*, could be found to be associated simply by chance. For example, it is natural to suppose such associations can be predicted at random if *go* or *d* are highly represented in the structure/sequence CIDs. In principle, in order to estimate the probability of getting our GO-Pfam associations by chance, one could generate random datasets by shuffling the relations between GO terms and CIDs on the one hand, and between Pfam domains and CIDs on the other hand. However, this is quite impractical given the very large numbers of CIDs, GO terms, and Pfam domains, and the complexity of the filtering procedure that would have to be repeated for each shuffled version of the dataset. Therefore, following [6], we assume that within each dataset (SIFTS, Swissprot, Swissprot-IEA, TrEMBL, or TrEMBL-IEA), the random hypothesis for the (*go*, *d*) association is represented by the hypergeometric distribution of the expected number of CIDs associated with both *go* and *d*.

Letting N denote the total number of CIDs, N_d the number of CIDs related to the Pfam domain d, and N_{go} the number of CIDs related to the GO term *go*, the hypergeometric probability distribution is given by

$$p(X_{go,d} \geqslant K_{go,d}) = \frac{\sum_{i=K_{go,d}}^{\min(N_d,N_{go})} \binom{N_{go}}{i}\binom{N-N_{go}}{N_d-i}}{\binom{N}{N_d}}, \tag{3}$$

where $p(X_{go,d} \geqslant K_{go,d})$ represents, in each dataset, the probability of having a number $X_{go,d}$ equal to or greater than the observed number $K_{go,d}$ of CIDs associated with both d and *go*. Traditionally, a p-value of less than 0.05 is taken to be statistically significant. However, because this test is applied to a large number of GO-Pfam associations, we apply a Bonferoni correction which takes into account the so-called family-wise error rate (FWER) [18]. We therefore consider any p-value less than $0.05/T$ as denoting a statistically significant inferred GO-Pfam association in a dataset, with T the total number of tested GO-Pfam associations for that dataset.

2.5 Gold, Silver, and Bronze Associations

In order to differentiate associations based on their quality and reliability, our method categorizes associations into three classes of "Gold", "Silver", and "Bronze" using their calculated similarity scores and p-values. An association belongs to the Gold class if all its available p-values are statistically significant. The Silver class consists of associations for which the number of statistically significant p-values among the five datasets is greater than or equal to the number of statistically insignificant p-values (e.g. GO-Pfam is a Silver associations if its

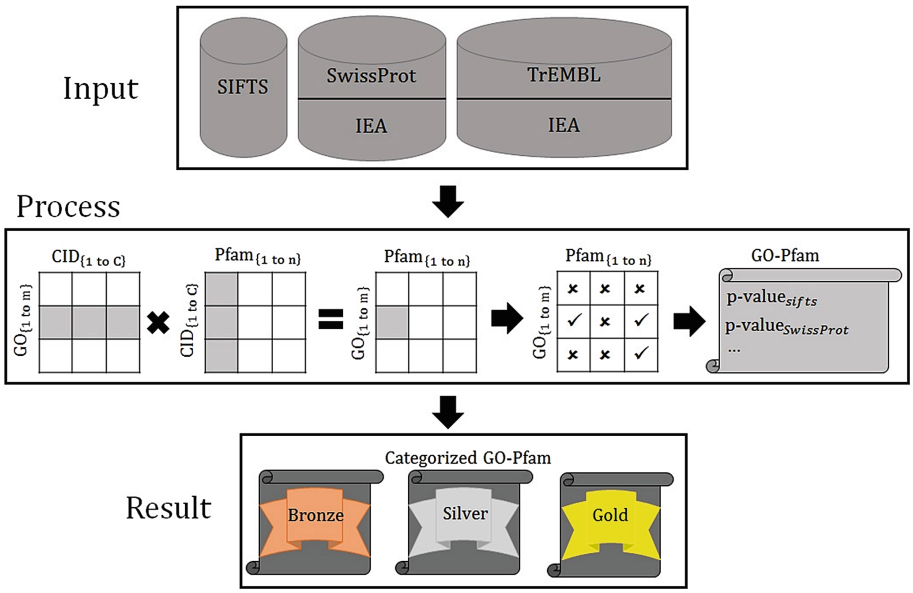


Fig. 2. A schematic overview of the GODM procedure.

p-values are significant in SIFTS, SwissProt, and TrEMBL-IEA). The remaining associations are assigned to the Bronze class. An illustration of the whole procedure is shown in Fig. 2.

3 Results

Our method takes as input five large datasets of MF GO-chain associations from SIFTS, and MF GO-sequence associations from SwissProt, SwissProt-IEA, TrEMBL and TrEMBL-IEA as well as five large datasets of Pfam-Chain and Pfam-sequence associations. These source datasets were merged to give a global dataset of 1,161,372 non-redundant GO-Pfam associations. Using the reference InterPro dataset of 1561 "true" associations against background associations, the best ROC-plot AUC value of 0.99 was obtained with the weights $w_{SIFTS} = 10$, $w_{SwissProt} = 1$, $w_{SwissProt-IEA} = 10$, $w_{TrEMBL} = 1$, and $w_{TrEMBL-IEA} = 8$. These weights clearly give a greater importance to the GO-Pfam associations from SIFTS and the IEA (Inferred from Electronic Annotation) section of SwissProt and TrEMBL compared to those derived from TrEMBL and the manually curated section of SwissProt.

In order to reduce the number of false associations predicted by our approach (and not just to simply optimise the overall AUC performance), various threshold values of the confidence score (using the above weights) were tested on the Training dataset using the F-measure (Sect. 2.3) with respect to the number of true and false associations having scores above or below the threshold. This gave an optimal threshold score of 0.01 for a maximum F-Measure of 0.99. Applying

this threshold to the Test dataset yielded a recall value of 0.965 and a precision value of 1.0 to give a F-measure of 0.98. This threshold was then used to filter GO-Pfam associations from the merged dataset according to their confidence score. It is worth noting that if the ranked list of Test associations is evaluated with respect to the median rank (since the dataset contains equal numbers of true and false instances), the threshold score is 0.0095 and our scoring function gives recall and precision values of 0.965, and thus a F-measure of only 0.965. This shows that using the chosen score threshold of 0.01 provides an objective way to achieve a very low rate of false positive associations while still maintaining very high recall and precision.

3.1 Analysis of Calculated GO-Pfam Associations

The summary of our calculated GO-Pfam associations is shown in Table 1. This table shows the numbers of GO-Pfam associations along with the numbers of distinct GO terms (leaf level) and Pfam entries involved in those associations for the five source datasets, our merged global dataset before and after filtering (the latter corresponding to our "GODM" GO-Pfam associations), and for the InterPro dataset of true associations. The overlap between these two last datasets is shown in the last line of the table.

Table 1. Statistics on the given and filtered MF GO-Pfam associations.

Dataset	GO-Pfam associations	GO terms	Pfam entries
SIFTS	10,064	2,763	3,370
SwissProt	22,435	4,220	4,669
SwissProt-IEA	28,982	3,228	4,469
TrEMBL	22,031	2,766	3,613
TrEMBL-IEA	1,136,711	4,254	9,342
Merged	1,161,372	5,510	9,929
Filtered associations (GODM)	20,318	5,047	6,154
Common with InterPro	1,519	586	1,362
InterPro	1,561	591	1,390

Overall, Table 1 shows that our approach yielded a total of $20,318$ GO-Pfam associations that include $1,519$ associations already present in InterPro. While this shows that our method finds 97.3% of the "correct" GO-Pfam associations in InterPro, it also shows that only 2.7% of the correct InterPro associations have confidence scores below our optimal score threshold of 0.01. This relatively high proportion of common associations reflects the fact that our method is designed to give relatively strong support (Confidence Score) to the correct associations in InterPro based on the five input sources. Concerning statistical significance, nearly half of the GO-Pfam associations belong to the Gold class (48%).

3.2 Comparison Between Our GODM and InterPro GO-Pfam Associations

Figure 3 (A) shows the average number of GO-Pfam associations per GO term and Pfam entry both for InterPro (shown in grey) and our calculated GODM dataset (in black). The ratio for our method is higher for GO terms (4.03 versus 2.64) and Pfam entries (3.3 versus 1.12), which reflects: (i) a significant enrichment in the annotation of Pfam domains; and (ii) participation of Pfam domains in different functions as either a single domain or a part of a complex.

Figure 3 (B) shows the distribution of GO terms (in grey) and Pfam entries (in black) according to the number of associations they are involved in. More than 1,800 GO terms and 2,500 Pfam entries are involved in single associations, i.e. associated with a single Pfam domain and a single GO term respectively. Intersection of these single association sets yields a list of 135 one-to-one GO-Pfam associations. Nevertheless, the distribution also shows that our collection of associations rather favours multiple associations, thereby reflecting the complex many-to-many relationships that exist within the original datasets.

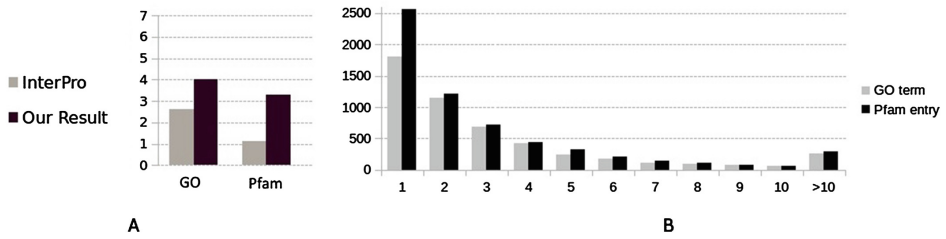


Fig. 3. A: average number of GO-Pfam associations per GO terms and per Pfam entry for the InterPro (grey) and our calculated GODM (black) datasets. B: distribution of GO terms according to their numbers of associations with Pfam entries (grey) and Pfam entries according to their numbers of associations with GO terms (black).

3.3 Comparing GODM and dcGO GO-Pfam Associations

In order to compare our results with dcGO [6], we extracted the Pfam2GO associations from the dcGO website (http://supfam.org/SUPERFAMILY/dcGO) where GO terms are leaves in the MF hierarchy of GO terms. This Pfam2GO dataset includes 3,086 GO-Pfam associations. Figure 4 shows that a total of 2,401 GO-Pfam associations are common to dcGO and our results (overlap B) while only 404 GO-Pfam associations are common between InterPro and dcGO (overlap C). Furthermore, this comparison shows that our GODM dataset contains 17,917 (20,318-2,401) additional GO-Pfam associations that are not available in the dcGO dataset. In a more detailed analysis, the overlap between the GODM and Pfam2GO datasets was studied with respect to our three quality classes. As summarized in the Table 2, the overlap between the two datasets contains 1,621, 600, and 180 Gold, Silver, and Bronze associations, respectively.

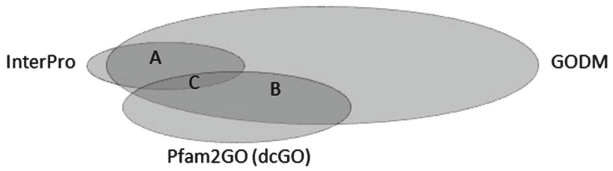


Fig. 4. Venn diagram showing the intersection between Pfam2GO (3,086 associations) from dcGO, our GODM associations (20,318 associations), and manually curated associations (1,561 associations) from InterPro. Region A (1,519 associations) is the overlap between our result and InterPro associations. Region B (2,401 associations) is the common associations between our result and Pfam2GO. Region C (404 associations) is the overlap between Pfam2GO and InterPro associations.

Table 2. Overlap between associations from GODM, Pfam2GO of dcGO, and InterPro.

Dataset	GODM	Overlap	
		With Pfam2GO	With InterPro
Gold	9,771	1,621	922
Silver	4,280	600	455
Bronze	6,267	180	72
Total	20,318	2401	1,519

3.4 Annotating PDB Chains with GO Terms

Our analysis of the July 2015 release of the SIFTS database reveals that some 41% of PDB entries currently lack a leaf GO term annotation. Indeed, we found that a total of 48,409 PDB chains lacking GO annotations in SIFTS include at least one of the 6,154 Pfam domains present in our calculated GODM associations. For those chains, GODM finds 19,371, 7,176 and 12,530 Gold, Silver, and Bronze GO-Pfam associations, respectively, giving a total of 39,077 PDB chains that could benefit from the annotations inferred by GODM. Moreover, 153 PDB chains could benefit from unambiguous one-to-one GO-Pfam associations.

To give an example, GODM finds a Gold association between PF03018 (Dirigent-like protein) and GO term GO:0042349 ("Guiding stereospecific synthesis activity"). Interestingly, the PF03018 domain is present in the PDB chain 4REV A ("Structure of the dirigent protein DRR206") which is not annotated by any GO term from the molecular function ontology. Consequently the GODM recommendation is to annotate the 4REV PDB entry with GO:0042349 term, which explicitly describes the possible function of this protein. Another example is PDB structure 2YRB, which is described only as "the solution structure of the first C2 domain from human KIAA1005 protein", and for which its previously assigned Pfam domain (PF11618) is annotated as a "protein of unknown function (DUF3250)". In this case, GODM finds a Gold association between PF11618 and GO:0031870 (thromboxane A2 receptor binding) thus indicating that this structure could be annotated with that GO term.

4 Conclusion

We have presented a systematic content-based filtering approach for assigning GO terms to protein domains and then categorizing those associations. This was achieved by first collecting existing annotations of protein chains or sequences, namely Pfam domain compositions on one hand and GO-MF leaf term annotations on the other. We then applied the content-based filtering method to find a list of direct associations between GO-MF leaf terms and Pfam domains. Our approach is able to infer a total of 20,318 direct GO-Pfam associations. Thus, compared to the 1,561 manually curated GO-Pfam associations from InterPro database, our approach discovers over 13 times as many associations in a completely automatic way. We have also proposed some possible ways to further analyze the coverage of the our approach. We believe that the large numbers of GO-Pfam associations calculated using our approach can considerably contribute to enriching the annotations of PDB protein chains, and that this will facilitate a better understanding and exploitation of structure-function relationships at the protein domain level.

Acknowledgments. This project is funded by Agence Nationale de la Recherche (grant number ANR-11-MONU-006-02), the Institut National de Recherche en Informatique et Automatique, and Région Lorraine.

