UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS

Identification of biotechnological potential on genomic nonfunctionalized orthologs elements

"Documento Definitivo"

Doutoramento em Biologia
Especialidade de Biologia de Sistemas

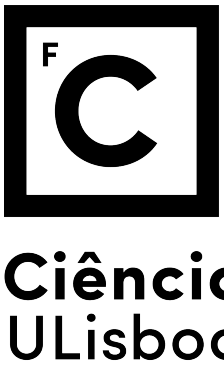
Pedro Miguel Agostinho Escudeiro

Tese orientada por:
Professor Doutor Ricardo Pedro Moreira Dias
Professor Doutor Christopher S. Henry

Documento especialmente elaborado para obtenção do grau de doutor

2022

UNIVERSIDADE DE LISBOA
FACULDADE DE CIÊNCIAS



Identification of biotechnological potential on genomic nonfunctionalized orthologs elements

Doutoramento em Biologia

Especialidade de Biologia de Sistemas

Pedro Miguel Agostinho Escudeiro

Tese orientada por:

Professor Doutor Ricardo Pedro Moreira Dias

Professor Doutor Christopher S. Henry

Júri:

Presidente:

- Doutor Rui Manuel dos Santos Malhó, Professor Catedrático e Presidente do Departamento de Biologia Vegetal da Faculdade de Ciências da Universidade de Lisboa.

Vogais:

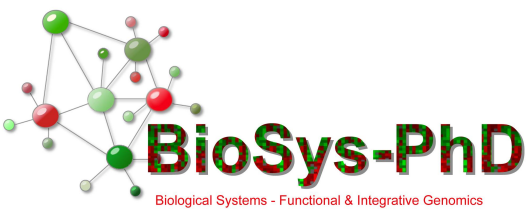
- Doutor Christopher S. Henry, *Computational Science Leader*, *Argonne National Laboratory* da *University of Chicago*, Estados Unidos, Orientador;
- Doutora Natacha Monge Gomes do Couto, Investigadora, *University of Bath*, Reino Unido;
- Doutor Daniel Vieira Noro e Silva Sobral, Investigador Auxiliar, Instituto Nacional de Saúde Doutor Ricardo Jorge - INSA;
- Doutor Francisco José Moreira Couto, Professor Associado com Agregação, Faculdade de Ciências da Universidade de Lisboa;
- Doutora Helena Margarida Moreira de Oliveira Vieira, Professora Associada Convidada, Faculdade de Ciências da Universidade de Lisboa.

Financiado pela Fundação para a Ciência e Tecnologia (PD/BD/131416/2017)

Documento especialmente elaborado para obtenção do grau de doutor

2022

Pedro Miguel Agostinho Escudeiro foi bolseiro de doutoramento no âmbito do programa doutoral BioSys – Sistemas Biológicos, Genómica Funcional & Integrativa (FCT /PD/00065/2012) da Faculdade de Ciências da Universidade de Lisboa, financiada pela Fundação para a Ciência e Tecnologia do Ministério da Educação e Ciência, Bolsa PD/BD/131416/2017.



Acknowledgements

I would like to thank:

- First and foremost, **Ricardo Pedro Moreira Dias, Ph.D.**, for accepting me as his doctoral student, thus granting me this once in a lifetime opportunity to pursue a long-standing goal of mine. I shall remain forever grateful for the trust he placed in me; for the years he invested in me; for encouraging me to chase after my curiosity, and for granting me the freedom to do so; yet to be mindful, and not to lose sight of "the forest". I offer him a heartfelt Thank You.

- **Christopher S. Henry, Ph.D.**, for wholeheartedly accommodating me in his group, for his humanity, compassion and understanding; for granting me the possibility to live and work abroad; for trusting my judgement and for giving me the autonomy to explore new methodologies and viewpoints. I am also grateful to him for bestowing me with access to a high-end server, without which a substantial amount of this work would not have been possible to execute.

- The **Fundação para a Ciência e a Tecnologia**, the **BioSys PhD Program**, and the **Argonne National Laboratory**, for the funding that enabled my to carry out my studies, travels, subsistence, and accommodation throughout these years.

- **Francisco M. Couto, Ph.D.**, for helping me conceptualize and structure part of the present work; for his availability in meeting with me; for the fruitful discussions; and for his kindness and patience.

- Last, but certainly not the least, **José Pedro Lopes Faria, Ph.D.**, for his friendship, support, guidance, and immeasurable forbearance; for being a kindhearted fellow countryman abroad; for helping me through some challenging times, when everything seemed bleak and Home stood so far away; and for keeping me company when I needed the most.

Aos meus Pais

Abstract

A staggering amount of gene products imputed from prokaryotic sequence data fail to be annotated by mainstream computational methods. This has led to an ever-increasing growth in gene products of unknown function within public databases. Evidence suggests that these gene products might provide biotechnological solutions to pressing societal concerns.

In this thesis we sought to understand if the putative functions of these uncharacterized gene products attained biotechnological interest. To this end, we collected $>$134 million protein sequences of unknown function from public databases. By doing so, we created the first worldwide repository of prokaryotic gene products of unknown molecular function. Upon clustering these sequences we generated a representative dataset containing $\sim$12 million proteins. We managed to annotate 99.97% of this dataset with at least one term. From the foregoing sum, 2.78% (351,917) were annotated with at least one Enzyme Commission (EC) number. We postulate that these are putative enzymes. We found that the most abundant enzymatic classes were those of Transferases (182,797) and Hydrolases (100,475). We also found that 9,622 putative enzymes might portray catalytic promiscuity, or multiple catalytic functions altogether. Afterwards we developed a new family of information-theoretic metrics that allow to quantify the annotation content and quality of a given sequence. These metrics enabled us to systemize our dataset according to distinct spectra of annotation. They also allow to expedite the selection of the best annotated sequences for ensuing experimental validation. We also provide a proof-of-concept for the usefulness of the work developed in this thesis by characterizing a putative enzyme subclass of utmost societal significance.

We conclude that there is both a tremendous quantity and diversity of uncharacterized gene products whose predicted functions are directly implicated with well-established biotechnological applications. This reservoir can be tapped into at the present time, conceivably allowing to solve several societal demands.

Keywords: Functional Genomics; Molecular Function Prediction; Microbial Dark Matter; Information Content; Biotechnology

Resumo

A tecnologia de sequenciação de última geração permitiu adquirir uma quantidade massiva de dados de cariz genético, levando a avanços sem precedentes no campo da genómica. O crescimento contínuo da aquisição deste tipo de dados genómicos levou a que bases de dados públicas fossem populadas com uma quantidade impressionante de sequências nucleotídicas. Apesar desta quantidade exorbitante de dados adquiridos pelas tecnologias de sequenciação atuais, a função molecular de mais de 35% dos genes provenientes de determinado genoma microbiano ainda permanece desconhecida.

Microrganismos conhecidos, e amplamente estudados, possuem uma fração considerável de genes de função molecular desconhecida codificados nos seus genomas. No entanto, foi descoberto que estes genes codificavam para propriedades fisiológicas essenciais à vida. Em alguns casos, a fração de genes cuja função molecular é desconhecida pode atingir 50% do conteúdo total genómico, como por exemplo em genomas recém-sequenciados ou pertencentes a microrganismos de filos candidatos sem exemplares em cultura. Em cenários mais extremos, conforme indicado por estudos metagenómicos, estas percentagens podem oscilar entre 85% a 99% do conteúdo total de genes identificados.

A caracterização experimental de cada um destes elementos genómicos de função desconhecida não é exequível. Os recursos necessários para se realizar tamanho empreendimento são inimagináveis, e tampouco seria possível acompanhar o ritmo a que estes elementos estão a ser descobertos, e se acumulam nas bases de dados públicas. Adicionalmente, a curadoria manual e individual destes elementos não pode ser levada em consideração como uma opção viável. Estas circunstâncias não deixam qualquer escolha senão empregar métodos de anotação automatizados, normalmente por meio de técnicas computacionais, como a única alternativa para prever a função molecular destes produtos. Além do mais, grande parte dos produtos da tradução imputados a partir destas sequências não são passíveis de ser anotados por parte de métodos de imputação convencionais, tais como os que se baseiam em homologia de sequências, muitas vezes estes primeiros carecendo mesmo de homólogos remotamente reconhecíveis em bases de dados públicas. Deste modo, os referidos elementos permanecem em carência de caracterização funcional. Estas circunstâncias colocam em questão até que ponto estes elementos enigmáticos

codificam realmente para produtos da tradução como os que conhecemos; bem como até que extensão poderão estes ser anotados e as suas funções desvendadas.

Porventura, existe cada vez mais evidência que sugere que estes elementos sem função molecular atribuída podem de facto ser uma mais valia na identificação de soluções biotecnológicas num mundo com crescentes exigências sociais. Estes elementos genéticos de função desconhecida podem-se provar excelentes ativos na identificação de novas proteínas produtoras de metabolitos, enzimas ou até mesmo novos mecanismos fisiológicos. Tal acontecimento poderia por sua vez, a devido tempo, proporcionar um aumento de soluções biotecnológicas inovadoras. É de suma importância desvendar este conhecimento inexplorado, a fim de divulgar novas soluções e respostas às mudanças ambientais e climáticas; de modo a superar desafios sociais como os presentes em áreas tais como a saúde, agronomia e indústria de alimentos; bem como desenvolver ferramentas de ponta que levem ao avanço de áreas como a engenharia metabólica e biologia sintética.

Deste modo, na presente tese doutoral, proposemo-nos prever a função molecular destes produtos da tradução sem caracterização funcional de origem procariótica, com o intuito de compreender se as funções putativas imputadas a partir destas sequências eram detentoras de interesse biotecnológico ou industrial.

Para alcançar estes objetivos, primeiro começámos por pesquisar bases de dados públicas em busca destes produtos da tradução. No decurso desta tarefa além de quantificarmos estes elementos, fomos também colecionando os dados incluídos nestas bases de dados que iam ao encontro dos nossos critérios de qualidade e acuidade. A conclusão desta tarefa levou à criação do primeiro repositório mundial de produtos da tradução imputados a partir de elementos genéticos de origem procariótica cuja função molecular é desconhecida, incorporando mais de 134 milhões de sequências de aminoácidos. Para nosso conhecimento, este repositório constitui a primeira coleção centralizada a nível mundial de "matéria negra funcional" de origem procariótica. Observámos também que a esmagadora maioria destas proteínas provêm de filos conhecidos, independentemente do domínio taxonómico. Em seguida reduzimos a dimensão do nosso repositório ao longo de 8 limiares de identidade aminoácidica global. A primeira etapa de aglomeração proporcionou uma redução de 79.51% no tamanho do nosso repositório, permitindo a criação de um conjunto de dados não redundante. A última etapa de agrupamento permitiu-nos alcançar uma redução de tamanho adicional de 54.2% do conjunto não redundante; implicando uma redução cumulativa de $\sim$90.62% em relação ao número inicial de $>$134

milhões de sequências. Estes resultados suscitam duas conclusões. Em primeiro lugar, de que é possível representar o nosso repositório por um conjunto não redundante de proteínas putativas representativas. Em segundo lugar, a de que pode haver um alto nível de redundância entre (ou dentro de) fontes de dados independentes, para sequências de proteínas de função desconhecida. Também notámos que a maioria das sequências representativas de cada conjunto não possuíam membros adicionais, independentemente do limite de identidade aminoacídica imposto. Além do mais, a maioria dos conjuntos de sequências compreende $<$100 membros, também independentemente do limite. Estes resultados levam-nos a concluir que estas as sequências não caracterizadas são muito distintas entre si.

Posteriormente empregámos várias ferramentas de imputação de função para anotar estas sequências e realizamos várias análises de dados de cariz exploratório de modo a vislumbrar as funções moleculares putativas destes elementos. Conseguimos anotar 99.97% (12,651,624) do conjunto de dados não redundante a 60% de identidade aminoacídica com pelo menos um termo semântico, oriundo de pelo menos um sistema de classificação. Este foi um resultado significativo, visto que em primeiro lugar iniciámos a presente tarefa com sequencias não anotadas. Apenas uma pequena fração destas proteínas (0.03%) não foi passível de anotação por qualquer um dos sistemas de classificação. Estas sequências poderão ser analisadas posteriormente pela comunidade científica, aquando da divulgação pública destes dados num futuro próximo.

Também desenvolvemos uma família de métricas baseadas na teoria da informação que nos permitissem qualificar numericamente as anotações atribuídas a uma dada sequência. Estas métricas classificam a representação semântica de uma determinada sequência de acordo com seu valor e qualidade informativa. Para testar e validar as métricas referidas, usámos um conjunto de dados padrão, bem como $\sim$12 milhões de proteínas que não cederam à anotação pelas suas bases de dados de origem. Ao calcularmos estas métricas para todas as proteínas anotadas, mostramos que é possível representá-las como uma coleção de distribuições. Porventura, ao fazermos uso destas métricas podemos explorar tanto o valor informativo "bruto" embutido numa dada representação semântica, como também avaliar o equilíbrio entre a especificidade desse valor e o seu grau de confiança. Este desenvolvimento permitiu-nos sistematizar estes produtos da tradução de acordo com espectros distintos de anotação. Adicionalmente, os valores incluídos nestas distribuições fornecem um panorama sobre o conteúdo informativo dessas anotações. Além do mais, esta família de métricas permite compreender se as anotações provenientes destas proteínas putativas representam múltiplas funções moleculares.

A combinação das métricas calculadas para cada sistema de classificação utilizado possibilitou a criação de um índice de anotação proteica. Este índice permite sistematizar as referidas anotações tendo em conta toda a extensão do seu valor informativo. Consequentemente, este índice permite a criação de uma distribuição global que caracteriza diferentes graus de anotação de proteínas putativas num determinado conjunto de dados. Também exemplificamos que, ao calcular o valor deste índice para todas as proteínas num conjunto de dados, podemos desvendar sub-distribuições distintas referentes a proteínas com diferentes níveis de anotação. Este desenvolvimento permitiu-nos distinguir quais as proteínas que foram anotadas de forma mais completa em cada um dos conjuntos de dados que usamos.

Depois de aplicarmos um algoritmo de limite, fomos capazes de selecionar um valor deste índice que permitiu a subsecção de cada conjunto de dados entre dois grupos principais de proteínas anotadas. Postulamos que esses grupos retratam proteínas que foram efetivamente anotadas, e aquelas que não o foram. Ao criar esta dicotomia, esperamos agilizar o direcionamento das sequências de proteínas putativas que foram mais flexíveis à anotação automatizada. Além do mais, pretendemos facilitar a seleção das sequências de proteínas mais promissoras para análises posteriores e subsequente validação experimental. Por último, apresentamos uma prova de conceito para a utilidade do trabalho desenvolvido nesta tese, caracterizando uma subclasse enzimática detentora de acrescida significância social.

Concluímos que existe uma enorme quantidade e diversidade de produtos da tradução cujas funções moleculares imputadas estão diretamente implicadas em possíveis aplicações biotecnológicas e industriais bem estabelecidas na literatura científica. O trabalho que desenvolvemos também permite agilizar a seleção das sequências cuja anotação foi mais completa, facilitando deste modo a seleção das mais promissoras para estudos posteriores. Este reservatório colossal de potencial biotecnológico pode ser explorado presentemente, possivelmente permitindo resolver vários desafios oriundos da sociedade contemporânea.

Palavras-Chave: Genómica Funcional; Imputação de Função Molecular; Matéria Negra Microbiana; Conteúdo de Informação; Biotecnologia

Table of Contents

List of Figures

List of Tables

List of Abbreviations

16s rRNA	16S ribosomal RNA
a.a.	Amino-acid
AAI	Amino-acid Identity
ABC	ATP-binding Cassette
ADH	Alcohol Dehydrogenase
AIP	Autoinducing Peptide
API	Application Programming Interface
ATP	Adenosine triphosphate
ATPase	Adenosine triphosphatase
BGC	Biosynthetic Gene Cluster
BH	Benjamini-Hochberg
Cas	CRISPR associated protein
CD	Conserved Domain
CDD	Conserved Domain Database
CDF	Cumulative Distribution Function
CD-HIT	Cluster Database at High Identity with Tolerance
CNN	Convolutional Neural Network
CRISPR	Clustered Regularly Interspaced Short Palindromic Repeats
D2V	Document to Vector
DNA	Deoxyribonucleic acid
EC	Enzyme Commission
EG	Ethylene glycol
EMBL	European Molecular Biology Laboratory
FACS	Fluorescence-activated Cell Sorting
FASTA	Fast-All
FD	Freedman-Diaconis
FsC	*Fusarium solani pisi* Cutinase
FTP	File Transfer Protocol
GC	Guanine-Cytosine

GenBank	NIH genetic sequence database
GenPept	GenBank gene products database
GFF3	General Feature Format version 3
GO	Gene Ontology
GRAVY	Grand average of hydropathicity
HiC	*Humicola insolens* Cutinase
HMM	Hidden Markov Model
Ichip	Isolation chip
IMG	Integrated Microbial Genomes
IQR	Interquartile range
JGI	Joint Genome Institute
kDa	kiloDalton
KDE	Kernel Density Estimation
K-S	Kolmogorov-Smirnov
MALDI-TOF	Matrix-Assisted Laser Desorption/Ionization Time-of-Flight mass-spectrometer
MF	Molecular Function
MHET	Mono-2-hydroxyethyl terephthalate
MHETase	Mono-2-hydroxyethyl terephthalate hydrolase
MSA	Multiple Sequence Alignment
MW	Molecular Weight
NAD+	Nicotinamide adenine dinucleotide
NADH	Reduced form of nicotinamide adenine dinucleotide
NADP+	Nicotinamide adenine dinucleotide phosphate
NADPH	Reduced form of nicotinamide adenine dinucleotide phosphate
NCBI	National Center for Biotechnology Information
NGS	Next-generation Sequencing
NoSQL	non-relational / non-SQL
NRP	Nonribosomal Peptide
NRPS	Nonribosomal Peptide Synthase
ORA	Over Representation Analysis
ORF	Open Reading Frame
ORFan	“orphan” ORF