References

1. Finn, R.D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E.L., Tate, J., Punta, M.: Pfam: the protein families database. Nucleic Acids Res. **42**(D1), D222–D230 (2014)
2. Berg, J.M., Tymoczko, J.L., Stryer, L.: Protein Structure and Function. W.H Freeman, New York (2002)
3. Bernstein, F.C., Koetzle, T.F., Williams, G.J., Meyer, E.F., Brice, M.D., Rodgers, J.R., Kennard, O., Shimanouchi, T., Tasumi, M.: The protein data bank. Eur. J. Biochem. **80**(2), 319–324 (1977)
4. Gutmanas, A., Alhroub, Y., Battle, G.M., Berrisford, J.M., Bochet, E., Conroy, M.J., Dana, J.M., Montecelo, M.A.F., van Ginkel, G., Gore, S.P., Haslam, P., Hatherley, R., Hendrickx, P.M.S., Hirshberg, M., Lagerstedt, I., Mir, S., Mukhopadhyay, A., Oldfield, T.J., Patwardhan, A., Rinaldi, L., Sahni, G., Sanz-García, E., Sen, S., Slowley, R.A., Velankar, S., Wainwright, M.E., Kleywegt, G.J.: PDBe: protein data bank in europe. Nucleic Acids Res. **42**(D1), D285–D291 (2014)
5. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.: Gene ontology: tool for the unification of biology. Nature Genet. **25**(1), 25–29 (2000)
6. Fang, H., Gough, J.: dcGO: database of domain-centric ontologies on functions, phenotypes, diseases and more. Nucleic Acids Res. **41**(D1), D536–D544 (2013)
7. Alborzi, S.Z., Devignes, M.D., Ritchie, D.W.: EC-PSI: associating enzyme commission numbers with Pfam domains. bioRxiv, 022343 (2015). doi:10.1101/022343

8. Alborzi, S.Z., Devignes, M.D., Ritchie, D.W.: ECDomainminer: discovering hidden associations between enzyme commission numbers and pfam domains. BMC Bioinform. **18**(1), 107 (2017)
9. Hanani, U., Shapira, B., Shoval, P.: Information filtering: overview of issues, research and systems. User Model. User-Adap. Interact. **11**(3), 203–259 (2001)
10. Ricci, F., Rokach, L., Shapira, B.: Introduction to recommender systems handbook. In: Ricci, F., Rokach, L., Shapira, B., Kantor, P.B. (eds.) Recommender Systems Handbook, pp. 1–35. Springer, Heidelberg (2011)
11. Breese, J.S., Heckerman, D., Kadie, C.: Empirical analysis of predictive algorithms for collaborative filtering. In: Proceedings of the Fourteenth Conference on Uncertainty in Artificial Intelligence, pp. 43–52. Morgan Kaufmann Publishers Inc. (1998)
12. Koren, Y., Bell, R.: Advances in collaborative filtering. In: Ricci, F., Rokach, L., Shapira, B. (eds.) Recommender Systems Handbook, pp. 77–118. Springer, Heidelberg (2015)
13. Basu, C., Hirsh, H., Cohen, W., et al.: Recommendation as classification: using social and content-based information in recommendation. In: AAAI/IAAI, pp. 714–720 (1998)
14. Velankar, S., Dana, J.M., Jacobsen, J., van Ginkel, G., Gane, P.J., Luo, J., Oldfield, T.J., ODonovan, C., Martin, M.J., Kleywegt, G.J.: SIFTS: structure integration with function, taxonomy and sequences resource. Nucleic Acids Res. **41**(D1), D483–D489 (2013)
15. Mitchell, A., Chang, H.Y., Daugherty, L., Fraser, M., Hunter, S., Lopez, R., McAnulla, C., McMenamin, C., Nuka, G., Pesseat, S., Sangrador-Vegas, A., Scheremetjew, M., Rato, C., Yong, S.Y., Bateman, A., Punta, M., Attwood, T.K., Sigrist, C.J.A., Redaschi, N., Rivoire, C., Xenarios, I., Kahn, D., Guyot, D., Bork, P., Letunic, I., Gough, J., Oates, M., Haft, D., Huang, H., Natale, D.A., Wu, C.H., Orengo, C., Sillitoe, I., Mi, H., Thomas, P.D., Finn, R.D.: The InterPro protein families database: the classification resource after 15 years. Nucleic Acids Res. **43**(D1), D213–D221 (2015)
16. Suzek, B.E., Huang, H., McGarvey, P., Mazumder, R., Wu, C.H.: UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics **23**(10), 1282–1288 (2007)
17. Fawcett, T.: An introduction to ROC analysis. Pattern Recogn. Lett. **27**(8), 861–874 (2006)
18. Cui, X., Churchill, G.A., et al.: Statistical tests for differential expression in cDNA microarray experiments. Genome Biol. **4**(4), 210 (2003)

C.1.4 Automatic Generation of Functional Annotation Rules Using Inferred GO-Domain Associations

There are millions of proteins with known sequences and unknown functions. The most reliable way to assign functions to proteins is by expert curators, but this is an expensive and time-consuming process. The huge gap between the small number of expert curators and the ever increasing number of new unannotated protein sequences has motivated the development of many automatic annotation approaches. These approaches aim for a balance between maximizing the number of annotations while minimizing the number of false assignments. However, achieving this aim in a reliable way remains an open research problem.

We present here a novel approach called CARDM (Combinatorial Association Rules Domain Miner) which exploits that fact that many proteins consist of one or more domains. CARDM combines a learning step in which functional annotations are assigned to protein domains, and a combinatorial step in which association rules are generated and filtered using previously validated annotations. The filtered rules are then aggregated to build predictive models that are used to automatically annotate protein sequences and structures. CARDM has been tested on the entire set of TrEMBL entries and on the dataset provided at the international 2013 CAFA (Critical Assessment of Functional Annotation) challenge. Overall, CARDM predicts 24 million EC numbers and 188 million GO terms for the protein entries in TrEMBL. We find that the performance of CARDM on the CAFA 2013 targets is similar to that of the best predictor groups in that round of CAFA.

Automatic Generation of Functional Annotation Rules Using Inferred GO-Domain Associations

Seyed Ziaeddin Alborzi [1,3]*, Sabeur Aridhi [1], Marie-Dominique Devignes [2], Rabie Saidi [4], Alexandre Renaux [4], Maria J. Martin [4], David W. Ritchie [3]
1 Universite de Lorraine, LORIA, UMR 7503, 54506 Vandœuvre-les-Nancy, France
2 CNRS, LORIA, UMR 7503, 54506 Vandœuvre-les-Nancy, France.
3 Inria Nancy Grand-Est, 54600 Villers-les-Nancy, France
4 European Molecular Biology Laboratory, European Bioinformatics Institute, Cambridge CB10 1SD, UK
*To whom correspondence should be addressed: seyed-ziaeddin.alborzi@inria.fr

1. INTRODUCTION

The GO ontology is widely used for functional annotation of genes and proteins. It describes biological processes (BP), molecular function (MF), and cellular components (CC) in three distinct hierarchical controlled vocabularies. At the molecular level, functions are often performed by highly conserved parts of proteins, identified by sequence or structure alignments and classified into domains or families (SCOP, CATH, PFAM, TIGRFAMs, etc.). The InterPro database provides a valuable integrated classification of protein sequences and domains which is linked to nearly all existing other classifications. Interestingly, several InterPro families have been manually annotated with GO terms using expert knowledge and the literature. However, the list of such annotations is incomplete (only 20% of Pfam domains and families possess MF GO functional annotation). We therefore developed the GODM approach to expand the available functional annotations of protein domains and families (1). Based on our ECDomainMiner approach (2), we use the respective associations of protein sequences with GO terms and protein domains to infer direct associations between GO terms and protein domains.

2. INFERRING GO-DOMAIN ASSOCIATIONS USING GODM

GODM finds associations between GO terms and protein domains from the known associations between (i) GO terms and protein sequences and (ii) the same protein sequences and the domains they are known to contain. The domains may belong to any domain classification such as Pfam. We used two types of datasets: (i) SIFTS for associations between PDB chains and GO terms or domains, (ii) the Swissprot and TrEMBL sections of UniProtKB for associations between sequence accession numbers (ANs) and GO terms or domains. Next, based on the evidence code of the GO term assignment, AN-GO term associations in the SwissProt and TrEMBL datasets are divided into two groups, namely associations for which GO terms were Inferred from Electronic Annotation (IEA) and the rest. These four input datasets are subsequently called Swissprot, Swissprot-IEA, TrEMBL, and TrEMBL-IEA. In order to exploit the GO hierarchy, associations involving ancestors of GO terms are also added to the datasets. Finally, PDB chains and ANs are grouped into non-redundant clusters having identical sequences using the Uniref100 resource.

In each dataset prepared in this way, each GO term and domain is assigned a feature vector of associated chain or AN clusters. This allows to calculate cosine similarities between GO terms and domains. The scores assigned to each vector pair in each of the five datasets are combined using a weighted average. The individual weights are optimised by calculating the ROC performance plot and maximizing the AUC with manually confirmed GO-Domain associations from InterPro as positive examples, against all others. Then, a threshold is chosen for the weighted score in order to eliminate weak GO-domain associations. Finally a *p-value* is calculated for each GO-domain association in each dataset using a hypergeometric distribution.

3. RESULTS FOR GO-PFAM ASSOCIATIONS

The GODM method infers 20,318 GO-Pfam associations where GO terms are leaves in the MF hierarchy of GO terms. Compared to the 1561 manually curated GO-Pfam associations in InterPro, this represents a 13-fold increase in the number of GO-Pfam associations. Furthermore, the GODM associations have been compared with the dcGO database (3) that includes 3,086 comparable GO-Pfam associations. A total of 2,401 GO-Pfam associations are common between dcGO and our results revealing that our GODM dataset contains 17,917 additional GO-Pfam associations. Moreover the overlap with the 1561 InterPro GO-Pfam associations is of 1519 for the GODM dataset versus only 404 for the dcGO dataset. The GODM method was also run with the SCOP and CATH classifications of domains or families and yielded very similar results.

4. USING THE GODM RESOURCE TO GENERATE ANNOTATION RULES

In this section, we present a systematic way to generate high confidence rules for protein annotation using the GODM associations. We first ran GODM several times to find associations between GO terms and domains from the different domain classifications (such as PFAM, TIGRFAMs, etc.). Then, all associations were grouped for each given GO term resulting in an association of the GO term with a set of domains pertaining from diverse classifications. We then generated all possible subsets of domains ($\{D_1,..., D_n\}$, $n \leq 4$) and associated them with the concerned GO term, GO_k. The subsets of domains were further diversified by adding a taxon (T_j) from a list of interest (one per subset). These complex associations, ($\{\{D_1,..., D_n\}, T_j\}, Go_k$), were converted into annotation rules:

IF a sequence S belongs to taxon T_j and S contains domains $\{D_1,..., D_n\}$
THEN S is annotated by GO_k.

In order to verify the quality of each generated rule, a confidence score was assigned as the ratio of the number of SwissProt sequences verifying the rule over the number of SwissProt sequences verifying the premise of the rule. Candidate rules with high confidence (usually 100%) are retained and used to assign GO terms to unannotated protein sequences. When using Pfam, SCOP, CATH, Panther, PROSITE, CDD, SMART, PRINTS, and TIGRFAM domain classification for GODM, and a set of 40 taxa from CAFA3 unannotated protein sequences, we obtained 6,357, 17,466, and 2,338 annotation rules for MF, BP, and CC GO terms with 100% confidence on SwissProt. These rules were used to annotate target protein sequences in the CAFA3 challenge (http://biofunctionprediction.org/cafa/). There were a total of 121,914 target sequences having at least one known domain present in our GODM-derived rules. Using our high confidence annotation rules, we obtained 188,549 MF, 315,310 BP, and 191,835 CC GO term predictions for 98,849, 106,346, and 105,274 distinct CAFA3 target sequences, respectively.

5. CONCLUSION

The GODM approach provides a substantial enrichment of functional annotations at the protein domain level which has been exploited here for protein functional annotation but can also be used to deepen our knowledge about structure-function relationships at the domain level.

6. REFERENCES

1. Alborzi, S.Z., Devignes, M.D. and Ritchie, D.W., 2017, April. Associating Gene Ontology Terms with Pfam Protein Domains. *In International Conference on Bioinformatics and Biomedical Engineering* (pp. 127-138). Springer, Cham.
2. Alborzi, S.Z., Devignes, M.D. and Ritchie, D.W., 2017. ECDomainMiner: discovering hidden associations between enzyme commission numbers and Pfam domains. *BMC bioinformatics*, 18(1), p.107.
3. Fang, H. and Gough, J., 2013. dcGO: database of domain-centric ontologies on functions, phenotypes, diseases and more. *Nucleic acids research*, 41(D1), D536-D544.

C.2 Posters

C.2.1 EC-PSI: Associating Enzyme Commission Numbers with Pfam Domains

With the growing number of protein structures in the protein data bank (PDB), there is a need to annotate these structures at the domain level in order to relate protein structure to protein function. Thanks to the SIFTS database, many PDB chains are now cross-referenced with Pfam domains and enzyme commission (EC) numbers. However, these annotations do not include any explicit relationship between individual Pfam domains and EC numbers. This article presents a novel statistical training-based method called EC-PSI that can automatically infer high confidence associations between EC numbers and Pfam domains directly from EC-chain associations from SIFTS and from EC-sequence associations from the SwissProt, and TrEMBL databases. By collecting and integrating these existing EC-chain/sequence annotations, our approach is able to infer a total of 8,329 direct EC-Pfam associations with an overall F-measure of 0.819 with respect to the manually curated InterPro database, which we treat here as a "Gold Standard" reference dataset. Thus, compared to the 1,493 EC-Pfam associations in InterPro, our approach provides a way to find over six times as many high quality EC-Pfam associations completely automatically.

C.2.2 Associating Gene Ontology Terms with Pfam Protein Domains

The fast growing number of protein structures in the protein data bank (PDB) raises new opportunities to study protein structure-function relationships. As the biological activity of many proteins often arises from specific domain-domain and domain-ligand interactions, there is a need to provide a direct mapping from structure to function at the domain level. Many protein entries in PDB and UniProt are annotated to show their component protein domains according to Pfam classification, as well as their molecular function through the Gene Ontology (MF GO) terms. We therefore hypothesize that relevant MF GO-domain associations are hidden in this complex dataset of annotations.

C.2.3 Using Content-Based Filtering to Infer Direct Associations between the CATH, Pfam, and SCOP Domain Databases

Protein domain structure classification systems such as CATH and SCOP provide a useful way to describe evolutionary structure-function relationships. Similarly, the Pfam sequence-based classification identifies sequence-function relationships. Nonetheless, there is no complete direct mapping from one classification to another. This means that functional annotations that have been assigned to one classification cannot always be assigned to another. Here, we present a novel content-based filtering approach called CAPS (Computing direct Associations between annotations of Protein Sequences and Structures) to systematically analyze multiple protein-domain relationships in the SIFTS and UniProt databases in order to infer direct mappings between CATH superfamilies, Pfam clans or families, and SCOP superfamilies. We then compare the result with existing mappings in Pfam, InterPro, and Genome3D.

C.2.4 Automatic Generation of Functional Annotation Rules Using Inferred GO-Domain Associations

The GO ontology is widely used for functional annotation of genes and proteins. It describes biological processes (BP), molecular function (MF), and cellular components (CC) in three distinct hierarchical

controlled vocabularies. At the molecular level, functions are often performed by highly conserved parts of proteins, identified by sequence or structure alignments and classified into domains or families (SCOP, CATH, PFAM, TIGRFAMs, etc.). The InterPro database provides a valuable integrated classification of protein sequences and domains which is linked to nearly all existing other classifications. Interestingly, several InterPro families have been manually annotated with GO terms using expert knowledge and the literature. However, the list of such annotations is incomplete (only 20% of Pfam domains and families possess MF GO functional annotation). We therefore developed the GODomainMiner approach to expand the available functional annotations of protein domains and families. Based on our ECDomainMiner approach, we use the respective associations of protein sequences with GO terms and protein domains to infer direct associations between GO terms and protein domains. Finally, we used our calculated GO-domain associations to devise a systematic way, called AutoProf-Annotator (* Changed to CARDM *), to generate high confidence rules for protein sequence (or structure) annotation.

EC-PSI: Associating Enzyme Commission Numbers with Pfam Domains

Seyed Ziaeddin Alborzi[1,2], Marie-Dominique Devignes[3], David W. Ritchie[1]
[1]INRIA Nancy Grand-Est, [2]CNRS Nancy, [3]Université de Lorraine

INTRODUCTION

With the growing number of protein structures in the protein data bank (PDB) [1], there is a need to annotate these structures at the domain level in order to relate protein structure to protein function. Thanks to the SIFTS database [2], many PDB chains are now cross-referenced with Pfam domains [3] and Enzyme Commission (EC) numbers [4]. However, these annotations do not include any explicit relationship between individual Pfam domains and EC numbers. This poster presents a novel statistical training-based method called EC-PSI (for EC-Pfam Statistical Inferring) that can automatically infer high confidence associations between EC numbers and Pfam domains directly from EC-chain associations from SIFTS and from EC-sequence associations from the SwissProt, and TrEMBL databases [5].

MATERIALS & METHODOLOGY

A. Data sources

- **SIFTS**: A database which provides relations between PDB structures and other resources.
- **UniProt**: A protein sequence resource which is divided into manually (Swissprot) and automatically curated (Trembl) databases.
- **Interpro**: An integrated database of protein domains with reviewed functional annotations (= available "Gold Standard" set of "true" EC-Pfam associations) [6].

B. Algorithm

1. Extract from SIFTS data associations between 4-digit EC numbers and PDB chains, and associations between PDB chains and Pfam domains, leading to many-to-many relationships between EC numbers and Pfam domains (Figure 1). Repeat this step with sequences instead of PDB chains using SwissProt and Trembl.

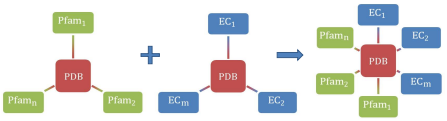


Figure 1. Building many-to-many relationships between EC numbers and Pfam domains.

2. Draw a tree-like set of relations for each EC number using all its associated PDB chains (Figure 2). Repeat this step with SwissProt and Trembl data.

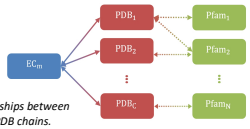


Figure 2. Tree-like representation of the relationships between an EC number EC_m and N Pfam domains via C PDB chains.

3. Calculate the EC-Pfam frequency score (PPFEC) for a given pair (EC_m, D_n) as the ratio between the number of PDB chain having Pfam domain D_n and the total number of PDB chains associated with EC_m(Formula 1). Calculate the corresponding frequencies from SwissProt (PSFEC) and Trembl (PTFEC).

$$PPFEC_n^m = \frac{|\{P_i^m; D_n \in P_i^m, i=1,\ldots,C^m\}|}{C^m} \quad (1)$$

4. Aggregate the three frequency scores into one confidence score (Formula 2).

$$ConfidenceScore_{m,n} = \frac{a \times PPFEC_{m,n} + b \times PSFEC_{m,n} + c \times PTFEC_{m,n}}{a+b+c} \quad (2)$$

5. Find the best values for weighting factors a, b, c using our InterPro-derived Gold Standard. Values for a, b and c varied from 1 to 10 in steps of 1. For each combination, the "true" associations retrieved from InterPro and an equivalent number of "false" associations were scored and a ROC plot was drawn. The highest AUC value (Area Under the Curve) was chosen to select the best three values : a= 1, b=10, c=1.

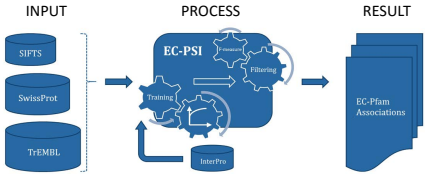


Figure 3. Flow-chart of the EC-PSI data processing and training procedure.

RESULTS

A. Statistics

The EC-PSI method inferred in a completely automatic manner nearly six-times more associations than in our InterPro "gold standard" dataset (Table 1).

Table 1. Statistics on the given and calculated EC-Pfam associations.

Dataset	EC-Pfam assoc.	4-digit EC no.	Pfam domains
SIFTS	6204	2575	2606
SwissProt	9879	3959	3147
TrEMBL	28572	3538	5839
Merged	32018	4588	6290
InterPro	1493	676	1273
EC-PSI (Calculated)	8329	4436	2462
Common to EC-PSI and InterPro	1089	593	944

The optimal score threshold found to be **0.08**. Applying this threshold to Test Dataset yielded F-measure, precision, and recall values of **0.81**, **0.948**, and **0.707,** respectively.

B. Increase in EC-Pfam associations depending on the top-level EC branch (first-digit)

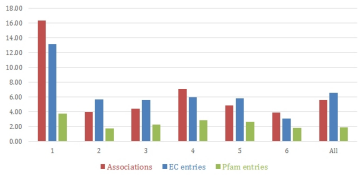


Figure 4. Scale-up factors for EC-PSI versus InterPro associations (red), EC entries (blue), Pfam domains (green), depending on the EC branch. 1: Oxydoreductases, 2: Transferases, 3: Hydrolases, 4: Lyases, 5: Isomerases, 6: Ligases, All: All EC branches.

CONCLUSIONS & PERSPECTIVES

We have developed a statistical method for inferring associations between EC number and Pfam Domains.
We are currently applying the method on the 3-digit level of EC classification.
The large numbers of EC-Pfam associations calculated using our approach can contribute considerably to enriching the annotations of PDB protein chains (Figure 5).
This will facilitate a better understanding and exploitation of structure-function relationships at the protein domain level.

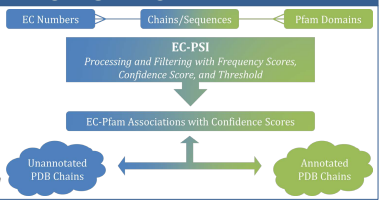


Figure 5. Using EC-PSI to transform EC-Chain/Sequence annotations into EC-Pfam annotations with confidence scores, thus enriching PDB chains annotations.

EXAMPLE

PDB entry **1JVN** is associated in SIFTS with two domains:

- N-terminal: Glutamine amidotransferase class-1 **(PF00117)**.
- C-terminal: Histidine biosynthesis protein **(PF00977)**.

And is annotated with:

- **EC 2.4.2.-** : Pentosyl transferase.

EC-PSI retrieved the following annotations for each domain:

- 8 EC numbers for **PF00117** with a majority of **EC 6.3.-.-**: Ligase forming carbon-nitrogen bonds.
- 1 EC number for **PF00977** : **EC 5.3.1.16** specific isomerase, part of the Histidine biosynthesis pathway.

These new annotations (not present in InterPro) enrich the global annotation of this PDB structure.

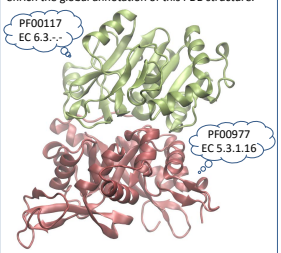


Figure 6. Schematic presentation of 1JVN.

LITERATURE CITED

1. F. C. Bernstein et al. The protein data bank. European Journal of Biochemistry, 80(2) :319–324, 1977.
2. S. Velankar et al. SIFTS: structure integration with function, taxonomy and sequences resource. Nucleic Acids Research, 41(D1): D483–D489, 2013.
3. R. D. Finn et al. Pfam: the protein families database. Nucleic Acids Research, 42(D1): D222–D230, 2014.
4. A. Fleischmann et al. IntEnz, the integrated relational enzyme database. Nucleic Acids Research, 32(suppl 1): D434–D437, 2004.
5. The UniProt Consortium. The universal protein resource (UniProt) in 2010. Nucleic Acids Research, 38(suppl 1): D142–D148, 2010
6. A. Mitchell et al. The InterPro protein families database : the classification resource after 15 years. Nucleic Acids Research, 43(D1): D213–D221, 2015.

ACKNOWLEDGMENTS

This project is funded by the Agence Nationale de la Recherche (grant reference ANR-11-MONU-006-02),
the Institut National de Recherche en Informatique et Automatique (Inria), and the Lorraine Region.

CONTACT

Zia Alborzi
INRIA Nancy Grand Est, LORIA, Bureau B133, France
Email: seyed-ziaeddin.alborzi@inria.fr
Website: www.loria.fr/~salborzi
Phone: +33 6 83 29 89 83

Associating Gene Ontology Terms with Pfam Protein Domains Using Content-Based Filtering

Seyed Ziaeddin Alborzi[1,3], Maxime Guyot[3], David W. Ritchie[1], Marie-Dominique Devignes[2,3]
[1]INRIA Nancy Grand-Est, [2]CNRS Nancy, [3]Université de Lorraine

INTRODUCTION

The fast growing number of protein structures in the protein data bank (PDB) [1] raises new opportunities to study protein structure-function relationships. As the biological activity of many proteins often arises from specific domain-domain and domain-ligand interactions, there is a need to provide a direct mapping from structure to function at the domain level.

Many protein entries in PDB and UniProt are annotated to show their component protein domains according to Pfam classification [2], as well as their molecular function through the Gene Ontology (GO_MF) terms [3]. We therefore hypothesize that relevant GO_MF-domain associations are hidden in this complex dataset of annotations.

MATERIALS & METHODOLOGY

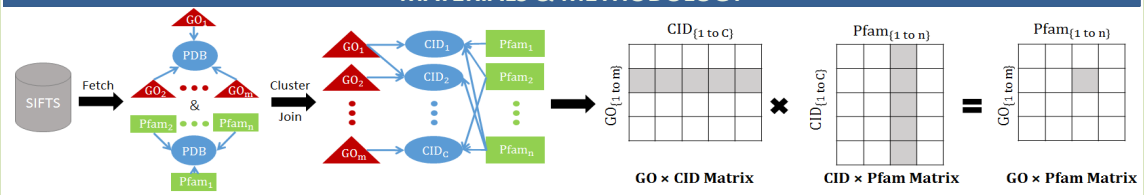


Fig 1. A graphical illustration of the content-based filtering workflow for calculating raw GO-Pfam from existing GO-Chain and Pfam-Chain associations.

A. Input Data Sources
- **SIFTS** [4] (PDB annotations) ; **UniProt** [5] (Swissprot + Trembl, each divided in two based on IEA evidence code) ;

B. Extract Associations as two bipartite graphs
- between protein instances and either "leaf" GO terms or Pfam domains.

C. Cluster identical sequences and join the two graphs.
- CID: Clusters of identifiers for proteins with 100% identical sequence
- Many-to-many relations between GO terms and Pfam with CIDs as intermediates

D. Map Graphs to Adjacency Matrices and Apply Content-Based Filtering
- The cosine similarity score of a given pair (GO_m, $Pfam_n$) is the ratio between the number of CIDs shared by GO_m and $Pfam_n$, and the square root product of the total number of CIDs associated with GO_m and the total number of CIDs associated with $Pfam_n$.

E. Repeat the Procedure for Other Sources
- Repeat steps B, C, and D using SwissProt, SwissProt-IEA, Trembl, and Trembl-IEA.

F. Create a Gold Standard
- From InterPro [6] (1561 reviewed GO-Pfam associations = *"positive examples"*)
- From shuffled SIFTS data (1561 *"negative samples"*)

G. Determine Source Weights
- Aggregate the fiver similarity scores into one confidence score:

$$ConfidenceScore_{go,pfam} = \frac{\sum_i w_i s(go,pfam)}{\sum_i w_i}$$

- Values for w_{sifts}, $w_{swissprot}$, $w_{swissprot-IEA}$, w_{trembl} and $w_{trembl-IEA}$ varied from 1 to 10 in steps of 1.
- For each combination, the positive examples from InterPro are scored against all other associations (background); ROC plot is drawn and maximal AUC (0.995) is obtained for : w_{sifts} = 10, $w_{swissprot}$ = 1, $w_{swissprot-IEA}$ = 10, w_{trembl} =1, and $w_{trembl-IEA}$ = 8

H. Determining the threshold score
- From 1 to 0.01 in steps of 0.01, F-measure is calculated with respect of InterPro *positive* and shuffled SIFTS dataset *negative samples*.
- Optimal F-measure (88.6%) for **threshold score at 0.02.**

I. Calculating P-Values for each association in each dataset
- Hypergeometric law + Bonferroni correction (according to Fang and Gough 2013),

J. Categorize the Inferred Associations (score > threshold)
- **Gold**: all P-values significant ; **Silver**: number of significant P-values ≥ non-significant ones; **Bronze**: the rest.

STATISTICS

Dataset	GO_MF - Pfam associations	GO_MF terms (leaves)	Pfam domains
SIFTS	10,064	2,763	3,370
SwissProt	22,435	4,220	4,669
SwissProt-IEA	28,982	3,228	4,469
TrEMBL	22,031	2,766	3,613
TrEMBL-IEA	1,136,711	4,254	9,342
Merged	1,161,372	5,510	9,929
GODM_MF	**15,508**	**4,661**	**5,306**
InterPro	1,561	591	1,390
Common with GODM	1,433	584	1,301

Table 1. Statistics on the calculated GO-Pfam associations versus InterPro.

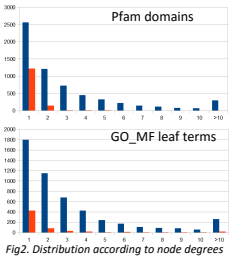


Fig2. Distribution according to node degrees

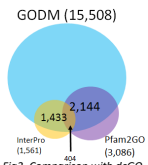


Fig3. Comparison with dcGO
Venn diagram showing the various intersections between the 3,086 Pfam2GO associations from dcGO [7], our 15,508 GODM_MF associations associations, and the 1,561 manually curated associations from InterPro.

GRAPH-BASED EXPLORATION

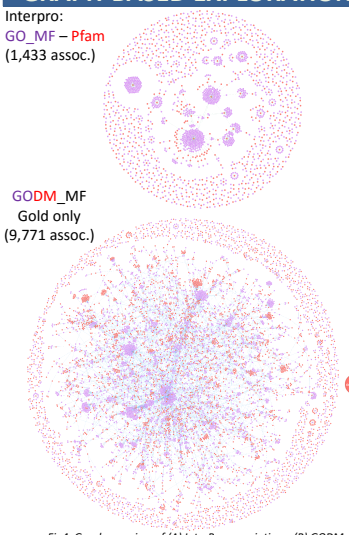


Fig4. Graph overview of (A) InterPro associations, (B) GODM_MF Gold-type associations, using Neo4J graph database.

TRANSFERING ANNOTATIONS

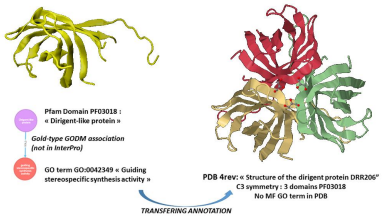


Fig 5. Example of transferring newly discovered domain annotation to PDB entry

Pfam ID	Pfam name	GO_MF ID	GO_MF name	Score
PF01040		8495		GOLD
PF02628		8495		GOLD
PF07798		8495		GOLD
PF01040		8412		GOLD
id		2083		GOLD
id		46408		GOLD
id		46428		GOLD
id		47293		GOLD
id		47295		GOLD
id		47292		BRONZE
id		10176		BRONZE
id		10355		BRONZE
id		10356		BRONZE
id		10357		BRONZE

Fig 6 and Table 2: Analysis of selected subgraphs (not in InterPro)

CONCLUSION

- The 13-fold increase in (GO-MF , Pfam) associations compared with interPro is a possible reservoir of functional annotations for structural domains of unknown function in the Pfam database.
- Multiple associations should be explored carefully.
- The GODM resource could be used to annotate thousands of PDB chains or protein sequences which currently lack any GO annotation although their domain composition is known.

LITERATURE CITED

1. F. C. Bernstein et al. The protein data bank. European Journal of Biochemistry, 80(2) :319–324, 1977.
2. R. D. Finn et al. Pfam: the protein families database. Nucleic Acids Research, 42(D1): D222–D230, 2014.
3. The GO consortium
4. S. Velankar et al. SIFTS: structure integration with function, taxonomy and sequences resource. Nucleic Acids Research, 41(D1): D483–D489, 2013.
5. The UniProt Consortium. The universal protein resource (UniProt) in 2010. Nucleic Acids Research, 38(suppl 1): D142–D148, 2010
6. A. Mitchell et al. The InterPro protein families database : the classification resource after 15 years. Nucleic Acids Research, 43(D1): D213–D221, 2015.
7. Fang, H. and Gough, J. (2013). dcGO: database of domain-centric ontologies on functions, phenotypes, diseases and more. Nucleic acids research, 41(D1):D536–D544.