PAS	Per-Arnt-Sim domain
PATRIC	Pathosystems Resource Integration Center
PCA	Principal Component Analysis
PDB	Protein Data Bank
PET	Polyethylene terephthalate
PETase	Polyethylene terephthalate hydrolase
pH	Potential of hydrogen
pI	Isoelectric Point
PK	Polyketide
PKS	Polyketide Synthase
PPI	Protein–protein Interaction
PSSM	Position-specific Scoring Matrix
PTM	Post-translational Modification
RAM	Random Access Memory
RefSeq	Reference sequence database
ρ	Spearman's rank correlation coefficient
RiPP	Ribosomally synthesized and post-translationally modified peptide
RNA	Ribonucleic acid
RPS-BLAST	Reversed Position Specific Basic Local Alignment Search Tool
RuBisCo	Ribulose-1,5-bisphosphate carboxylase-oxygenase
SAG	Single Amplified Genome
SeCys	Selenocysteine
SM	Secondary Metabolite
SQL	Structured Query Language
SRA	Sequence Read Archive
SS	Disulfide bond
S-W	Shapiro-Wilk
T-Coffee	Tree-based Consistency Objective Function for Alignment Evaluation
TCS	Transitive Consistency Score
TPA	Terephthalic acid
TrEMBL	Translated EMBL Nucleotide Sequence Data Library
tRNA	Transfer RNA

UniProt	Universal Protein resource
USD	United States Dollar
W/A	Without Annotation

Glossary

In order to expedite the comprehension of this thesis, we provide the following definitions:

Anc	Ancestors. The parent nodes of a given term t in a hierarchical topology.
annot	A function that represents the input-output logic for a prediction tool. Given a protein x from a corpus X as input, it outputs a set that contains terms t paired with prediction scores s.
b	A histogram bin.
b'	A smoothed histogram bin.
B	A histogram. A collection of bins b.
B'	A smoothed histogram. A collection of smoothed bins b'.
BPS	Biotechnological Potential Subset. Group of proteins from each dataset with EC, GO, and CD annotations; whose *PAI* value is greater than the threshold generated by the Minimum Method Algorithm; annotated with a single 4th-digit EC number; and whose 4th-digit EC number is represented on Table 1.1.
C	Classification System. Contains a set of terms T, and a set of relations between those terms R. In this thesis, a Classification System can be either the Gene Ontology (GO), Enzyme Commission (EC) numbers, or Conserved Domains (CD).
CA_x	Common Ancestors. Subset containing the ancestor terms shared by the terms in a given set.
Corpus	Resource consisting in a large collection of texts. In this thesis a corpus refers exclusively to a collection of protein sequences (i.e., a dataset).

DAG	Directed Acyclic Graph. Hierarchical topology consisting on nodes and edges, with each edge directed from one node to another, such that it never forms a closed loop.
Desc	Descendants. The child nodes of a given term t in a hierarchical topology.
FDM	Functional Dark Matter. Genomic sequences of microbial origin with unknown molecular function.
$\mathbf{f_X(t)}$	The frequency of a term t in a corpus X. This frequency is calculated by taking into account the number of times a term t occurs, along with the number of times all of its descendants occur.
$\mathbf{h_{FD}}$	The bin-width calculated by the Freedman-Diaconis rule.
Hypotheticals	Dataset comprising 12,654,843 prokaryotic protein sequences whose public datasources of origin were unable to annotate. These sequences are the cluster representatives from clusters whose members are at 60% global sequence identity resolution.
IC	Information Content. The negative log likelihood of the probability of occurrence of a term within a given corpus.
$\mathbf{IC_u}$	Uniformized Information Content. The value of *IC* divided by the log likelihood of the frequency of the *root* term.
$\mathbf{k^*}$	The optimal number of histogram bins. The maximum between the number of histogram bins calculated by Sturges formula, and the number of histogram bins calculated using the bin-width h_{FD}.
$\mathbf{k_{FD}}$	The number of histogram bins calculated using the bin-width h_{FD}.
$\mathbf{k_{Sturges}}$	The number of histogram bins calculated by the Sturges formula.
***k*-ary tree**	A rooted tree where each node has at most k children. Where k is the number of descendents of the parent node with the most offspring. All of its leaf nodes are at the same depth of the graph.
L	A logical proposition.

$\mathbf{M_x}$	The subset of a semantic representation T_x, containing the terms for which the value of depth in the sub-graph is maximized.
MDM	Microbial Dark Matter. The uncharted fraction of microbial diversity, and its undisclosed coding potential.
MICA	Most Informative Common Ancestor. The subset of common ancestors CA for which the IC_u is maximal.
$\mathbf{N_X(C)}$	The total number of distinct terms t annotated to proteins x in a corpus X, for a Classification System C.
p(t)	Probability of occurrence of a term t within a given corpus.
PAI	Protein Annotation Index. A weighted arithmetic mean, whose weights w_C are specific for each Classification System C. This weighted arithmetic mean is calculated for a given protein x, using the quantile-transformed min-max scaled PIC_S values gathered for each Classification System C in Z.
PIC	Protein Information Content. An equation that calculates the overall Information Content of a protein x in relation to its semantic representation T_x. It might have different outcomes depending on the cardinality of M_x, whether a $MICA_x$ exists, and whether the semantic representation T_x issues from a Classification System C whose R contains subsumption relations.
$\mathbf{PIC_S}$	Prediction-score-weighted version of the Protein Information Content equation.
$\boldsymbol{\Phi^{-1}}$	Probit function. The quantile function of the standard normal distribution.
Q	Equation that subsumes a normalization procedure that includes both a quantile transformation and min-max feature scaling.
$\mathbf{Q \circ PIC_S}$	A quantile-transformed min-max scaled version of the PIC_S equation.

R	The set of all relations between the terms t enclosed in a given Classification System C.
R^+	Transitive closure. The smallest relation on T that contains R and is transitive.
ref	A predicate that evaluates to true when a protein x in a corpus X refers a term t in T.
root	The root term. The single common ancestor to all terms in a Classification System.
s	A prediction score associated with a term t that was annotated to a protein x.
S	A function that returns the prediction score s calculated by *annot* given a term t as input.
sub-DAG	Subset of a Directed Acyclic Graph (DAG).
Subsumption relation	Relation between two entities, where one entity is a subclass of the other.
SwissProt	Manually annotated protein sequence database. In this thesis the "SwissProt" dataset refers to 235,544 prokaryotic protein sequences.
t	Semantic term. A label used to describe a given entity. In this thesis, a semantic term can be either a Gene Ontology (GO) term, a Enzyme Commission (EC) number, or an identifier for a Conserved Domain (CD).
T	The set of all terms t enclosed in a given Classification System C.
T_x	Semantic representation of a protein x. A set of terms that describe its molecular function.
$T_{x,Z}$	Superset of the semantic representations T_x of a protein x for each Classification System C in Z.

TDM	Taxonomic Dark Matter. Genomic sequences of microbial origin with unknown taxonomic provenance.
τ	Threshold value generated by the Minimum Method thresholding Algorithm.
$\mathbf{x}$	Symbolic representation of a protein sequence.
$\mathbf{X}$	A corpus. A symbolic representation of a collection of protein sequences (i.e., a dataset).
$\mathbf{w_C}$	Weight measure of a Classification System C. The total number of distinct terms in a Classification System C, annotated to proteins x in a corpus X; divided by the total number of distinct terms in all Classification Systems C in Z, annotated to proteins x in a corpus X.
$\mathbf{Z}$	A triple that contains the three Classification Systems C used in this thesis: the Gene Ontology (GO), Enzyme Commission (EC) numbers, and Conserved Domains (CD).

Thesis Outline

Objectives

The main goal of this doctoral thesis was to understand whether prokaryotic amino-acid sequences of unknown molecular function coded for enzymes whose function was of biotechnological interest. In order to meet this goal, we established three core objectives:

- ◇ **Quantification.** In this objective we sought to determine the amount of amino-acid sequences of unknown molecular function gathered from public databases that issued from the Archaea and Bacteria domains. Additionally, we wanted to store these sequences and establish the world's first repository of functionally-uncharacterized prokaryotic sequences. Lastly, we aspired to reduce the size of our repository and comprehend how these sequences related to one another in the process.

- ◇ **Classification.** The second objective of this thesis was to classify the amino-acid sequences from the repository we had populated in the first objective. We wanted to use both a gold-standard in addition to state-of-the-art function prediction tools, so that we would be able to collect diverse, non-overlapping, and complementary annotations from distinct sources whose architecture was heterogeneous.

- ◇ **Validation.** The last objective of this thesis was manifold. First, we sought to validate the classification that had taken place in the preceding objective. To this end, we developed a new family of information-theoretic measures that would enable us to numerically qualify each set of annotations for a given sequence. Second, we sought to focus our endeavors on a putative enzyme subclass of pressing societal significance. In doing so we strived to provide a proof-of-concept for the possible real-world usefulness of the work developed hitherto.

Structure

The present doctoral thesis is organized as follows:

- ◇ **Chapter 1**, provides a contextualization for the work developed in this thesis. We reference foundational publications, and outline the challenges, opportunities, and biotechnological potential encased within uncharacterized prokaryotic proteins. We discuss the progress and limitations of current approaches. We also outline the potential opportunities and directions for future research.

- ◇ **Chapter 2**, tackles the creation of a new repository and a non-redundant dataset issuing from the latter; as well as the molecular function imputation of the sequences enclosed therein. We address the retrieval, data-processing, size reduction, and classification procedures comprised by the first and second objectives of this thesis. In addition, we present an overview of the potential functions of these elusive sequences.