ACKNOWLEDGMENTS

This project is funded by the Agence Nationale de la Recherche (grant reference ANR-11-MONU-006-02), Inria and the Lorraine Region.

CONTACT

Zia Alborzi
INRIA Nancy Grand Est, LORIA, Bureau B133, France
Email: seyed-ziaeddin.alborzi@inria.fr
Website: www.loria.fr/~salborzi
Phone: +33 (0)7 83 29 89 83

Using Content-Based Filtering to Infer Direct Associations between the CATH, Pfam, and SCOP Domain Databases

Seyed Ziaeddin Alborzi[1,3], David W. Ritchie[1], Marie-Dominique Devignes[2,3]
[1]INRIA Nancy Grand-Est, [2]CNRS Nancy, [3]Université de Lorraine

INTRODUCTION

Protein domain structure classification systems such as CATH and SCOP provide a useful way to describe evolutionary structure-function relationships. Similarly, the Pfam sequence-based classification identifies sequence-function relationships. Nonetheless, there is no complete direct mapping from one classification to another. This means that functional annotations that have been assigned to one classification cannot always be assigned to another. Here, we present a novel content-based filtering approach called **CAPS (Computing direct Associations between annotations of Protein Sequences and Structures)** to systematically analyze multiple protein-domain relationships in the SIFTS and UniProt databases in order to infer direct mappings between CATH superfamilies, Pfam clans or families, and SCOP superfamilies. We then compare the result with existing mappings in Pfam, InterPro, and Genome3D.

MATERIALS & METHODOLOGY

Content-Based Filtering Intuition

User and Film Profiles

	Star1	Star2	Star3	Star4	Star5
User1	0	0	0	1	1
User2	1	1	0	0	1

	Star1	Star2	Star3	Star4	Star5
Film1	0	0	0	0	1
Film2	0	1	0	1	0

(Cosine) Similarity

	Film1	Film2
User1	0.7	0.5
User2	0.57	0.4

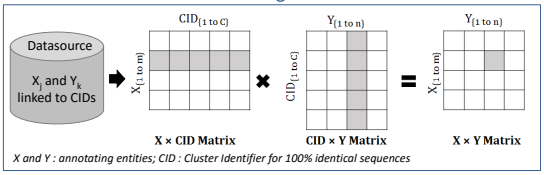
	CATH1	CATH2
SCOP1	0.7	0.5
SCOP2	0.57	0.4

SCOP and CATH Profiles

	CID1	CID2	CID3	CID4	CID5
SCOP1	0	0	0	1	1
SCOP2	1	1	0	0	1

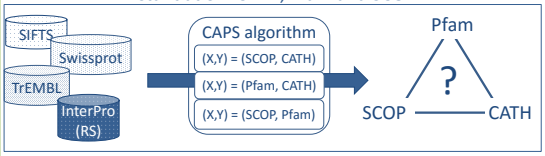
	CID1	CID2	CID3	CID4	CID5
CATH1	0	0	0	0	1
CATH2	0	1	0	1	0

Generalization: Scoring Matrix Calculation



X and Y : annotating entities; CID : Cluster Identifier for 100% identical sequences

Instantiation: CATH, Pfam and SCOP



The CAPS algorithm

Given

- X and Y, two sets of annotating entities,
- RS+, a reference set of confirmed associations between elements of X and Y,
- S_1 to S_n, n sources of relations between X and protein chains or sequences (CIDs), and between Y and protein chains or sequences,

1. For each datasource Si

- Extract X-CID and Y-CID relations (CID: clusters of 100% identical sequences)
- Compute dot product of normalized $(X \times CID)_i$ and $(CID \times Y)_i$ matrices to get the $(X \times Y)_i$ cosine similarity matrix for source S_i.

2. Aggregate similarity scores of all sources

$$ConfidenceScore_{X_j,Y_k} = \frac{\sum_i^n w_i s(X_j,Y_k)}{\sum_i^n w_i}$$

- Determine sources weights w_i that maximize AUC in ROC plots using RS positive examples of X-Y associations against background.

3. Determine the threshold confidence score

- Build a set of negative examples RS- (by random shuffling of relations supporting RS+) and build training and test sets of positive and negative examples
- Select confidence score threshold that maximizes F-measure on the training set of positive and negative X-Y associations and evaluate on the test set.

4. (Optional) Calculate P-Values for each association in each source S_i

- Hypergeometric law + Bonferroni correction
- Categorize the CAPS-inferred associations (Gold: all P-values significant ; Silver: more significant P-values than non significant ones ; Bronze: The rest).

Return

- "CAPS-inferred" X-Y associations (score > threshold), score and category.

CAPS-INFERRED ASSOCIATIONS

SCOP-CATH | Pfam-CATH | SCOP-Pfam

Table 1. CAPS mappings versus InterPro.

Dataset	SCOP-CATH Mappings	SCOP SupFam	CATH SupFam	Dataset	Pfam-CATH Mappings	Pfam Clans/Fam	CATH SupFam	Dataset	SCOP-Pfam Mappings	SCOP SupFam	Pfam Clans/Fam
Merged	580,763	1,851	2,604	Merged	1,068,601	7,228	2,754	Merged	1,004,741	2,111	7,165
CAPS	5,576	1,817	2,549	CAPS	7,623	3,033	2,745	CAPS	6,618	2,109	3,168
InterPro	2,856	1,637	2,231	InterPro	3,573	2,008	2,494	InterPro	2,100	1,537	1,752
Common with CAPS	2,764	1,634	2,225	Common with CAPS	3,494	1,998	2,489	Common with CAPS	2,053	1,532	1,745

Fig1. Distribution according to node degrees (Number of Associations)

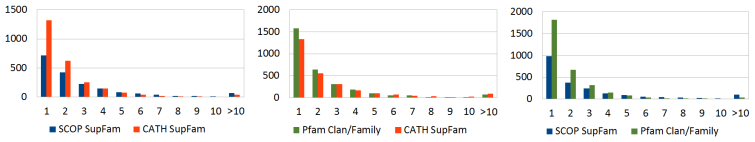
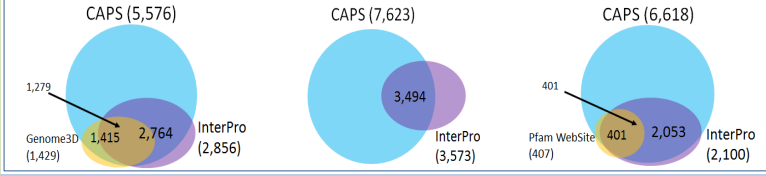
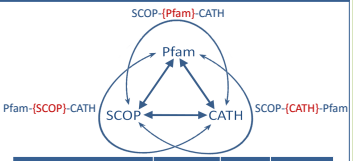


Fig 2. Intersection between our result (CAPS), InterPro, Genome3D, and Pfam website mappings.



TRIANGULAR VERIFICATION



	Mappings	SCOP	CATH
SCOP-CATH	5,576	1,817	2,549
Common with SCOP-{Pfam}-CATH	**5,438**	**1786**	**2518**
1:1 SCOP-CATH	506	506	506
Common with SCOP-{Pfam}-CATH	492	492	492

	Mappings	SCOP	CATH
Pfam-CATH	7,623	3,033	2,745
Common with Pfam-{SCOP}-CATH	**6,768**	**2629**	**2518**
1:1 Pfam-CATH	457	457	457
Common with Pfam-{SCOP}-CATH	393	393	393

	Mappings	SCOP	CATH
SCOP-Pfam	6,618	2,109	3,168
Common with SCOP-{CATH}-Pfam	**5,628**	1786	2629
1:1 Pfam-CATH	635	635	635
Common with Pfam-{SCOP}-CATH	478	478	478

CONCLUSION

- Over 90% of all associations found, are self-consistent with respect to triangular (SCOP-CATH-Pfam) associations.
- Overall, our approach finds 4 times as many SCOP-CATH superfamily associations than currently exist in Genome3D. These new associations will be beneficial to:
  1. Transfer annotations from one classification scheme to another.
  2. Investigate annotation consistency between different classifications.
- We are currently extending our approach to
  1. Analyze multiple associations in more detail.
  2. Confirm the associations using 3D structure alignment

LITERATURE CITED

1. F. C. Bernstein et al. The protein data bank. European Journal of Biochemistry, 80(2) :319–324, 1977.
2. R. D. Finn et al. Pfam: the protein families database. Nucleic Acids Research, 42(D1): D222–D230, 2014.
3. The GO consortium
4. A. G. Murzin et al. SCOP: a structural classification of proteins database for the investigation of sequences and structures. Journal of molecular biology 247.4 (1995): 536-540
5. C. A. Orengo, et al. CATH–a hierarchic classification of protein domain structures. Structure 5.8 (1997): 1093-1109.
6. S. Velankar et al. SIFTS: structure integration with function, taxonomy and sequences resource. Nucleic Acids Research, 41(D1): D483–D489, 2013.
7. The UniProt Consortium. The universal protein resource (UniProt) in 2010. Nucleic Acids Research, 38(suppl 1): D142–D148, 2010
8. A. Mitchell et al. The InterPro protein families database : the classification resource after 15 years. Nucleic Acids Research, 43(D1): D213–D221, 2015.
9. T. E. Lewis, et al. Genome3D: a UK collaborative project to annotate genomic sequences with predicted 3D structures based on SCOP and CATH domains. Nucleic acids research 41.D1 (2013): D499-D507.

ACKNOWLEDGMENTS

This project is funded by the Agence Nationale de la Recherche (grant reference ANR-11-MONU-006-02), Inria and the Lorraine Region.

CONTACT

Zia Alborzi
INRIA Nancy Grand Est, LORIA, Bureau B133, France
Email: seyed-ziaeddin.alborzi@inria.fr
Website: www.loria.fr/~salborzi
Phone: +33 (0)7 83 29 89 83

Automatic Generation of Functional Annotation Rules Using Inferred GO-Domain Associations

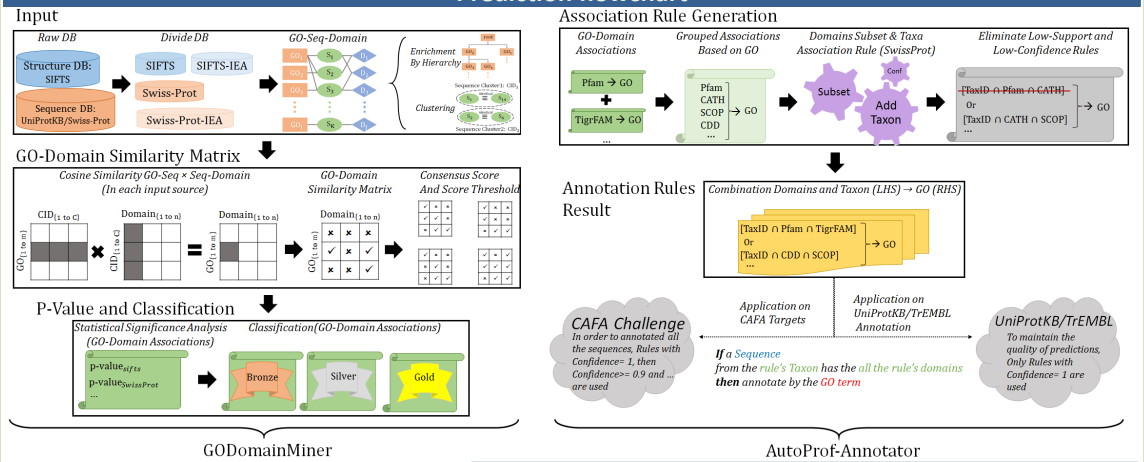
Seyed Ziaeddin Alborzi[1,3], Sabeur Aridhi[3], Marie-Dominique Devignes[2,3] Rabie Saidi[4], Alexandre Renaux[4], Maria J. Martin[4], David W. Ritchie[1]

[1]INRIA Nancy Grand-Est, [2]CNRS Nancy, [3]Université de Lorraine, [4]European Bioinformatics Institute

Introduction

The GO ontology is widely used for functional annotation of genes and proteins. It describes biological processes (BP), molecular function (MF), and cellular components (CC) in three distinct hierarchical controlled vocabularies. At the molecular level, functions are often performed by highly conserved parts of proteins, identified by sequence or structure alignments and classified into domains or families (SCOP, CATH, PFAM, TIGRFAMs, etc.). The InterPro database provides a valuable integrated classification of protein sequences and domains which is linked to nearly all existing other classifications. Interestingly, several InterPro families have been manually annotated with GO terms using expert knowledge and the literature. However, the list of such annotations is incomplete (only 20% of Pfam domains and families possess MF GO functional annotation). We therefore developed the GODomainMiner approach to expand the available functional annotations of protein domains and families (1). Based on our ECDomainMiner approach (2), we use the respective associations of protein sequences with GO terms and protein domains to infer direct associations between GO terms and protein domains. Finally, we used our calculated GO-Domain associations to devise a systematic way, called AutoProf-Annotator, to generate high confidence rules for protein sequence (or structure) annotation.

Prediction flowchart



Rules Statistics

AR Confidence > 0.5	Molecular Function	Biological Process	Cellular Component
Combination of Domains	1,723,497	1,841,000	1,543,333
Distinct Taxon	8,337	8,237	8,276
Prediction Rules	4,705	11,676	1,870

Table 1. Numbers of rules and the combination of domains result in the all rules.

AR Confidence = 1	Molecular Function	Biological Process	Cellular Component
Combination of Domains	1,692,547	1,826,347	1,4,96,772
Distinct Taxon	8,332	7966	8,266
Prediction Rules	4,673	11,582	1,853

Association Rule Samples

Rule (Confidence = 1)
- *{{{PF02423 ∩ CATH:3.30.1780.10} ∩ Mammalia} → GO:0047127)*
- PF02423: Ornithine cyclodeaminase/mu-crystallin family.
- CATH: 3.30.1780.10: Ornithine cyclodeaminase.
- MF GO:0047127: hiomorpholine-carboxylate dehydrogenase.

UniProtKB/Swiss-Prot:
- Hits: 5 sequences, all are annotated with the GO term.

UniProtKB/TrEMBL Annotation
- Hits: 47 Sequences.
- 1 Sequence is annotated with the GO term.
- 7 Sequence are annotated with ancestors of the GO term (General)
- 39 Sequence are annotated by AutoProf-Annotator

Rule (Confidence = 1)
- *{{{CD01399} ∩ Proteobacteria} → GO:0046348)*
- *CD01399* : GlcN6P_deaminase.
- BP GO:0046348 : amino sugar catabolic process.

UniProtKB/Swiss-Prot:
- Hits: 103 sequences, all are annotated with the GO term.

UniProtKB/TrEMBL Annotation
- Hits: 1930 Sequences.
- 1171 Sequence is annotated with the GO term.
- 569 Sequence are annotated with ancestors of the GO term (General)
- 190 Sequence are annotated by AutoProf-Annotator

Summary of CAFA Challenge Predictions

Molecular Function GO | Biological Process GO | Cellular Component GO

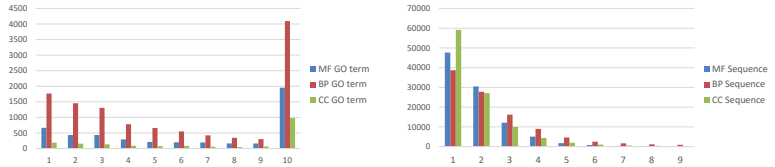
Table 2. GO function prediction for 130,000 CAFA Targets.

	CAFA Targets	CAFA Targets (Conf. 1)
Prediction	188,549	164,359
Sequence	98,849	81,248
GO term	4,705	4,673
Common to existing GO terms	ISMB/ECCB Function SIG	ISMB/ECCB Function SIG

	CAFA Targets	CAFA Targets (Conf. 1)
Prediction	315,310	229,006
Sequence	106,346	72,543
GO term	11,676	11,582
Common to existing GO terms	ISMB/ECCB Function SIG	ISMB/ECCB Function SIG

	CAFA Targets	CAFA Targets (Conf. 1)
Prediction	191,835	150,411
Sequence	105,274	76,233
GO term	1,870	1,853
Common to existing GO terms	ISMB/ECCB Function SIG	ISMB/ECCB Function SIG

Fig2. CAFA3: Distribution according to the number of GO terms for each sequence (right), and sequences for each GO term (left)



Annotation Examples

PRS56_Human is a target of CAFA3
- Very well annotated protein sequence in UniProtKB/Swiss-Prot
- Annotation Score: 5 - Experimental evidence at protein level
- Existing information in UniProtKB/Swiss-Prot:
  - MF **GO:0004252**
  - BP GO:0043010
  - BP GO:0006508
  - CC GO:0005783
- AutoProf-Annotator predicts following GO terms:
  - MF **GO:0004252** (Exact Match) (Conf. = 1)
  - BP GO:0044699 (Ancestor of GO:0043010) (Conf. = 0.6)
  - BP GO:0019538 (Parent of GO:0006508) (Conf. = 0.7)
  - CC GO:0044464 (Ancestor of GO:0005783) (Conf. = 0.7)

6PGL_SALCH is a target of CAFA3
- Annotated protein sequence in UniProtKB/Swiss-Prot
- Annotation Score: 2 – Protein inferred from homology
- Existing information in UniProtKB/Swiss-Prot :
  - MF **GO:0017057**
  - BP GO:0006006
  - BP GO:0006508
- AutoProf-Annotator predicts following GO terms:
  - MF **GO:0017057** (Exact Match) (Conf. = 1)
  - BP GO:0006006 (Exact Match) (Conf. = 1)
  - BP GO:0006508 (Exact Match) (Conf. = 1)
  - CC GO:0042597 (New Prediction) (Conf. = 0.9)

CONCLUSION

Our GODomainMiner approach provides a substantial enrichment of functional annotations at the protein domain level which has been exploited to develop a novel system here called AutoProf-Annotator for protein functional annotation. We used the AutoProf-Annotator to annotate target sequences in CAFA challenge.

LITERATURE CITED

1. Alborzi, Seyed Ziaeddin, Marie-Dominique Devignes, and David W. Ritchie. "Associating Gene Ontology Terms with Pfam Protein Domains." International Conference on Bioinformatics and Biomedical Engineering. Springer, Cham, 2017.
2. Alborzi, Seyed Ziaeddin, Marie-Dominique Devignes, and David W. Ritchie. "ECDomainMiner: discovering hidden associations between enzyme commission numbers and Pfam domains." BMC bioinformatics 18.1 (2017): 107.

ACKNOWLEDGMENTS

This project is funded by the Agence Nationale de la Recherche (grant reference ANR-11-MONU-006-02), Inria and the Lorraine Region. Travel to the meeting was made possible, in part, by a travel award from the NSF.

Contact INRIA Nancy Grand Est, LORIA, Bureau B133, France *Email*: seyed-ziaeddin.alborzi@inria.fr *Website*: https://members.loria.fr/SAlborzi/ *Phone*: +33 (0)7 83 29 89 83

Bibliography

[Afantenos et al., 2005] Afantenos, S., Karkaletsis, V., and Stamatopoulos, P. (2005). Summarization from medical documents: a survey. *Artificial intelligence in medicine*, 33(2):157–177.

[Agrawal et al., 1993] Agrawal, R., Imieliński, T., and Swami, A. (1993). Mining association rules between sets of items in large databases. In *Acm sigmod record*, volume 22, pages 207–216. ACM.

[Agrawal et al., 1994] Agrawal, R., Srikant, R., et al. (1994). Fast algorithms for mining association rules. In *Proc. 20th int. conf. very large data bases, VLDB*, volume 1215, pages 487–499.

[Alborzi et al., 2017a] Alborzi, S. Z., Aridhi, S., Devignes, M.-D., Saidi, R., Renaux, A., Martin, M. J., and Ritchie, D. W. (2017a). Automatic Generation of Functional Annotation Rules Using Inferred GO-Domain Associations. In *Function-SIG ISMB/ECCB 2017*.

[Alborzi et al., 2015] Alborzi, S. Z., Devignes, M.-D., and Ritchie, D. W. (2015). EC-PSI: Associating Enzyme Commission Numbers with Pfam Domains. *bioRxiv*.

[Alborzi et al., 2017b] Alborzi, S. Z., Devignes, M.-D., and Ritchie, D. W. (2017b). Associating Gene Ontology Terms with Pfam Protein Domains. In *International Conference on Bioinformatics and Biomedical Engineering*, pages 127–138. Springer.

[Alborzi et al., 2017c] Alborzi, S. Z., Devignes, M.-D., and Ritchie, D. W. (2017c). ECDomainMiner: discovering hidden associations between enzyme commission numbers and Pfam domains. *BMC bioinformatics*, 18(1):107.

[Alborzi et al., 2014] Alborzi, S. Z., Maduranga, D., Fan, R., Rajapakse, J. C., and Zheng, J. (2014). CUDAGRN: Parallel speedup of inferring large gene regulatory networks from expression data using random forest. In *IAPR International Conference on Pattern Recognition in Bioinformatics*, pages 85–97. Springer.

[Alborzi et al., 2012] Alborzi, S. Z., Raji, F., Saraee, M. H., et al. (2012). Privacy preserving mining of association rules on horizontally distributed databases. In *2012 International Conference on Software and Computer Applications (ICSCA 2012)*, pages 158–164. IACSIT Press, Singapore.

[Alborzi et al., 2016] Alborzi, S. Z., Ritchie, D. W., and Devignes, M.-D. (2016). Using Content-Based Filtering to Infer Direct Associations between the CATH, Pfam, and SCOP Domain Databases. In *ECCB 2016*.

[Alborzi et al., 2018] Alborzi, S. Z., Ritchie, D. W., and Devignes, M.-D. (2018). Computational Discovery of Direct Associations between GO terms and Protein Domains. *BMC bioinformatics*.

[Altschul et al., 1997] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped blast and psi-blast: a new generation of protein database search programs. *Nucleic acids research*, 25(17):3389–3402.

[Andreeva et al., 2013] Andreeva, A., Howorth, D., Chothia, C., Kulesha, E., and Murzin, A. G. (2013). SCOP2 prototype: a new approach to protein structure mining. *Nucleic acids research*, 42(D1):D310–D314.

[Apic et al., 2001] Apic, G., Gough, J., and Teichmann, S. A. (2001). Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *Journal of molecular biology*, 310(2):311–325.

[Apweiler et al., 2001] Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Birney, E., Biswas, M., Bucher, P., Cerutti, L., Corpet, F., Croning, M. D., et al. (2001). The interpro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic acids research*, 29(1):37–40.

[Apweiler et al., 2010] Apweiler, R. et al. (2010). The universal protein resource (uniprot) in 2010. *Nucleic acids research*, 38(suppl 1):D142–D148.

[Apweiler et al., 2017] Apweiler, R. et al. (2017). Uniprot: the universal protein knowledgebase. *Nucleic acids research*, 45(D1):D158–D169.

[Ashburner et al., 2000] Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., et al. (2000). Gene ontology: tool for the unification of biology. *Nature genetics*, 25(1):25–29.

[Asratian et al., 1998] Asratian, A. S., Denley, T. M., and Häggkvist, R. (1998). *Bipartite graphs and their applications*, volume 131. Cambridge University Press.

[Attwood et al., 2003] Attwood, T. K., Bradley, P., Flower, D. R., Gaulton, A., Maudling, N., Mitchell, A. L., Moulton, G., Nordle, A., Paine, K., Taylor, P., et al. (2003). Prints and its automatic supplement, preprints. *Nucleic acids research*, 31(1):400–402.

[Bairoch, 2000] Bairoch, A. (2000). The enzyme database in 2000. *Nucleic acids research*, 28(1):304–305.

[Balabanović and Shoham, 1997] Balabanović, M. and Shoham, Y. (1997). Fab: content-based, collaborative recommendation. *Communications of the ACM*, 40(3):66–72.

[Banerjee et al., 2009] Banerjee, A., Chitnis, U., Jadhav, S., Bhawalkar, J., and Chaudhury, S. (2009). Hypothesis testing, type i and type ii errors. *Industrial psychiatry journal*, 18(2):127.

[Bashton and Chothia, 2007] Bashton, M. and Chothia, C. (2007). The generation of new protein functions by the combination of domains. *Structure*, 15(1):85–99.

[Bateman et al., 2002] Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S. R., Griffiths-Jones, S., Howe, K. L., Marshall, M., and Sonnhammer, E. L. (2002). The pfam protein families database. *Nucleic acids research*, 30(1):276–280.

[Bateman et al., 2014] Bateman, A. et al. (2014). Uniprot: a hub for protein information. *Nucleic acids research*, page gku989.

[Bauer and Kuster, 2003] Bauer, A. and Kuster, B. (2003). Affinity purification-mass spectrometry. *The FEBS Journal*, 270(4):570–578.

[Ben-Hur et al., 2008] Ben-Hur, A., Ong, C. S., Sonnenburg, S., Schölkopf, B., and Rätsch, G. (2008). Support vector machines and kernels for computational biology. *PLoS computational biology*, 4(10):e1000173.

[Benabderrahmane et al., 2010] Benabderrahmane, S., Smail-Tabbone, M., Poch, O., Napoli, A., and Devignes, M.-D. (2010). Intelligo: a new vector-based semantic similarity measure including annotation origin. *BMC bioinformatics*, 11(1):588.

[Benjamini and Hochberg, 1995] Benjamini, Y. and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the royal statistical society. Series B (Methodological)*, pages 289–300.

[Benson et al., 2017] Benson, D. A., Cavanaugh, M., Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., and Sayers, E. W. (2017). Genbank. *Nucleic acids research*, 45(Database issue):D37.

[Berg et al., 2002] Berg, J. M., Tymoczko, J. L., and Stryer, L. (2002). Protein structure and function. *W.H. Freeman*.

[Berkhin et al., 2006] Berkhin, P. et al. (2006). A survey of clustering data mining techniques. *Grouping multidimensional data*, 25:71.

[Berman et al., 2006] Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2006). The protein data bank, 1999–. In *International Tables for Crystallography Volume F: Crystallography of biological macromolecules*, pages 675–684. Springer.

[Bernstein et al., 1977] Bernstein, F. C., Koetzle, T. F., Williams, G. J., Meyer, E. F., Brice, M. D., Rodgers, J. R., Kennard, O., Shimanouchi, T., and Tasumi, M. (1977). The protein data bank. *European Journal of Biochemistry*, 80(2):319–324.

[Bhaskara and Srinivasan, 2011] Bhaskara, R. M. and Srinivasan, N. (2011). Stability of domain structures in multi-domain proteins. *Scientific reports*, 1.

[Binns et al., 2009] Binns, D., Dimmer, E., Huntley, R., Barrell, D., O'donovan, C., and Apweiler, R. (2009). Quickgo: a web-based tool for gene ontology searching. *Bioinformatics*, 25(22):3045–3046.

[Bollobas, 2012] Bollobas, B. (2012). *Graph theory: an introductory course*, volume 63. Springer Science & Business Media.

[Bordner and Abagyan, 2005] Bordner, A. J. and Abagyan, R. (2005). Statistical analysis and prediction of protein–protein interfaces. *Proteins: Structure, Function, and Bioinformatics*, 60(3):353–366.

[Bork et al., 1998] Bork, P., Dandekar, T., Diaz-Lazcoz, Y., Eisenhaber, F., Huynen, M., and Yuan, Y. (1998). Predicting function: from genes to genomes and back. *Journal of molecular biology*, 283(4):707–725.

[Boudellioua et al., 2016] Boudellioua, I., Saidi, R., Hoehndorf, R., Martin, M. J., and Solovyev, V. (2016). Prediction of metabolic pathway involvement in prokaryotic UniProtKB data by association rule mining. *PloS one*, 11(7):e0158896.

[Boutet et al., 2007] Boutet, E., Lieberherr, D., Tognolli, M., Schneider, M., and Bairoch, A. (2007). Uniprotkb/swiss-prot: the manually annotated section of the uniprot knowledgebase. *Plant bioinformatics: methods and protocols*, pages 89–112.

[Browne et al., 2007] Browne, F., Wang, H., Zheng, H., and Azuaje, F. (2007). Supervised statistical and machine learning approaches to inferring pairwise and module-based protein interaction networks. In *Bioinformatics and Bioengineering, 2007. BIBE 2007. Proceedings of the 7th IEEE International Conference on*, pages 1365–1369. IEEE.

[Brückner et al., 2009] Brückner, A., Polge, C., Lentze, N., Auerbach, D., and Schlattner, U. (2009). Yeast two-hybrid, a powerful tool for systems biology. *International journal of molecular sciences*, 10(6):2763–2788.

[Burke, 2002] Burke, R. (2002). Hybrid recommender systems: Survey and experiments. *User modeling and user-adapted interaction*, 12(4):331–370.

[Bursteinas et al., 2016] Bursteinas, B., Britto, R., Bely, B., Auchincloss, A., Rivoire, C., Redaschi, N., O'Donovan, C., and Martin, M. J. (2016). Minimizing proteome redundancy in the uniprot knowledgebase. *Database*, 2016.

[Carbon et al., 2008] Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., Lewis, S., Hub, A., and Group, W. P. W. (2008). Amigo: online access to ontology and annotation data. *Bioinformatics*, 25(2):288–289.

[Chandola et al., 2009] Chandola, V., Banerjee, A., and Kumar, V. (2009). Anomaly detection: A survey. *ACM computing surveys (CSUR)*, 41(3):15.

[Chapelle et al., 2003] Chapelle, O., Weston, J., and Schölkopf, B. (2003). Cluster kernels for semi-supervised learning. In *Advances in neural information processing systems*, pages 601–608.

[Chawla et al., 2004] Chawla, N. V., Japkowicz, N., and Kotcz, A. (2004). Special issue on learning from imbalanced data sets. *ACM Sigkdd Explorations Newsletter*, 6(1):1–6.

[Chen et al., 2012] Chen, C., Zhao, J.-F., Huang, Q., Wang, R.-S., and Zhang, X.-S. (2012). Inferring domain-domain interactions from protein-protein interactions in the complex network conformation. *BMC systems biology*, 6(1):S7.

[Chen et al., 2011] Chen, H.-H., Gou, L., Zhang, X., and Giles, C. L. (2011). Collabseer: a search engine for collaboration discovery. In *Proceedings of the 11th annual international ACM/IEEE joint conference on Digital libraries*, pages 231–240. ACM.

[Chen and Liu, 2005] Chen, X.-W. and Liu, M. (2005). Prediction of protein–protein interactions using random decision forest framework. *Bioinformatics*, 21(24):4394–4400.

[Cheng and Perocchi, 2015] Cheng, Y. and Perocchi, F. (2015). Protphylo: identification of protein–phenotype and protein–protein functional associations via phylogenetic profiling. *Nucleic acids research*, 43(W1):W160–W168.

[Chothia, 1992] Chothia, C. (1992). One thousand families for the molecular biologist. *Nature*, 357(6379):543–544.