- ◇ **Chapter 3**, addresses the first mark of the third objective. We go through the rationale that led us to create a novel family of information-theoretic measures. We provide evidence suggesting that these measures allow to compute the information content of an annotated protein. We apply these measures on the non-redundant dataset created in Chapter 2, and compare the results against a gold-standard dataset. We also make the case that our results allow to create a dichotomy between the sequences that were effectively annotated, and those that were not.

- ◇ **Chapter 4**, addresses the second mark of the third objective. We identify and showcase which sequences from our non-redundant dataset belong to enzyme classes with known biotechnological potential. We perform additional analyses to characterize a promising enzymatic subclass of crucial societal significance. We conclude that the representative sequences of this class are prime candidates for further characterization and future biotechnological applications.

- ◇ In **Chapter 5**, we discuss the work that was developed throughout this doctoral study. We also outline the main contributions of this thesis. In addition, we provide retrospective critiques, and insights into future endeavors.

- ◇ **Chapter 6**, provides a few closing remarks.

Chapter 1

General Introduction

This chapter partially transcribes the contents of the following manuscript:

> **Escudeiro P, Henry CS, Tenreiro R & Dias R** (2022). *Functional characterization of prokaryotic dark matter: the road so far and what lies ahead*. submitted to Current Research in Microbial Sciences.

Abstract

Earth is predicted to be home to over one trillion microbial species. Yet, fewer than one hundred thousand prokaryotic species have been described. Such an uncharted fraction of the microbial diversity, and its undisclosed coding potential, is known as the "microbial dark matter" (MDM).

Next-generation sequencing has enabled a massive collection of genome sequence data, leading to unprecedented advances in the field of genomics. Still, over 35% of the orthologous elements of prokaryotic genomes cannot be functionally assigned by standard classification methods, highlighting the need for novel approaches for sequencing data analysis.

Harnessing novel information from unknown prokaryotic species is often limited by the use of annotation methods that rely on sequence similarity searches based on cultured species. This hinders the discovery of unique genetic elements that are missing in cultivated species. An additional challenge to MDM exploration is that the genomes of uncultured prokaryotes usually diverge from those of well-characterized species.

Here, we outline the challenges, opportunities, and the potential hidden within the functional dark matter (FDM) of prokaryotes. We also discuss the hurdles and pitfalls surrounding the molecular and computational approaches currently used to probe these uncharted waters, and discuss future opportunities for research and applications.

1.1 What is the Microbial Dark Matter?

Ever since the dawn of life ∼3.5 billion years ago [1, 2], the Earth's environmental, geochemical, and biological systems of all levels of complexity have relied on microorganisms [3]. Microorganisms are the most abundant, ubiquitous, functionally and metabolically diverse forms of life [4]; they are responsible for a substantial part of the planet's biomass [5] and an overwhelming fraction of its biodiversity [4, 5].

Estimating the number of prokaryotic species has been a demanding task [6]. Previous studies have estimated that approximately four hundred thousand (4×10^5) bacterial and archaeal species exist [7], whereas more recent estimates point toward over one trillion (10^{12}) [4]. The actual number, however, is still a subject of debate [8, 9]. According to modern classification, a very small fraction of all prokaryotic species (up to 10^5) has been described in the literature or in public sequence repositories [6], thus exposing our ignorance regarding their diversity. Roughly 85–99% of prokaryotic taxa are unamenable to axenic culture [10], precluding the *in vitro* characterization of such species, including their metabolic pathways, ecological profiles and, perhaps most importantly, their coding potential.

Bringing this knowledge to light is of paramount importance, as this may guide the development of cutting-edge tools with applications in metabolic engineering and synthetic biology, and may help devise novel biotic solutions to overcome environmental or societal challenges in health, agronomy, and food industries. Insights gained from microbiome research and innovation (R&I) [11] could ultimately contribute to the implementation of the United Nations 2030 Agenda for Sustainable Development [12].

It has been known for decades that most microorganisms cannot be cultured [13–15]. In 1985, Norman Pace's group showed that the diversity of yet-uncultured prokaryotes could be probed by molecular biology methods [16] namely by using the 16S ribosomal RNA (16S rRNA) as a marker [17]. Since then, microbiology has benefited tremendously from further advances in the field, which became collectively known as "molecular microbial ecology" [18].

More recently, next-generation sequencing brought a whole breadth of novel insights into microbial genomics. Moreover, modern methodologies such as metagenomics, metatranscriptomics, and single-cell genomics, have enabled the study of microorganisms in their natural community and biochemical conditions, either by bulk sequencing of the entire community or by capturing and amplifying the genome of individual cells from environmental samples, with

no need for laboratory cultivation [5].

Recent approaches (e.g., "shotgun" metagenomics) further streamlined the analysis of microbial diversity, bypassing cloning and amplification steps and shortening sample preparation times [5]. Moreover, the emergence of fourth-generation sequencing technologies [19] currently allows the gathering of massive amounts of single-molecule data with unprecedented detail and at a relatively low cost [19], overcoming amplification, sequencing, and compositional biases inherent to other molecular approaches [20]. More recently, the combination of high-throughput culturing, coupled to MALDI-TOF mass spectrometry and 16S rRNA sequencing (i.e., Culturomics), have allowed the isolation of hundreds of new prokaryotic species [21].

As a result, a modest part of the unknown microbial "black box" has been unveiled over the past decades. Together with the continuous improvement in bioinformatics, it has been possible to prompt the discovery of novel functional and metabolic features that have (i) bolstered natural product discovery [22–29], (ii) challenged preconceived boundaries among the three domains of life [30, 31], and (iii) reshaped our understanding of microbial life forms [32–34].

The unexploited fraction of microbial diversity, along with its functional and metabolic potential, is commonly referred to as the microbial dark matter (MDM). The term was coined by Marcy and colleagues in 2007 [35], alluding to the large amount of uncharacterized microbial taxa and respective genomes inferred by culture-independent approaches. Although most authors use the term interchangeably to refer to the taxonomic tapestry and the coding potential of MDM, we believe that either concept should be employed separately.

As Bernard et al. pointed out [36], there are four basic categories of sequence novelty based on annotation (Figure 1.1): (i) sequences with known taxonomic provenance and molecular function (e.g., a hydrolase gene from *Escherichia coli*); (ii) sequences with known taxonomic provenance but unknown molecular function (e.g., an *E. coli* gene without ascribed function); (iii) sequences with unknown taxonomic provenance but known molecular function (e.g., a hydrolase gene from an uncharacterized bacterium); and (iv) sequences with unknown taxonomic provenance and unknown molecular function (e.g., a gene with no ascribed function from an uncharacterized bacterium, i.e., the true dark matter).

Still abiding by the rationale set forth by Bernard and colleagues [36], and for clarification purposes, we propose the classification of microbial genomic sequences with unknown taxonomic provenance as "taxonomic dark matter" (TDM), regardless of their functional anno-

tation; and the classification of sequences with unknown molecular function as "functional dark matter" (FDM), regardless of their taxonomic provenance or phylogenetic context. Throughout this review we use the term "microbial" to refer to prokaryotes only, comprising the Archaea and the Bacteria domains. Even though viral dark matter has been the focus of increasing advances and research (see [37] for a review), that topic is beyond the scope of this review.

Most researchers employ phylogenetic-driven techniques and canonical molecular markers, such as the 16S rRNA [7], to study MDM by targeted community profiling approaches. Still, the coding potential of these bacterial communities often remains inaccessible.

Harnessing sequence data from such microorganisms is often limited by gene annotation methods reliant on sequence similarities with proteins characterized from microbial cultures [38]. This approach makes it difficult to study the FDM and identify unique functions particular to uncultured organisms [38]. Adding to these shortcomings, the genome of these elusive microorganisms is typically divergent from well-known species [39–41].

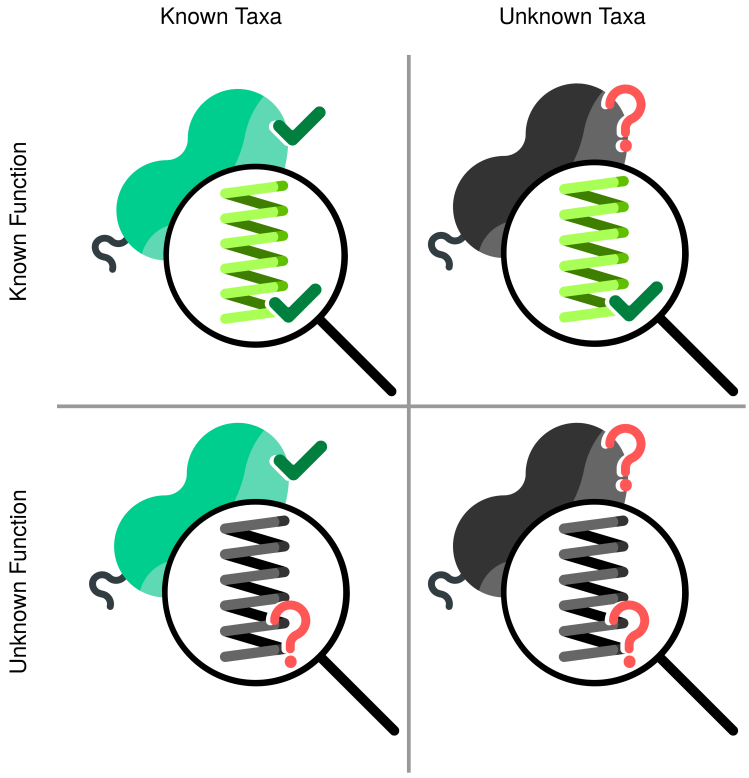


Figure 1.1: Four possible combinations when addressing a protein regarding its taxonomic provenance and molecular function. Both rows and columns depict a binary range: known and unknown. The rows refer to the molecular function of the protein and the columns refer to its taxonomic provenance. Each quadrant represents one combination that results from the intersection of the rows and columns. The bean-shape represents a prokaryotic cell and the coil-shape represents a protein. A green foreground with a green checkmark represents "known" and a grey foreground with a red question mark represents "unknown". Adapted from [36]. The icons used in this figure were retrieved from `flaticon.com`.

Notwithstanding, several remarkable studies have explored the phylogenetic novelty hidden in the TDM, offering insights into the putative functions it encodes [25, 42–47]. Some reports even predict the existence of community-wide metabolic or biogeochemical profiles [48–52].

In fact, in order to find unknown gene functions one does not need to venture into the uncultivated myriad of microorganisms, nor that of candidate phyla for that matter. As of 2021, and exclusively referring to the proteins present in NCBI's RefSeq database, less than 3% of Archaeal (Figure 1.2.A) and 1% of Bacterial (Figure 1.2.B) FDM proteins issue from candidate phyla.

Indeed, in 2016, a minimal synthetic bacterial genome based on that from *Mycoplasma mycoides*, a well-studied mammalian parasite, was generated [53]. This synthetic genome contained only 473 genes, of which 149 had unknown biological functions. Yet, each of those genes was considered essential, as deletion of any of them was lethal [53].

Therefore, even in a controlled laboratory environment, the function of nearly one-third of essential genes is unknown. In fact, despite the advances in sequencing technologies and the progress made in our understanding of microbial genomes, the biological functions of roughly 35% of microbial gene products remain a mystery [54].

Equally impressive is the diversity of secondary metabolites and enzymes produced by microorganisms, well-known to mediate the Earth's fundamental biogeochemical cycles [1, 27]. Recent evidence suggests that microorganisms may harbor up to one million biosynthetic gene clusters (BGCs) [23, 55], few of which have been thoroughly described [23].

One can anticipate that these genetic elements may lead to the discovery of novel compounds and biological pathways with potential applications in agronomic, biotechnological, environmental, and pharmaceutical industries [1, 26–29]. In fact, the key to solve some of the current societal challenges may lie on novel metabolite-producing or catalytic gene products of biotechnological value [40, 41, 56–58], such as those hidden in the FDM, as previously summarized [40, 59].

Here, we outline the challenges, opportunities, and unique potential of the hidden world of FDM. We also discuss the limitations of molecular and computational approaches that currently preclude the comprehensive characterization of the MDM. Finally, we offer a perspective on the potential opportunities and directions for future research.

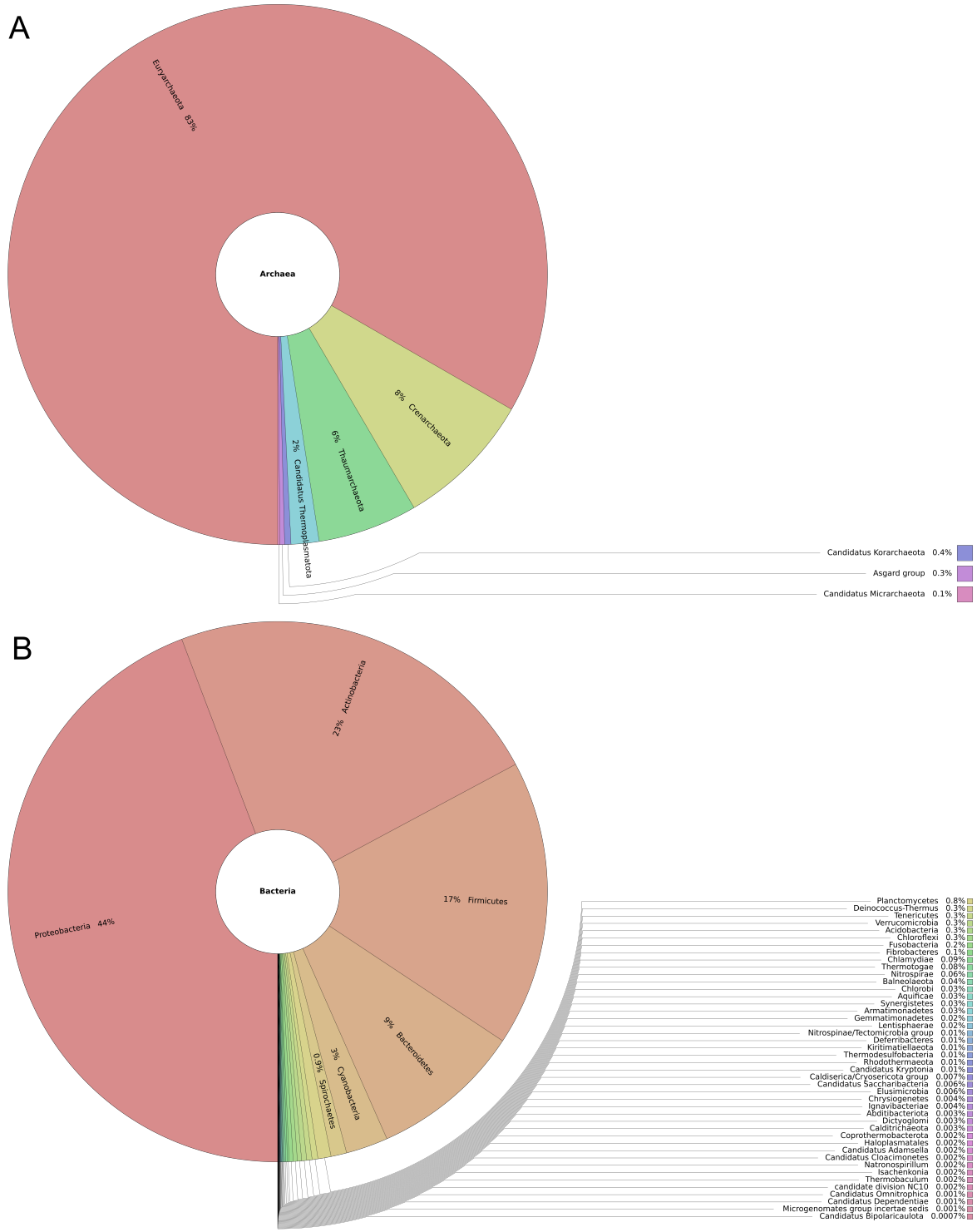


Figure 1.2: Krona plots depicting the taxonomic provenance at the phylum level of FDM proteins from NCBI's RefSeq database, for Archaea (**A**) and Bacteria (**B**) domains. The total number of proteins is 566,678 for Archaea (**A**), and 27,001,444 for Bacteria (**B**). These counts were gathered as follows. First we downloaded all gunzipped, non-redundant protein GenPept flat-files (i.e., `*.nonredundant_protein.*.protein.gpff.gz`) from NCBI's FTP server for RefSeq's release 206 (downloaded on 26th of May 2021), for both Archaea and Bacteria domains, available at `ftp.ncbi.nlm.nih.gov/refseq/release/archaea/` and `ftp.ncbi.nlm.nih.gov/refseq/release/bacteria/`, respectively. The total amount of files retrieved was 1,287. Next we parsed each file programmatically and retrieved the taxonomy lineage field for the proteins whose description matched any of the following words associated with FDM: 'hypothetical', 'predicted', 'putative', 'uncharacterized', or 'unknown function'. We then grouped these proteins by their domain and phylum as issuing from the taxonomy lineage field, counted the number of occurrences for each group, and saved this data to a `.tsv` file to be used by Krona [60] as input.

1.2 Revealing the hidden potential of functional dark matter

Over the past decade, the growth in genomic data acquisition has been staggering. Estimates indicate that the total amount of sequencing data doubles approximately every seven months, spanning from small groups producing a couple of terabases per year, to large, dedicated institutes generating several petabases a year [61].

However, the quality of the data is heterogeneous and the information itself is largely spread among databases and repositories, making the task of centralizing all sequencing information in a single public archive a burdensome and nearly infeasible task. Thus, most sequencing data repositories and databases contain large amounts of data requiring further analysis. However, the gene products imputed from those data are often classified as "uncharacterized", "putative", "predicted", "hypothetical", or simply "unknown" [62, 63].

1.2.1 Current estimates

In some cases the fraction of genes of unknown function can amount to as much as 50%, such as in newly sequenced genomes [64] and uncultured microbial candidate taxa [35, 36, 42, 65, 66]. Archaeal genes, for instance, are reported to encode from 30% to 80% of FDM [63]. In comparison to Bacteria, this represents a greater content of FDM harbored by archaeal genomes, as accounted by Makarova et al. [63].

The same authors also announced that in most of the Archaea, the amount of FDM scales linearly with the genome size [63]. These occurrences were attributed to the difficulty in isolating and cultivating most Archaea, which in turn further hinders the experimental characterization of their genes [63].