[Chothia and Lesk, 1986] Chothia, C. and Lesk, A. M. (1986). The relation between the divergence of sequence and structure in proteins. *EMBO Journal*, 5(4):823.

[Chunjie et al., 2017] Chunjie, L., Qiang, Y., et al. (2017). Cosine normalization: Using cosine similarity instead of dot product in neural networks. *arXiv preprint arXiv:1702.05870*.

[Cooper et al., 2009] Cooper, T. A., Wan, L., and Dreyfuss, G. (2009). Rna and disease. *Cell*, 136(4):777–793.

[Cui et al., 2003] Cui, X., Churchill, G. A., et al. (2003). Statistical tests for differential expression in cDNA microarray experiments. *Genome Biol*, 4(4):210.

[Dandekar et al., 1998] Dandekar, T., Snel, B., Huynen, M., and Bork, P. (1998). Conservation of gene order: a fingerprint of proteins that physically interact. *Trends in biochemical sciences*, 23(9):324–328.

[Daraselia et al., 2004] Daraselia, N., Yuryev, A., Egorov, S., Novichkova, S., Nikitin, A., and Mazo, I. (2004). Extracting human protein interactions from medline using a full-sentence parser. *Bioinformatics*, 20(5):604–611.

[Davidson et al., 2010] Davidson, J., Liebald, B., Liu, J., Nandy, P., Van Vleet, T., Gargi, U., Gupta, S., He, Y., Lambert, M., Livingston, B., et al. (2010). The youtube video recommendation system. In *Proceedings of the fourth ACM conference on Recommender systems*, pages 293–296. ACM.

[de Beer et al., 2014] de Beer, T. A. P., Berka, K., Thornton, J. M., and Laskowski, R. A. (2014). PDBsum additions. *Nucleic Acids Research*, 42(D1):D292–D296.

[Deng et al., 2002] Deng, M., Mehta, S., Sun, F., and Chen, T. (2002). Inferring domain–domain interactions from protein–protein interactions. *Genome research*, 12(10):1540–1548.

[Doğan et al., 2016] Doğan, T., MacDougall, A., Saidi, R., Poggioli, D., Bateman, A., O'Donovan, C., and Martin, M. J. (2016). Uniprot-daac: domain architecture alignment and classification, a new method for automatic functional annotation in uniprotkb. *Bioinformatics*, 32(15):2264–2271.

[Donaldson et al., 2003] Donaldson, I., Martin, J., De Bruijn, B., Wolting, C., Lay, V., Tuekam, B., Zhang, S., Baskin, B., Bader, G. D., Michalickova, K., et al. (2003). Prebind and textomy–mining the biomedical literature for protein-protein interactions using a support vector machine. *BMC bioinformatics*, 4(1):11.

[Eddy, 1998] Eddy, S. R. (1998). Profile hidden Markov models. *Bioinformatics (Oxford, England)*, 14(9):755–763.

[Enault et al., 2005] Enault, F., Suhre, K., and Claverie, J.-M. (2005). Phydbac" gene function predictor": a gene annotation tool based on genomic context analysis. *BMC bioinformatics*, 6(1):247.

[Engelhardt et al., 2005] Engelhardt, B. E., Jordan, M. I., Muratore, K. E., and Brenner, S. E. (2005). Protein molecular function prediction by bayesian phylogenomics. *PLoS computational biology*, 1(5):e45.

[Fang and Gough, 2013] Fang, H. and Gough, J. (2013). dcGO: database of domain-centric ontologies on functions, phenotypes, diseases and more. *Nucleic acids research*, 41(D1):D536–D544.

[Fariselli et al., 2002] Fariselli, P., Pazos, F., Valencia, A., and Casadio, R. (2002). Prediction of protein–protein interaction sites in heterocomplexes with neural networks. *The FEBS Journal*, 269(5):1356–1361.

[Fawcett, 2006] Fawcett, T. (2006). An introduction to ROC analysis. *Pattern recognition letters*, 27(8):861–874.

[Fayyad et al., 1996] Fayyad, U., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI magazine*, 17(3):37.

[Felfernig et al., 2007] Felfernig, A., Isak, K., Szabo, K., and Zachar, P. (2007). The vita financial services sales support environment. In *Proceedings of the National Conference on Artificial Intelligence*, volume 22, page 1692. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999.

[Feuermann et al., 2016] Feuermann, M., Gaudet, P., Mi, H., Lewis, S. E., and Thomas, P. D. (2016). Large-scale inference of gene function through phylogenetic annotation of gene ontology terms: case study of the apoptosis and autophagy cellular processes. *Database*, 2016(0):baw155.

[Finn et al., 2016a] Finn, R. D., Attwood, T. K., Babbitt, P. C., Bateman, A., Bork, P., Bridge, A. J., Chang, H.-Y., Dosztányi, Z., El-Gebali, S., Fraser, M., et al. (2016a). Interpro in 2017—beyond protein family and domain annotations. *Nucleic acids research*, 45(D1):D190–D199.

[Finn et al., 2016b] Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., et al. (2016b). The pfam protein families database: towards a more sustainable future. *Nucleic acids research*, 44(D1):D279–D285.

[Finn et al., 2013] Finn, R. D., Miller, B. L., Clements, J., and Bateman, A. (2013). ipfam: a database of protein family and domain interactions found in the protein data bank. *Nucleic acids research*, 42(D1):D364–D373.

[Fleischmann et al., 2004] Fleischmann, A., Darsow, M., Degtyarenko, K., Fleischmann, W., Boyce, S., Axelsen, K. B., Bairoch, A., Schomburg, D., Tipton, K. F., and Apweiler, R. (2004). Intenz, the integrated relational enzyme database. *Nucleic acids research*, 32(suppl_1):D434–D437.

[Fogel, 2008] Fogel, G. B. (2008). Computational intelligence approaches for pattern discovery in biological systems. *Briefings in bioinformatics*, 9(4):307–316.

[Forslund and Sonnhammer, 2008] Forslund, K. and Sonnhammer, E. L. (2008). Predicting protein function from domain content. *Bioinformatics*, 24(15):1681–1687.

[Frawley et al., 1992] Frawley, W. J., Piatetsky-Shapiro, G., and Matheus, C. J. (1992). Knowledge discovery in databases: An overview. *AI magazine*, 13(3):57.

[Friedberg, 2006] Friedberg, I. (2006). Automated protein function prediction—the genomic challenge. *Briefings in bioinformatics*, 7(3):225–242.

[Gavin et al., 2002] Gavin, A.-C., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J. M., Michon, A.-M., Cruciat, C.-M., et al. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868):141–147.

[George et al., 2004] George, R. A., Spriggs, R. V., Thornton, J. M., Al-Lazikani, B., and Swindells, M. B. (2004). SCOPEC: a database of protein catalytic domains. *Bioinformatics*, 20(suppl 1):i130–i136.

[Ghoorah et al., 2011] Ghoorah, A. W., Devignes, M.-D., Smaïl-Tabbone, M., and Ritchie, D. W. (2011). Spatial clustering of protein binding sites for template based protein docking. *Bioinformatics*, 27(20):2820–2827.

[Ghoorah et al., 2013a] Ghoorah, A. W., Devignes, M.-D., Smaïl-Tabbone, M., and Ritchie, D. W. (2013a). Kbdock 2013: a spatial classification of 3d protein domain family interactions. *Nucleic acids research*, 42(D1):D389–D395.

[Ghoorah et al., 2013b] Ghoorah, A. W., Devignes, M.-D., Smaïl-Tabbone, M., and Ritchie, D. W. (2013b). Protein docking using case-based reasoning. *Proteins: Structure, Function, and Bioinformatics*, 81(12):2150–2158.

[Godsil and Royle, 2013] Godsil, C. and Royle, G. F. (2013). *Algebraic graph theory*, volume 207. Springer Science & Business Media.

[Goldberg and Roth, 2003] Goldberg, D. S. and Roth, F. P. (2003). Assessing experimentally derived interactions in a small world. *Proceedings of the National Academy of Sciences*, 100(8):4372–4376.

[Gomez-Uribe and Hunt, 2016] Gomez-Uribe, C. A. and Hunt, N. (2016). The netflix recommender system: Algorithms, business value, and innovation. *ACM Transactions on Management Information Systems (TMIS)*, 6(4):13.

[González and Liao, 2010] González, A. J. and Liao, L. (2010). Predicting domain-domain interaction based on domain profiles with feature selection and support vector machines. *BMC bioinformatics*, 11(1):537.

[Guimarães et al., 2006] Guimarães, K. S., Jothi, R., Zotenko, E., and Przytycka, T. M. (2006). Predicting domain-domain interactions using a parsimony approach. *Genome biology*, 7(11):R104.

[Guo et al., 2008] Guo, Y., Yu, L., Wen, Z., and Li, M. (2008). Using support vector machine combined with auto covariance to predict protein–protein interactions from protein sequences. *Nucleic acids research*, 36(9):3025–3030.

[Gutmanas et al., 2014] Gutmanas, A., Alhroub, Y., Battle, G. M., Berrisford, J. M., Bochet, E., Conroy, M. J., Dana, J. M., Montecelo, M. A. F., van Ginkel, G., Gore, S. P., Haslam, P., Hatherley, R., Hendrickx, P. M. S., Hirshberg, M., Lagerstedt, I., Mir, S., Mukhopadhyay, A., Oldfield, T. J., Patwardhan, A., Rinaldi, L., Sahni, G., Sanz-García, E., Sen, S., Slowley, R. A., Velankar, S., Wainwright, and J., M. E. K. G. (2014). PDBe: protein data bank in europe. *Nucleic Acids Research*, 42(D1):D285–D291.

[Hadley and Jones, 1999] Hadley, C. and Jones, D. T. (1999). A systematic comparison of protein structure classifications: SCOP, CATH and FSSP. *Structure*, 7(9):1099–1112.

[Haft et al., 2012] Haft, D. H., Selengut, J. D., Richter, R. A., Harkins, D., Basu, M. K., and Beck, E. (2012). Tigrfams and genome properties in 2013. *Nucleic acids research*, 41(D1):D387–D395.

[Haft et al., 2003] Haft, D. H., Selengut, J. D., and White, O. (2003). The tigrfams database of protein families. *Nucleic acids research*, 31(1):371–373.

[Han et al., 2011] Han, J., Pei, J., and Kamber, M. (2011). *Data mining: concepts and techniques*. Elsevier.

[Han et al., 2000] Han, J., Pei, J., and Yin, Y. (2000). Mining frequent patterns without candidate generation. In *ACM sigmod record*, volume 29, pages 1–12. ACM.

[Hanani et al., 2001] Hanani, U., Shapira, B., and Shoval, P. (2001). Information filtering: Overview of issues, research and systems. *User Modeling and User-Adapted Interaction*, 11(3):203–259.

[Harris et al., 2004] Harris, M. et al. (2004). The gene ontology (go) database and informatics resource. *Nucleic acids research*, 32(suppl 1):D258–D261.

[Hipp et al., 2000] Hipp, J., Güntzer, U., and Nakhaeizadeh, G. (2000). Algorithms for association rule mining—a general survey and comparison. *ACM sigkdd explorations newsletter*, 2(1):58–64.

[Hooper et al., 2014] Hooper, C. M., Tanz, S. K., Castleden, I. R., Vacher, M. A., Small, I. D., and Millar, A. H. (2014). Subacon: a consensus algorithm for unifying the subcellular localization data of the arabidopsis proteome. *Bioinformatics*, 30(23):3356–3364.

[Hsin Liu et al., 2012] Hsin Liu, C., Li, K.-C., and Yuan, S. (2012). Human protein–protein interaction prediction by a novel sequence-based co-evolution method: co-evolutionary divergence. *Bioinformatics*, 29(1):92–98.

[Huang et al., 2014] Huang, L., Li, F., Sheng, J., Xia, X., Ma, J., Zhan, M., and Wong, S. T. (2014). DrugComboRanker: drug combination discovery based on target network analysis. *Bioinformatics*, 30(12):i228–i236.

[Huang et al., 2004] Huang, M., Zhu, X., Hao, Y., Payan, D. G., Qu, K., and Li, M. (2004). Discovering patterns to extract protein–protein interactions from full texts. *Bioinformatics*, 20(18):3604–3612.

[Hulo et al., 2007] Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., Cuche, B. A., De Castro, E., Lachaize, C., Langendijk-Genevaux, P. S., and Sigrist, C. J. (2007). The 20 years of prosite. *Nucleic acids research*, 36(suppl_1):D245–D249.

[Isono and Schwechheimer, 2010] Isono, E. and Schwechheimer, C. (2010). Co-immunoprecipitation and protein blots. *Plant Developmental Biology: Methods and Protocols*, pages 377–387.

[Jeong et al., 2001] Jeong, H., Mason, S. P., Barabási, A.-L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, 411(6833):41–42.

[Jones et al., 2014] Jones, P., Binns, D., Chang, H.-Y., Fraser, M., Li, W., McAnulla, C., McWilliam, H., Maslen, J., Mitchell, A., Nuka, G., et al. (2014). Interproscan 5: genome-scale protein function classification. *Bioinformatics*, 30(9):1236–1240.

[Jothi et al., 2006] Jothi, R., Cherukuri, P. F., Tasneem, A., and Przytycka, T. M. (2006). Co-evolutionary analysis of domains in interacting proteins reveals insights into domain–domain interactions mediating protein–protein interactions. *Journal of molecular biology*, 362(4):861–875.

[Kanehisa et al., 2016] Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., and Tanabe, M. (2016). Kegg as a reference resource for gene and protein annotation. *Nucleic acids research*, 44(D1):D457–D462.

[Kenworthy, 2001] Kenworthy, A. K. (2001). Imaging protein-protein interactions using fluorescence resonance energy transfer microscopy. *Methods*, 24(3):289–296.

[Kerrien et al., 2006] Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuermann, M., Friedrichsen, A., Huntley, R., et al. (2006). Intact—open source resource for molecular interaction data. *Nucleic acids research*, 35(suppl_1):D561–D565.

[Kerrien et al., 2011] Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., Duesbury, M., Dumousseau, M., Feuermann, M., Hinz, U., et al. (2011). The intact molecular interaction database in 2012. *Nucleic acids research*, 40(D1):D841–D846.

[Khor, 2014] Khor, S. (2014). Inferring domain-domain interactions from protein-protein interactions with formal concept analysis. *PloS one*, 9(2):e88943.

[Kobe et al., 2008] Kobe, B., Guncar, G., Buchholz, R., Huber, T., Maco, B., Cowieson, N., Martin, J. L., Marfori, M., and Forwood, J. K. (2008). Crystallography and protein–protein interactions: biological interfaces and crystal contacts.

[Koren and Bell, 2015] Koren, Y. and Bell, R. (2015). Advances in collaborative filtering. In *Recommender systems handbook*, pages 77–118. Springer.

[Kotsiantis and Pintelas, 2009] Kotsiantis, S. and Pintelas, P. (2009). Predictive data mining: A survey of regression methods. In *Encyclopedia of Information Science and Technology, Second Edition*, pages 3105–3110. IGI Global.

[Laskowski, 2001] Laskowski, R. A. (2001). PDBsum: summaries and analyses of PDB structures. *Nucleic Acids Research*, 29(1):221–222.

[Lee et al., 2007] Lee, D., Redfern, O., and Orengo, C. (2007). Predicting protein function from sequence and structure. *Nature reviews. Molecular cell biology*, 8(12):995.

[Lee et al., 2006] Lee, H., Deng, M., Sun, F., and Chen, T. (2006). An integrated approach to the prediction of domain-domain interactions. *BMC bioinformatics*, 7(1):269.