In more extreme cases, as indicated by metagenomic studies, these percentages can span from 85% [67] to 99% of total gene content [68, 69]. These estimates, however, highly depend on several factors, including the (meta)genome complexity, sequencing depth, coverage, data quality, read length, the extent of taxonomic novelty, (meta)genomic and environmental context, and the sensitivity and accuracy of the computational methods used for sequence annotation [67]. Moreover, for the prokaryotes that cannot be cultured yet, the sole source of information available to annotate their genome is oftentimes based on comparison to known operon contexts and protein sequences [63].

However, conventional homology-based methods often fail to annotate uncharacterized genes due to a lack of recognizable homologs in public databases, precluding meaningful biological interpretations [67]. As noted by Lobb et al., genes not sharing homology to any known sequences are the most challenging to annotate [67]. In some occasions, such genes have been predicted to be open reading frames (ORFs) of hypothetical function, the so-called "orphan" ORFs, or simply "ORFans" [70].

Figure 1.3 shows the evolution over time of the number of protein sequences belonging to the Archaea and Bacteria FDM, as gathered from NCBI's "protein" database. Note that the number of FDM sequences, as well as the total number of sequences, tend to increase over time, regardless of biological domain. As of 2020, the percentage of FDM sequences in NCBI's protein database is greater than 30% for Archaea, and nearly 20% for Bacteria.

Since the commercialization of next-generation sequencing (NGS; i.e., 2005, see [71]), these percentages have been steady for Archaea but decreased in the long run for Bacteria. We speculate that the decrease in the proportion of FDM sequences for the Bacteria domain is a product of the representativeness of the total number of sequences from this domain in public databases and, consequently, the improvement and fine-tuning of annotation pipelines aimed particularly at this domain.

1.2.2 Possible functions

Uncharacterized prokaryotic genes were initially considered to be mere "junk" elements, pseudogenes, or misannotations [72–74] owing to the narrow understanding at the time of the functional sequence space [67]. But increasing evidence indicates that yet-unclassified elements do encode for specific functions [75].

For instance, Hanson et al. showed that nearly 15% of *E. coli* enzymes of unknown function could play roles in metabolite repair, as described in the authors' enzyme role classification survey from 1998–2015 [76]. Another metabolic role thought to be played by these elements is the addition and removal of posttranslational modifications (PTMs) [77]. In this respect, it should be emphasized that enzymes responsible for recently discovered PTMs usually remain unidentified [77, 78]. Likewise, one can speculate that some uncharacterized proteins may have been enzymes that lost their catalytic properties throughout evolution and then acquired allosteric regulation properties thereafter [77, 79].

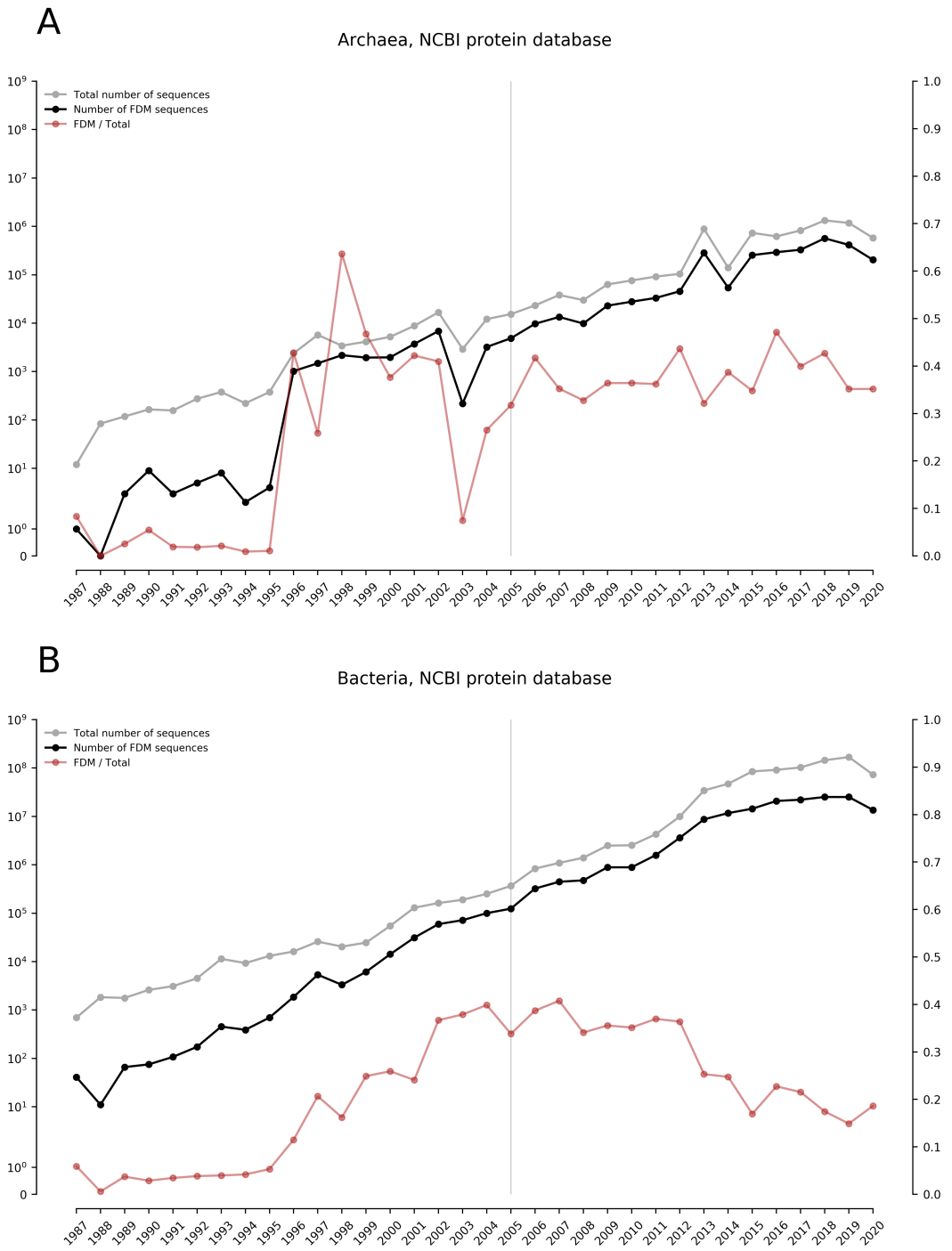


Figure 1.3: Time-series plots showing the number of protein sequences of unknown function in comparison to the total number of sequences in NCBI's protein database. These counts refer to sequences in the Archaea (**A**) and Bacteria (**B**) domains from 1987 to 2020, inclusively. The x-axis represents the date of submission, the y-axis to the left represents the number of sequences in "symlog" scale (i.e., linear from 0 to 1, and logarithmic from 1 upwards), and the y-axis to the right represents the ratio between the number of FDM sequences and the total number of sequences. The light-grey vertical line placed in 2005 marks the commercial availability of the first NGS platform (the pyrosequencing method by 454 Life Sciences, now Roche). In order to obtain these counts we submitted queries to NCBI's Esearch utility at Entrez (`eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?`). For each year and for each domain, we retrieved the total number of sequences with the following query: `db=protein&term=YEAR[pdat]+AND+DOMAIN[orgn]&rettype=count`; whereas as to gather the number of sequences of unknown function we used: `db=protein&term=YEAR[pdat]+AND+DOMAIN[orgn]+AND+(hypothetical[title]+OR+predicted[title]+OR+putative[title]+OR+uncharacterized[title]+OR+unknown+function[title])&rettype=count`.

Other proteins, commonly referred to as "moonlighting", play multiple biochemical or biophysical roles which are not associated with gene fusion events nor proteolytic fragments [80]. Moonlighting proteins are thought to include a few hundred members and play an extensive range of functions [80, 81]. These circumstances, as adverted by Ellens et al., may hold back possible breakthroughs in the field, owing to a substantial increasing number of uncharacterized catalytic elements revealed by sequenced genomes [77].

Proteins of unknown function have also been presumed to encode for ecological or taxon-specific functions [82], including morphological and developmental adaptations [67, 83, 84], which could explain their lack of homology to annotated genes [67]. However, this latter point can also arise as a consequence of the severe undersampling of Earth's microbiomes, despite the advances in metagenomics in the past decades.

Furthermore, previous studies have hypothesized that ORFans, or sequences for which no specific function has been assigned, can actually be of viral, integrative, or mobilomic origin [85, 86]. Additional support for this possibility is elsewhere [87]. This idea makes sense from the biological point of view if one considers the fast mutation rates observed in viral DNA (and RNA), and the underrepresentation of sequences from viral origin in most public databases [67].

Some of these uncharacterized proteins are also thought to be gene fragments, or pseudogenes arising naturally through gene degeneration [63]. There is also evidence that these unknown elements might be genes that evolve fast, like those associated with anti-parasite defense [44], and those that encode small proteins [63, 88]. Indeed, it is often difficult to discriminate between protein coding ORFs and random in-frame genomic fragments [88]. For this reason, most functional prediction tools require a minimum sequence length, which culminates in incomplete databases [88].

Focusing on the study of small proteins ($\leqslant$ 50 amino acids in length), Sberro and colleagues probed more than one thousand human-associated metagenomes [88]. They reported more than four thousand conserved protein families, most of them novel. In the process, they described an abundance of putative functions, namely: housekeeping, mammalian-specific, cell-cell crosstalk, adaptation, anti-parasite defense, secreted or transmembrane proteins, as well as possible products of horizontal gene transfer [88]. Despite this breakthrough, they outlined that over 90% of these families lack a domain assignment, and nearly half of them are absent from reference genomes [88].

1.2.3 Biotechnological significance

Consequently, one can assume that, just as genomic elements of yet unassigned function might encode for any potential function, they can also code for a myriad of unknown properties and functions still waiting to be charted in the microbial sequence space. One such property regarded as being crucial is that of biotechnological potential [28, 36, 40, 59]. Indeed, this variety of genes of unknown function may be an outstanding asset for identifying novel metabolite-producing proteins and enzymes responsible for new physiological mechanisms in a broader sense.

These new functions may lead to escalating innovative biotechnological solutions in a world of ever-increasing societal demands. Indeed, Pascoal et al. argued that bioremediation and bioprospecting are the two areas holding the most promise from within the MDM, where innovative approaches in biotechnology might arise [89], like new solutions for the decontamination of environments [90].

A prominent example that is as promising as it is challenging is the prospect of bioremediating ecosystems polluted with plastic [56]. For instance, polyethylene terephthalate (PET) is reported to be the most abundant polyester plastic [91], being mainly used in the textile and packaging industries [56, 91]. Nearly 70 million tons of PET are manufactured worldwide per year [91]. The common recycling process of PET is through thermomechanical means, resulting in its loss of mechanical properties [92]. As such, PET is preferentially synthesized *de novo*, and its waste continues to accumulate in ecosystems throughout the globe [91, 93].

In 2016 a research team screened natural microbial communities at a PET bottle recycling site and managed to isolate a new bacterium named *Ideonella sakaiensis* (strain 201-F6) [93]. Further characterization of *I. sakaiensis* revealed that it could use PET as a primary energy and carbon source [93]. After assembling the draft genome sequence of *I. sakaiensis*, the team identified an ORF that putatively encoded a hydrolase [93].

Upon recombinant expression of this protein, the team observed that it exhibited PET-hydrolytic activity, thus naming it PET hydrolase (PETase) [93]. PETase catalyzes the hydrolysis of PET into its monomeric component mono-2-hydroxyethyl terephthalate (MHET) [93]. The genome of *I. sakaiensis* also coded for another enzyme that was capable of degrading MHET, which was designated MHET hydrolase (MHETase) [56, 93]. MHETase hydrolyzes MHET into its two monomers, terephthalic acid (TPA) and ethylene glycol (EG) [93], which

are used by *I. sakaiensis* in its metabolism [56]. Moreover, TPA and EG have the potential to serve as novel substrates that can be converted into value-added products [94]. EG, for instance, can be used for numerous applications, such as a coolant in antifreeze [94].

It is worth noting that even though multiple studies have described enzymes that can degrade PET, the connection of extracellular enzymatic PET degradation to catabolism in a single microbe was hitherto unheard of [95]. However, the recently reported crystal structure resolution of MHETase is likely to possess a scaffold that is unprecedented for plastic-degrading enzymes [96]. This example shows that the prospection of the diversity harbored by the MDM offers a promising source for the identification of pollutant-degrading enzymes that could be used for bioremediation [56].

Another relevant example of the biotechnological potential of the MDM is that of the isolation chip (Ichip) [97]. In 2010 the Ichip was developed by a research group for the *in situ* cultivation of microbes that had eluded previous standard culture efforts [97, 98]. This isolation chip comprises several hundreds of miniaturized chambers and each chamber harbors one or few cells from a given environmental sample. Each chip harbors cells from a single environment.

Chamber incubation is carried out in the environment from where the cells were taken, allowing growth factors and other molecules to diffuse throughout the semipermeable membranes covering the chambers, thus facilitating growth and increasing the recovery of uncultured microbes [59]. The discovery of the antibiotic Teixobactin [24] not only demonstrates the efficacy of the Ichip but also highlights the potential of uncultured bacterial communities.

1.2.4 Catalytic prospectives

Enzymes are the backbone of numerous industries [58, 99–106]. Reactions catalyzed by enzymes are thought to follow the rules of green chemistry—they are safer, faster, and generate less waste than traditional methods [40]. The unmatched eco-friendly potential of enzymes is of vital use in the industry to mitigate the rampant overconsumption of our planet's resources [40].

Presently, there are more than two-hundred types of enzymes of microbial origin that are commercially available [106], of which about 20 types are produced on industrial scales [58, 99]. Therefore, the discovery of novel enzymes of biotechnological or industrial interest is

critical for the growth of the industrial enzymes market. This market amounted to 9.9 billion USD in 2019, and it is projected to reach 14.9 billion USD by 2027 [40].

Yet, most of industrial enzymes currently originate from fungi or mesophilic bacteria [40, 41]. Moreover, the majority of enzymes of industrial relevance are hydrolytic in nature (i.e., hydrolases) [58]. Examples of industrially-relevant enzymes that were previously unearthed from the MDM include cellulases [54], lipases [58], alcohol dehydrogenases [41, 57], enzymes that catalyze organophosphorus compounds [107], along with other relevant ones displaying enhanced stability under industrial conditions, as recently reviewed [40]. Table 1.1 presents a list of enzymes of known biotechnological and/or industrial interest, gathered and manually curated from the literature.

As highlighted by Bernard and colleagues, science has profited tremendously from numerous breakthroughs that relied on enzymes of microbial origin, as substantiated by the work of numerous Nobel Prize laureates [36]. Many of these early studies relied on restriction enzymes [108], then on DNA polymerases [109] coupled to the advent of the polymerase chain reaction [110], and more recently on the CRISPR-cas9 system [111]. Hence, microbial gene discovery can greatly push progress and development of new mechanisms and compounds of pharmaceutical, biotechnological, and biomedical relevance.

In a seminal report, Grotzinger et al. proposed a workflow aimed at targeted protein production based on single amplified genomes (SAGs) from species that we cannot culture yet [41]. As a proof-of-concept, they used the method described in their paper to unearth an alcohol dehydrogenase (ADH) from an uncharacterized polyextremophilic archaeon sampled from a brine pool at the bottom of the Red Sea [41].

ADHs are of industrial interest, given their ability to produce chiral compounds for pharmaceuticals and fine chemicals [57]. ADHs can also be used in biosensor-based diagnostics and fuel-cell technologies [41]. Experimental characterization of this ADH not only demonstrated its thermostability, halotolerance, and the ability to withstand the presence of different solvents, but also the prospect of it being stored and used as a powder; all of which are features of biotechnological significance [41].

A more recent paper identified and characterized another ADH of similar polyextremophilic nature and with solvent tolerance, presumed to be a member of a rare enzyme family—that of microbial cinnamyl alcohol dehydrogenases [57].

Table 1.1: Enzymes of biotechnological and/or industrial interest and their applications sorted by alphabetical order. Adapted and manually curated from [99–104].

EC number	Enzyme name	Application
EC:4.1.1.5	Acetolactate decarboxylase	Converting α-acetolactate to acetoin directly. Decreasing fermentation time by avoiding formation of diacetyl.
EC:2.7.4.3	Adenylate kinase	Biological indicator for validation of procedures to inactivate transmissible spongiform encephalopathy agents.
EC:1.4.1.1	Alanine dehydrogenase	Candidate for enantioselective production of optically active amino acids.
EC:1.1.1.1	Alcohol dehydrogenase	Candidate for asymmetric synthesis. Reduction of C-O and C-C bonds.
EC:3.1.3.1	Alkaline phosphatase	Candidate for molecular biology application: dephosphorylation of DNA.
EC:3.2.1.212	Alpha-L-fucosidase	Establishing glycosidic bonds.
EC:3.2.1.1	Alpha-amylase	Additive in food, textile, detergent and bioremediation industries. Waste-water treatment, drainage. Molecular biology applications. Treatment for digestive disorders.
EC:3.2.1.22	Alpha-galactosidase	Additive in soybean foodstuff.
EC:3.5.1.4	Amidase	Acylation, deacylation, enantioseparation. Degradation of nitrile-containing wastes.
EC:3.2.1.3	Amyloglucosidase	Glucose production. Increasing glucose content in beverages. Additive in toothpastes, mouthwashes, and bioremediation.
EC:4.3.1.1	Aspartase	L-aspartic acid production.
EC:3.2.1.2	Beta-amylase	Producing low-molecular weight carbohydrates. Starch hydrolysis. Cleaving α-1,4-linkages from non-reducing ends of amylose, amylopectin and glycogen molecules.