[Leinonen et al., 2004] Leinonen, R., Diez, F. G., Binns, D., Fleischmann, W., Lopez, R., and Apweiler, R. (2004). Uniprot archive. *Bioinformatics*, 20(17):3236–3237.

[Letunic et al., 2014] Letunic, I., Doerks, T., and Bork, P. (2014). Smart: recent updates, new developments and status in 2015. *Nucleic acids research*, 43(D1):D257–D260.

[Li et al., 2007] Li, J., Halgamuge, S. K., Kells, C. I., and Tang, S.-L. (2007). Gene function prediction based on genomic context clustering and discriminative learning: an application to bacteriophages. *BMC bioinformatics*, 8(4):S6.

[Li and Godzik, 2006] Li, W. and Godzik, A. (2006). Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. *Bioinformatics*, 22(13):1658–1659.

[Li et al., 2001] Li, W., Jaroszewski, L., and Godzik, A. (2001). Clustering of highly homologous sequences to reduce the size of large protein databases. *Bioinformatics*, 17(3):282–283.

[Li et al., 2013] Li, X., Ng, S.-K., and Wang, J. T. (2013). *Biological data mining and its applications in healthcare*, volume 8. World Scientific.

[Licata et al., 2011] Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., Sacco, F., Palma, A., Nardozza, A. P., Santonico, E., et al. (2011). Mint, the molecular interaction database: 2012 update. *Nucleic acids research*, 40(D1):D857–D861.

[Lin and Chen, 2013] Lin, X. and Chen, X.-w. (2013). Heterogeneous data integration by tree-augmented naïve bayes for protein–protein interactions prediction. *Proteomics*, 13(2):261–268.

[Lindley, 2000] Lindley, D. V. (2000). The philosophy of statistics. *Journal of the Royal Statistical Society: Series D (The Statistician)*, 49(3):293–337.

[Liolios et al., 2009] Liolios, K., Chen, I.-M. A., Mavromatis, K., Tavernarakis, N., Hugenholtz, P., Markowitz, V. M., and Kyrpides, N. C. (2009). The genomes on line database (gold) in 2009: status of genomic and metagenomic projects and their associated metadata. *Nucleic acids research*, 38(suppl_1):D346–D354.

[Liu et al., 2016] Liu, X., Yang, S., Li, C., Zhang, Z., and Song, J. (2016). Spar: a random forest-based predictor for self-interacting proteins with fine-grained domain information. *Amino acids*, 48(7):1655–1665.

[Lops et al., 2011] Lops, P., De Gemmis, M., and Semeraro, G. (2011). Content-based recommender systems: State of the art and trends. In *Recommender systems handbook*, pages 73–105. Springer.

[Mann et al., 2017] Mann, C. M., Muppirala, U. K., and Dobbs, D. (2017). Computational prediction of rna-protein interactions. *Promoter Associated RNA: Methods and Protocols*, pages 169–185.

[Marchler-Bauer et al., 2005] Marchler-Bauer, A., Anderson, J. B., Cherukuri, P. F., DeWeese-Scott, C., Geer, L. Y., Gwadz, M., He, S., Hurwitz, D. I., Jackson, J. D., Ke, Z., et al. (2005). Cdd: a conserved domain database for protein classification. *Nucleic acids research*, 33(suppl_1):D192–D196.

[Marchler-Bauer et al., 2016] Marchler-Bauer, A., Bo, Y., Han, L., He, J., Lanczycki, C. J., Lu, S., Chitsaz, F., Derbyshire, M. K., Geer, R. C., Gonzales, N. R., et al. (2016). Cdd/sparcle: functional classification of proteins via subfamily domain architectures. *Nucleic acids research*, 45(D1):D200–D203.

[Martin, 2004] Martin, A. C. R. (2004). PDBSprotEC: a web-accessible database linking PDB chains to EC numbers via SwissProt. *Bioinformatics*, 20(6):986–988.

[Martin et al., 1998] Martin, A. C. R., Orengo, C. A., Hutchinson, E. G., Jones, S., Karmirantzou, M., Laskowski, R. A., Mitchell, J. B. O., Taroni, C., and Thornton, J. M. (1998). Protein folds and functions. *Structure*, 6(7):875–884.

[Massjouni et al., 2006] Massjouni, N., Rivera, C. G., and Murali, T. (2006). Virgo: computational prediction of gene functions. *Nucleic acids research*, 34(suppl_2):W340–W344.

[Meyer et al., 2013] Meyer, M. J., Das, J., Wang, X., and Yu, H. (2013). Instruct: a database of high-quality 3d structurally resolved protein interactome networks. *Bioinformatics*, 29(12):1577–1579.

[Mi et al., 2017] Mi, H., Huang, X., Muruganujan, A., Tang, H., Mills, C., Kang, D., and Thomas, P. D. (2017). Panther version 11: expanded annotation data from gene ontology and reactome pathways, and data analysis tool enhancements. *Nucleic acids research*, 45(D1):D183–D189.

[Mittelhammer et al., 2000] Mittelhammer, R. C., Judge, G. G., and Miller, D. J. (2000). *Econometric Foundations Pack with CD-ROM*, volume 1. Cambridge University Press.

[Mogotsi, 2010] Mogotsi, I. (2010). Christopher d. manning, prabhakar raghavan, and hinrich schütze: Introduction to information retrieval.

[Morell et al., 2009] Morell, M., Ventura, S., and Avilés, F. X. (2009). Protein complementation assays: approaches for the in vivo analysis of protein interactions. *FEBS letters*, 583(11):1684–1691.

[Morgat et al., 2011] Morgat, A. et al. (2011). Ongoing and future developments at the universal protein resource. *Nucleic acids research*, 39(suppl 1):D214–D219.

[Mosca et al., 2013] Mosca, R., Céol, A., Stein, A., Olivella, R., and Aloy, P. (2013). 3did: a catalog of domain-based interactions of known three-dimensional structure. *Nucleic acids research*, 42(D1):D374–D379.

[Muflikhah and Baharudin, 2009] Muflikhah, L. and Baharudin, B. (2009). Document clustering using concept space and cosine similarity measurement. In *Computer Technology and Development, 2009. ICCTD'09. International Conference on*, volume 1, pages 58–62. IEEE.

[Murzin et al., 1995] Murzin, A. G., Brenner, S. E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of molecular biology*, 247(4):536–540.

[Nagarajan et al., 2013] Nagarajan, R., Ahmad, S., and Michael Gromiha, M. (2013). Novel approach for selecting the best predictor for identifying the binding sites in dna binding proteins. *Nucleic acids research*, 41(16):7606–7614.

[Nariai et al., 2007] Nariai, N., Kolaczyk, E. D., and Kasif, S. (2007). Probabilistic protein function prediction from heterogeneous genome-wide data. *PLoS One*, 2(3):e337.

[Nye et al., 2004] Nye, T. M., Berzuini, C., Gilks, W. R., Babu, M. M., and Teichmann, S. A. (2004). Statistical analysis of domains in interacting protein pairs. *Bioinformatics*, 21(7):993–1001.

[Omelchenko et al., 2010] Omelchenko, M. V., Galperin, M. Y., Wolf, Y. I., and Koonin, E. V. (2010). Non-homologous isofunctional enzymes: a systematic analysis of alternative solutions in enzyme evolution. *Biology direct*, 5(1):31.

[Orengo et al., 1997] Orengo, C. A., Michie, A., Jones, S., Jones, D. T., Swindells, M., and Thornton, J. M. (1997). CATH–a hierarchic classification of protein domain structures. *Structure*, 5(8):1093–1109.

[Oughtred et al., 2016] Oughtred, R., Chatr-aryamontri, A., Breitkreutz, B.-J., Chang, C. S., Rust, J. M., Theesfeld, C. L., Heinicke, S., Breitkreutz, A., Chen, D., Hirschman, J., et al. (2016). Biogrid: a resource for studying biological interactions in yeast. *Cold Spring Harbor Protocols*, 2016(1):pdb–top080754.

[Pagel et al., 2004] Pagel, P., Wong, P., and Frishman, D. (2004). A domain interaction map based on phylogenetic profiling. *Journal of molecular biology*, 344(5):1331–1346.

[Pearl et al., 2003] Pearl, F. M. G., Bennett, C., Bray, J. E., Harrison, A. P., Martin, N., Shepherd, A., Sillitoe, I., Thornton, J., and Orengo, C. A. (2003). The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic acids research*, 31(1):452–455.

[Pedruzzi et al., 2013] Pedruzzi, I., Rivoire, C., Auchincloss, A. H., Coudert, E., Keller, G., De Castro, E., Baratin, D., Cuche, B. A., Bougueleret, L., Poux, S., et al. (2013). HAMAP in 2013, new developments in the protein family classification and annotation system. *Nucleic acids research*, 41(D1):D584–D589.

[Peng et al., 2014] Peng, W., Wang, J., Cai, J., Chen, L., Li, M., and Wu, F.-X. (2014). Improving protein function prediction using domain and protein complexes in PPI networks. *BMC systems biology*, 8(1):35.

[Peri et al., 2003] Peri, S., Navarro, J. D., Amanchy, R., Kristiansen, T. Z., Jonnalagadda, C. K., Surendranath, V., Niranjan, V., Muthusamy, B., Gandhi, T., Gronborg, M., et al. (2003). Development of human protein reference database as an initial platform for approaching systems biology in humans. *Genome research*, 13(10):2363–2371.

[Phyu, 2009] Phyu, T. N. (2009). Survey of classification techniques in data mining. In *Proceedings of the International MultiConference of Engineers and Computer Scientists*, volume 1, pages 18–20.

[Pitre et al., 2008] Pitre, S., Alamgir, M., Green, J. R., Dumontier, M., Dehne, F., and Golshani, A. (2008). Computational methods for predicting protein–protein interactions. In *Protein–Protein Interaction*, pages 247–267. Springer.

[Punta and Ofran, 2008] Punta, M. and Ofran, Y. (2008). The rough guide to in silico function prediction, or how to use sequence and structure information to predict protein function. *PLoS computational biology*, 4(10):e1000160.

[Puton et al., 2012] Puton, T., Kozlowski, L., Tuszynska, I., Rother, K., and Bujnicki, J. M. (2012). Computational methods for prediction of protein–rna interactions. *Journal of structural biology*, 179(3):261–268.

[Radivojac et al., 2013] Radivojac, P., Clark, W. T., Oron, T. R., Schnoes, A. M., Wittkop, T., Sokolov, A., Graim, K., Funk, C., Verspoor, K., Ben-Hur, A., et al. (2013). A large-scale evaluation of computational protein function prediction. *Nature methods*, 10(3):221–227.

[Raghavachari et al., 2007] Raghavachari, B., Tasneem, A., Przytycka, T. M., and Jothi, R. (2007). Domine: a database of protein domain interactions. *Nucleic acids research*, 36(suppl_1):D656–D661.

[Raghavan and Wong, 1986] Raghavan, V. V. and Wong, S. M. (1986). A critical analysis of vector space model for information retrieval. *Journal of the American Society for information Science*, 37(5):279.

[Raza, 2012] Raza, K. (2012). Application of data mining in bioinformatics. *arXiv preprint arXiv:1205.1125*.

[Reichert et al., 2000] Reichert, J., Jabs, A., Slickers, P., and Sühnel, J. (2000). The IMB Jena image library of biological macromolecules. *Nucleic Acids Research*, 28(1):246–249.

[Rentzsch and Orengo, 2009] Rentzsch, R. and Orengo, C. A. (2009). Protein function prediction–the power of multiplicity. *Trends in biotechnology*, 27(4):210–219.

[Ricci et al., 2011] Ricci, F., Rokach, L., and Shapira, B. (2011). Introduction to recommender systems handbook. In *Recommender systems handbook*, pages 1–35. Springer.

[Richardson, 1981] Richardson, J. S. (1981). The anatomy and taxonomy of protein structure. *Advances in protein chemistry*, 34:167–339.

[Riley et al., 2005] Riley, R., Lee, C., Sabatti, C., and Eisenberg, D. (2005). Inferring protein domain interactions from databases of interacting proteins. *Genome biology*, 6(10):R89.

[Robillard et al., 2014] Robillard, M. P., Maalej, W., Walker, R. J., and Zimmermann, T. (2014). *Recommendation systems in software engineering*. Springer Science & Business.

[Rost et al., 2003] Rost, B., Liu, J., Nair, R., Wrzeszczynski, K. O., and Ofran, Y. (2003). Automatic prediction of protein function. *Cellular and Molecular Life Sciences*, 60(12):2637–2650.

[Roy et al., 2012] Roy, A., Yang, J., and Zhang, Y. (2012). COFACTOR: an accurate comparative algorithm for structure-based protein function annotation. *Nucleic acids research*, 40(W1):W471–W477.

[Sammut et al., 2008] Sammut, S. J., Finn, R. D., and Bateman, A. (2008). Pfam 10 years on: 10 000 families and still growing. *Briefings in bioinformatics*, 9(3):210–219.

[Samuel, 2000] Samuel, A. L. (2000). Some studies in machine learning using the game of checkers. *IBM Journal of research and development*, 44(1.2):206–226.

[Sarwar et al., 2001] Sarwar, B., Karypis, G., Konstan, J., and Riedl, J. (2001). Item-based collaborative filtering recommendation algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 285–295. ACM.

[Schomburg et al., 2002] Schomburg, I., Chang, A., and Schomburg, D. (2002). BRENDA, enzyme data and metabolic information. *Nucleic acids research*, 30(1):47–49.

[Schultz et al., 1998] Schultz, J., Milpetz, F., Bork, P., and Ponting, C. P. (1998). Smart, a simple modular architecture research tool: identification of signaling domains. *Proceedings of the National Academy of Sciences*, 95(11):5857–5864.

[Segura et al., 2015] Segura, J., Sorzano, C. O. S., Cuenca-Alba, J., Aloy, P., and Carazo, J. M. (2015). Using neighborhood cohesiveness to infer interactions between protein domains. *Bioinformatics*, 31(15):2545–2552.

[Sharan et al., 2007] Sharan, R., Ulitsky, I., and Shamir, R. (2007). Network-based prediction of protein function. *Molecular systems biology*, 3(1):88.

[Sigrist et al., 2002] Sigrist, C. J., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A., and Bucher, P. (2002). Prosite: a documented database using patterns and profiles as motif descriptors. *Briefings in bioinformatics*, 3(3):265–274.

[Sigrist et al., 2005] Sigrist, C. J., De Castro, E., Langendijk-Genevaux, P. S., Le Saux, V., Bairoch, A., and Hulo, N. (2005). Prorule: a new database containing functional and structural information on prosite profiles. *Bioinformatics*, 21(21):4060–4066.

[Singhal, 2001] Singhal, A. (2001). Modern information retrieval: A brief overview. *IEEE Data Eng. Bull.*, 24(4):35–43.

[Sprinzak and Margalit, 2001] Sprinzak, E. and Margalit, H. (2001). Correlated sequence-signatures as markers of protein-protein interaction. *Journal of molecular biology*, 311(4):681–692.

[Stark et al., 2006] Stark, C., Breitkreutz, B.-J., Reguly, T., Boucher, L., Breitkreutz, A., and Tyers, M. (2006). Biogrid: a general repository for interaction datasets. *Nucleic acids research*, 34(suppl_1):D535–D539.

[Stein et al., 2010] Stein, A., Céol, A., and Aloy, P. (2010). 3did: identification and classification of domain-based interactions of known three-dimensional structure. *Nucleic acids research*, 39(suppl_1):D718–D723.

[Stein et al., 2005] Stein, A., Russell, R. B., and Aloy, P. (2005). 3did: interacting protein domains of known three-dimensional structure. *Nucleic acids research*, 33(suppl_1):D413–D417.

[Stoesser et al., 2002] Stoesser, G., Baker, W., van den Broek, A., Camon, E., Garcia-Pastor, M., Kanz, C., Kulikova, T., Leinonen, R., Lin, Q., Lombard, V., et al. (2002). The embl nucleotide sequence database. *Nucleic acids research*, 30(1):21–26.

[Suzek et al., 2007] Suzek, B. E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C. H. (2007). Uniref: comprehensive and non-redundant uniprot reference clusters. *Bioinformatics*, 23(10):1282–1288.

[Suzek et al., 2014] Suzek, B. E., Wang, Y., Huang, H., McGarvey, P. B., Wu, C. H., et al. (2014). Uniref clusters: a comprehensive and scalable alternative for improving sequence similarity searches. *Bioinformatics*, 31(6):926–932.

[Tamames et al., 1997] Tamames, J., Casari, G., Ouzounis, C., and Valencia, A. (1997). Conserved clusters of functionally related genes in two bacterial genomes. *Journal of molecular evolution*, 44(1):66–73.

[Tateno et al., 2000] Tateno, Y., Miyazaki, S., Ota, M., Sugawara, H., and Gojobori, T. (2000). Dna data bank of japan (ddbj) in collaboration with mass sequencing teams. *Nucleic Acids Research*, 28(1):24–26.

[Thomas et al., 2003] Thomas, P. D., Kejariwal, A., Campbell, M. J., Mi, H., Diemer, K., Guo, N., Ladunga, I., Ulitsky-Lazareva, B., Muruganujan, A., Rabkin, S., et al. (2003). Panther: a browsable database of gene products organized by biological function, using curated protein family and subfamily classification. *Nucleic acids research*, 31(1):334–341.

[Thompson et al., 1994] Thompson, J. D., Higgins, D. G., and Gibson, T. J. (1994). Clustal w: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic acids research*, 22(22):4673–4680.

[Tsiptsis and Chorianopoulos, 2011] Tsiptsis, K. K. and Chorianopoulos, A. (2011). *Data mining techniques in CRM: inside customer segmentation*. John Wiley & Sons.

[Tuschl, 2003] Tuschl, T. (2003). Functional genomics: Rna sets the standard. *Nature*, 421(6920):220–221.

[Valencia and Pazos, 2002] Valencia, A. and Pazos, F. (2002). Computational methods for the prediction of protein interactions. *Current opinion in structural biology*, 12(3):368–373.

[Vazquez et al., 2003] Vazquez, A., Flammini, A., Maritan, A., and Vespignani, A. (2003). Global protein function prediction in protein-protein interaction networks. *arXiv preprint cond-mat/0306611*.

[Velankar et al., 2012] Velankar, S., Dana, J. M., Jacobsen, J., van Ginkel, G., Gane, P. J., Luo, J., Oldfield, T. J., O'Donovan, C., Martin, M.-J., and Kleywegt, G. J. (2012). Sifts: structure integration with function, taxonomy and sequences resource. *Nucleic acids research*, 41(D1):D483–D489.

[Wang et al., 2001] Wang, J. T.-L., Ma, Q., Shasha, D., and Wu, C. H. (2001). New techniques for extracting features from protein sequences. *IBM Systems Journal*, 40(2):426–441.

[Wasserman, 2013] Wasserman, L. (2013). *All of statistics: a concise course in statistical inference*. Springer Science & Business Media.

[Watson et al., 2005] Watson, J. D., Laskowski, R. A., and Thornton, J. M. (2005). Predicting protein function from sequence and structural data. *Current opinion in structural biology*, 15(3):275–284.

[Webb et al., 1992] Webb, E. C., of Biochemistry, I. U., and Biology, M. (1992). *Enzyme nomenclature 1992. Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes*. Number Ed. 6. Academic Press.

[Wei et al., 2016] Wei, Z.-S., Han, K., Yang, J.-Y., Shen, H.-B., and Yu, D.-J. (2016). Protein–protein interaction sites prediction by ensembling svm and sample-weighted random forests. *Neurocomputing*, 193:201–212.

[Whisstock and Lesk, 2003] Whisstock, J. C. and Lesk, A. M. (2003). Prediction of protein function from protein sequence and structure. *Quarterly reviews of biophysics*, 36(3):307–340.

[Witten et al., 2016] Witten, I. H., Frank, E., Hall, M. A., and Pal, C. J. (2016). *Data Mining: Practical machine learning tools and techniques*. Morgan Kaufmann.

[Wurman, 1989] Wurman, R. S. (1989). *Information anxiety*. Doubleday.

[Xenarios et al., 2000] Xenarios, I., Rice, D. W., Salwinski, L., Baron, M. K., Marcotte, E. M., and Eisenberg, D. (2000). Dip: the database of interacting proteins. *Nucleic acids research*, 28(1):289–291.

[Xenarios et al., 2002] Xenarios, I., Salwinski, L., Duan, X. J., Higney, P., Kim, S.-M., and Eisenberg, D. (2002). Dip, the database of interacting proteins: a research tool for studying cellular networks of protein interactions. *Nucleic acids research*, 30(1):303–305.

[Xin and Radivojac, 2011] Xin, F. and Radivojac, P. (2011). Computational methods for identification of functional residues in protein structures. *Current Protein and Peptide Science*, 12(6):456–469.

[Xu and Dunbrack Jr, 2012] Xu, Q. and Dunbrack Jr, R. L. (2012). Assignment of protein sequences to existing domain and family classification systems: Pfam and the pdb. *Bioinformatics*, 28(21):2763–2772.

[Yang, 2010] Yang, Z. R. (2010). *Machine learning approaches to bioinformatics*, volume 4. World scientific.

[Yates et al., 2016] Yates, A., Akanni, W., Amode, M. R., Barrell, D., Billis, K., Carvalho-Silva, D., Cummins, C., Clapham, P., Fitzgerald, S., Gil, L., et al. (2016). Ensembl 2016. *Nucleic acids research*, 44(D1):D710–6.

[Yellaboina et al., 2010] Yellaboina, S., Tasneem, A., Zaykin, D. V., Raghavachari, B., and Jothi, R. (2010). Domine: a comprehensive collection of known and predicted domain-domain interactions. *Nucleic acids research*, 39(suppl_1):D730–D735.

[You et al., 2017] You, R., Zhang, Z., Xiong, Y., Sun, F., Mamitsuka, H., and Zhu, S. (2017). Golabeler: Improving sequence-based large-scale protein function prediction by learning to rank. *bioRxiv*, page 145763.

[Zahiri et al., 2013] Zahiri, J., Hannon Bozorgmehr, J., and Masoudi-Nejad, A. (2013). Computational prediction of protein–protein interaction networks: algorithms and resources. *Current genomics*, 14(6):397–414.

[Zaki, 2000] Zaki, M. J. (2000). Scalable algorithms for association mining. *IEEE Transactions on Knowledge and Data Engineering*, 12(3):372–390.

[Zanzoni et al., 2002] Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M., and Cesareni, G. (2002). Mint: a molecular interaction database. *FEBS letters*, 513(1):135–140.

[Zdobnov and Apweiler, 2001] Zdobnov, E. M. and Apweiler, R. (2001). Interproscan–an integration platform for the signature-recognition methods in interpro. *Bioinformatics*, 17(9):847–848.

[Zhang et al., 2014] Zhang, S.-W., Hao, L.-Y., and Zhang, T.-H. (2014). Prediction of protein–protein interaction with pairwise kernel support vector machine. *International journal of molecular sciences*, 15(2):3220–3233.

[Zhu et al., 2007] Zhu, M., Gao, L., Guo, Z., Li, Y., Wang, D., Wang, J., and Wang, C. (2007). Globally predicting protein functions based on co-expressed protein–protein interaction networks and ontology taxonomy similarities. *Gene*, 391(1):113–119.

Résumé

L'un des moyens les plus intéressants et les plus fructueux d'inférer des principes à partir de grands ensembles de données est l'utilisation des techniques d'exploration (ou de fouille) de données. Cette thèse aborde le problème de la découverte d'associations cachées dans des ensembles de données complexes en utilisant la similarité vectorielle et les résultats obtenus ont été appliqués à la prédiction des annotations biologiques grâce à l'ingénierie de règles d'association.

Les protéines sont des macromolécules qui exercent des fonctions biologiques dans les organismes vivants. Une protéine consiste en une séquence d'acides aminés qui adopte une forme tridimensionnelle (3D) particulière, largement responsable de sa fonction moléculaire. Les fonctions des protéines peuvent être décrites par différentes ontologies, termes ou classifications, dans lesquelles les relations entre ces fonctions peuvent être hiérarchiques (Gene Ontology (GO), Enzyme Commission Numbers (EC)). Au niveau moléculaire, les fonctions sont souvent effectuées par des parties de protéines hautement conservées, identifiées à partir d'alignements de séquence ou de structure, qui peuvent être classés en domaines ou familles. Cependant, il existe des millions de protéines composées de plusieurs domaines, dans lesquelles un domaine seul ou une combinaison de domaines sont responsables d'une fonction. Par conséquent, annoter les domaines responsables d'une fonction spécifique est une tâche non triviale. En outre, l'affectation manuelle des fonctions protéiques aux domaines correspondants en utilisant des connaissances spécialisées prend beaucoup de temps. Une méthode de calcul devrait donc être développée pour aborder le problème de l'association des domaines protéiques avec des fonctions protéiques.

Avec la croissance rapide du nombre de structures et de séquences de protéines découvertes, le nombre de séquences de protéines qui ne comportent pas d'annotations fonctionnelles augmente énormément. La prédiction automatique des fonctions protéiques est un des grands défis de la bioinformatique.

Cette thèse présente: 1) le développement d'une nouvelle approche pour trouver des associations directes entre des paires d'éléments liés indirectement à travers diverses caractéristiques communes, 2) l'utilisation de cette approche pour associer directement des fonctions biologiques aux domaines protéiques (ECDomainMiner et GODomainMiner) et pour découvrir des interactions domaine-domaine, et enfin 3) l'extension de cette approche pour annoter de manière à partir des domaines complète les structures et les séquences des protéines.

Au total, 20 728 et 20 318 associations EC-Pfam et GO-Pfam non redondantes ont été découvertes, avec des F-mesures de plus de 0,95 par rapport à un ensemble de référence Gold Standard extrait d'une source d'associations connues (InterPro). Par rapport à environ 1500 associations déterminées manuellement dans InterPro, ECDomainMiner et GODomainMiner produisent une augmentation de 13 fois du nombre d'associations EC-Pfam et GO-Pfam disponibles.

Ces associations domaine-fonction sont ensuite utilisées pour annoter des milliers de structures de protéines et des millions de séquences de protéines pour lesquelles leur composition de domaine est connue mais qui manquent actuellement d'annotations fonctionnelles. En utilisant des associations de domaines ayant acquis des annotations fonctionnelles inférées, et en tenant compte des informations de taxonomie, des milliers de règles d'annotation ont été générées automatiquement. Ensuite, ces règles ont été utilisées pour annoter des séquences de protéines dans la base de données TrEMBL. Nous avons également utilisé ces règles d'annotation pour participer à un défi intitulé L'évaluation critique des algorithmes d'annotation de fonctions protéiques (CAFA) afin de découvrir les termes GO pour 121 914 séquences cibles ayant au

moins un domaine connu présent dans nos règles dérivées de GODomainMiner. L'annotation fonctionnelle automatique des séquences protéiques a été réalisée en collaboration avec l'équipe UniProt au European Bioinformatics Institute (EBI) où j'ai passé troi mois pendant ma thèse.

Au cours de cette thèse, deux articles évalués par des pairs ont été publiés : « ECDomainMiner: la découverte d'associations cachées entre les numéros de commission enzymatique et les domaines de Pfam » et « Associer les termes de l'ontologie des gènes aux domaines protéiques Pfam » (accepté). Trois autres manuscrits sont en préparation. Les bases de données des résultats ECDomainMiner et GODomainMiner sont publiquement disponibles à http://ecdm.loria.fr/, http://godm.loria.fr/, respectivement.

Mots-clés: Graphes tripartites, similarité vectorielle, règles d'associations, bases de données biologiques, domaines protéiques, annotation fonctionnelle des protéines, interactions domaine-domaine.

Abstract

One of the most interesting and powerful ways of inferring principles out of large datasets is usage of data mining. This thesis addresses the problem of discovering hidden associations in complex datasets using vector similarity and the method proposed has been applied to the prediction of biological annotations.

Proteins are macromolecules which carry out biological functions in living organisms. A protein consists of a sequence of amino acids which fold into a particular three-dimensional (3D) shape that is largely responsible for its molecular function. The functions of proteins can be described by different ontologies, terms, or classifications, whereas the relationships between these functions can be hierarchical (Gene Ontology (GO), Enzyme Commission Numbers (EC)) or flat. At the molecular level, functions are often performed by highly conserved parts of proteins, identified from sequence or structure alignments, which may be classified into domains or families (such as SCOP, CATH, PFAM, TIGRFAMs). The known functions of a whole protein can easily be transferred to a domain if proteins comprise single domain. However, there are millions of proteins with multiple domains in which a domain alone or a combination of domains are responsible for a function. Therefore, annotating which domains carry out a specific function is a non-trivial task.

Several direct associations between protein domains and functions have been annotated manually. Nevertheless, the list of such annotations is incomplete. In addition, manual assignment of protein functions to the corresponding domains using expert knowledge is very time-consuming. A computational method should thus be developed to tackle the problem of associating protein domains to protein functions.

With the prompt growth in the number of discovered protein structures and sequences, the number of protein sequences that lack functional annotations from in vitro experiments is increasing enormously. More than 99% of protein sequences in UniProtKB have no experimental functional annotations. Thus, it is indispensable to bridge this widening functional annotation gap by computational prediction of protein functions.

This thesis presents: 1) the development of a novel approach to find direct associations between pairs of elements linked indirectly through various common features, 2) the use of this approach to directly associate biological functions to protein domains (ECDomainMiner and GODomainMiner), and to discover domain-domain interactions, and finally 3) the extension of this approach to comprehensively annotate protein structures and sequences.

ECDomainMiner and GODomainMiner are two applications to discover new associations between EC Numbers and GO terms to protein domains, respectively. They find a total of 20,728 and 20,318 non-redundant EC-Pfam and GO-Pfam associations, respectively, with F-measures of more than 0.95 with respect to a "Gold Standard" test set extracted from InterPro. Compared to around 1500 manually curated associations in InterPro, ECDomainMiner and GODomainMiner infer a 13-fold increase in the number of available EC-Pfam and GO-Pfam associations.

These function-domain associations are then used to annotate thousands of protein structures and millions of protein sequences for which their domain composition is known but that currently lack experimental functional annotations. Using inferred function-domain associations and taking taxonomy information into account, thousands of annotation rules have automatically been generated. Then, these rules have been utilized to annotate protein sequences in the TrEMBL database. We also used these annotation rules for participating in a challenge called "The Critical Assessment of protein Function Annotation algorithms (CAFA)" in order to discover GO terms for 121,914 target sequences having at least

one known domain present in our GODomainMiner-derived rules. Automatic functional annotation protein sequences has been done in collaboration with UniProt team at European Bioinformatics Institute (EBI).

During the course of this thesis, two peer-reviewed articles of "ECDomainMiner: discovering hidden associations between enzyme commission numbers and Pfam domains" and Associating Gene Ontology Terms with Pfam Protein Domains have been published. Three further manuscripts are in preparation. The ECDomainMiner and GODomainMiner result databases are publicly available at http://ecdm.loria.fr/, http://godm.loria.fr/, respectively.

Keywords: Tripartite graphs, vector similarity, association rules, biological databases, protein domains, functional annotation of proteins, domain-domain interactions.