Continued on next page

Table 1.1 – Continued from previous page

EC number	Enzyme name	Application
EC:3.2.1.21	Beta-glucosidase	Production of ginseng compounds for medical applications.
EC:3.5.2.6	Beta-lactamase	Molecular biology applications by conferring antibiotic resistance to Beta-lactam antibiotics.
EC:3.4.22.32	Bromelain	Additive in the cosmetic industry.
EC:4.2.1.1	Carbonic anhydrase	Candidate for biomedical applications.
EC:1.11.1.6	Catalase	Candidate for textile and cosmetic industries. Antioxidants. Bleach termination. Cheese processing.
EC:3.2.1.4	Cellulase	Additive in food, detergent and textile industries. Deinking. Drainage improvement. Degradation of cellulose in the textile industry.
EC:3.4.23.4	Chymosin	Cheese manufacturing.
EC:3.1.1.74	Cutinase	Triglyceride removal. Degradation of plastics, polycaprolactone. Additive in the textile industry.
EC:6.5.1.1	DNA ligase (ATP)	Candidate for molecular biology applications.
EC:2.7.7.7	DNA-directed DNA polymerase	DNA amplification used in the polymerase chain reaction and recombinant DNA technologies.
EC:3.3.2.10	Epoxide hydratase	Candidate for the production of enantiopure epoxides in the pharmaceutical industry.
EC:3.1.11.1	Exodeoxyribonuclease I	Candidate for molecular biology application: 3'-5' exonuclease specific for single-stranded DNA.
EC:3.1.13.1	Exoribonuclease II	Antiviral agent. Candidate for molecular biology applications.
EC:1.17.1.9	Formate dehydrogenase	Oxidation of alcohols and oxygenation of C–H and C–C bonds.
EC:4.1.2.13	Fructose-bisphosphate aldolase	Establishes C–C coupling.

Continued on next page

Table 1.1 – Continued from previous page

EC number	Enzyme name	Application
EC:1.1.3.4	Glucose oxidase	Dough strengthening. Used in toothpastes and mouthwashes. Oxygen removal from beer. Polymerization of anilines. Detection of glucose in blood. Bleaching agent.
EC:3.5.1.2	Glutaminase	Cancer chemotherapy, particularly for leukemia.
EC:1.8.3.3	Glutathione oxidase	Used in hair waving.
EC:1.11.1.9	Glutathione peroxidase	Antioxidant properties.
EC:1.8.1.7	Glutathione reductase	Candidate as an antioxidant enzyme in heterologous systems.
EC:3.2.1.68	Isoamylase	Hydrolyzing α-1,6-linkages in glycogen and amylopectin.
EC:3.5.1.1	L-asparaginase	Cancer chemotherapy, particularly for leukemia.
EC:1.10.3.2	Laccase	Non-chlorine bleaching, delignification. Additive in food, textile, cosmetic, and pesticide industries. Degradation of waste containing olefin unit, polyurethane and phenolic compounds.
EC:3.2.1.108	Lactase	Lactose hydrolysis in dairy products or whey to avoid lactose intolerance. Antitumor agent.
EC:1.1.2.4	Lactic acid dehydrogenase	Reduction of C–O and C–C bonds.
EC:1.4.1.9	Leucine dehydrogenase	Candidate for medical and pharmaceutical industry applications.
EC:1.11.1.14	Lignin peroxidase	Degradation of phenolic compounds.
EC:3.1.1.3	Lipase	Additive in the food, detergent, cosmetic, textile, pharmaceutical, polymer, biodiesel, biosurfactant, pulping, and fossil-fuel industries.
EC:1.13.12.8	Luciferase	Molecular biology applications such as bioluminescent assays involving ATP.

Continued on next page

Table 1.1 – Continued from previous page

EC number	Enzyme name	Application
EC:3.2.1.17	Lysozyme	Antibiotic. Disruption of mucopeptide in bacterial cell walls. Cheese manufacturing.
EC:1.1.1.37	Malate dehydrogenase	Candidate for detection and production of malate.
EC:3.2.1.20	Maltase	Additive in detergent and food industries. Production of glucose from maltose.
EC:3.2.1.133	Maltogenic alpha-amylase	Enhances shelf life of bread.
EC:1.11.1.13	Manganese peroxidase	Degradation of phenolic compounds.
EC:3.2.1.25	Mannanase	Additive in food, detergent and textile industries.
EC:3.4.24.3	Microbial collagenase	Treatment for skin ulcers. Wool finishing.
EC:3.1.1.102	Mono(ethylene terephthalate) hydrolase	Conversion of PET monomers into terephthalic acid and ethylene glycol.
EC:3.5.1.14	N-acyl-aliphatic-L-amino acid amidohydrolase	Production of L-amino acids.
EC:3.2.1.40	Naringinase (alpha-L-rhamnosidase)	Acting on compounds that cause bitterness in citrus juices. Debittering.
EC:3.2.1.135	Neopullulanase	Acting on both α-1,6- and α-1,4-linkages.
EC:4.2.1.84	Nitrile hydratase	Degradation of nitrile-containing wastes. Used in acylation, deacylation, enantioseparation. Synthesis of acrylamide, butyramide, and nicotinamide.
EC:3.4.22.2	Papain	Additive in the cosmetic industry.
EC:4.2.2.2	Pectate lyase	Bioscouring. Candidate for the detergent industry.

Continued on next page

Table 1.1 – Continued from previous page

EC number	Enzyme name	Application
EC:3.2.1.15	Pectinase	Destabilizing the outer cell layer to improve fiber extraction via depectinization. Additive in food industries, such as clarification of juice and increasing its overall production, in the process of vinification, and the mashing of fruits.
EC:3.5.1.11	Penicillin acylase	Semi-synthetic penicillin production/broad-spectrum antibiotic production.
EC:1.11.1.7	Peroxidase	Hair dyeing. Quantification of hormones and antibodies.
EC:1.11.1.24	Peroxiredoxin	Candidate for food and pharmaceutical industries.
EC:3.1.3.26	Phytase	Candidate for feed applications, especially in aquaculture. Hydrolysis of phytic acid to release phosphorus, calcium, and magnesium cations.
EC:3.1.1.101	Poly(ethylene terephthalate) hydrolase	Biodegradation of PET polyester plastic into monomers.
EC:1.10.3.1	Polyphenol oxidase	Hair dyeing.
EC:5.3.4.1	Protein disulfide-isomerase	Hair waving.
EC:3.2.1.41	Pullulanase	Additive in food and biofuel industries. Attacking α-1,6-linkages, liberating straight-chain oligosaccharides of glucose residues linked by α-1,4-bonds.
EC:2.8.1.1	Rhodanese	Cyanide poisoning treatment.
EC:1.10.3.6	Rifamycin-B oxidase	Antibiotic synthesis.
EC:3.1.2.12	S-formylglutathione hydrolase	Candidates for chemical synthesis and industrial pharmaceutics.
EC:5.3.1.28	Sedoheptulose-7-phosphate isomerase	Candidate for biocatalysis under low water conditions.

Continued on next page

Table 1.1 – Continued from previous page

EC number	Enzyme name	Application
EC:2.1.2.1	Serine hydroxymethyl-transferase	Candidate as a pharmaceutical, agrochemical and food additive.
EC:3.4.24.40	Serralysin	Antiviral and anti-inflammatory properties.
EC:3.2.1.18	Sialidase	Hydrolysis of glycosidic linkages of terminal sialic acid residues in oligosaccharides, glycoproteins, glycolipids, colominic acid and synthetic substrates.
EC:3.4.21.62	Subtilisin	Additive in food, textile, leather, detergent, and cosmetic industries. Degrading protein into its constituent peptides and amino acids to overcome antinutritional factors.
EC:1.15.1.1	Superoxide dismutase	Anti-inflammatory and antioxidant properties. Free radical scavenging. Candidate for applications in agriculture, cosmetics, food, healthcare products and medicines.
EC:3.4.24.27	Thermolysin	Aspartame production.
EC:2.3.2.13	Transglutaminase	Hair waving. Protein cross linking. Laminated dough strengthening.
EC:5.3.1.1	Triosephosphate isomerase	Candidate for biocatalysis under low water conditions.
EC:3.4.21.4	Trypsin	Anti-inflammatory and anti-coagulant properties. Molecular biology applications. Food processing.
EC:1.14.18.1	Tyrosinase	Tumor-associated antigen. Polymerization of lignin and chitosan.
EC:3.2.2.27	Uracil-DNA glycosylase	Candidate for molecular biology application: release of free uracil from uracil-containing single-stranded or double-stranded DNA.
EC:3.5.1.5	Urease	Urea quantification in body fluids.

Continued on next page

Table 1.1 – Continued from previous page

EC number	Enzyme name	Application
EC:1.7.3.3	Uricase	Treatment of hyperuricemia.
EC:3.4.21.73	Urokinase	Removal of fibrin clots from bloodstream. Anti-coagulant properties.
EC:3.4.24.25	Vibriolysin	Additive in food, textile, leather, and detergent industries.
EC:3.2.1.32	Xylanase	Additive in food, textile, detergent, pulp and bioremediation industries. Hydrolyzing pentosans of malt, barley and wheat. Enhancing pulp-bleaching efficiency.
EC:5.3.1.5	Xylose isomerase	Production of high-fructose corn syrup. Catalyzing isomerization of glucose to fructose.

1.2.5 Biosynthetic Gene Clusters

Prokaryotes and other microorganisms (e.g., Fungi) are known to produce many secondary metabolites (SM) [28]. These SMs are natural products that encompass diverse chemical structures [28]. This chemical diversity allows SMs to perform a plethora of functions [28]. SMs may have antibiotic, anti-cancer, anti-viral, antifungal, antioxidant, anti-trypanosome, cholesterol-lowering, immunosuppressant, insecticide, and herbicide properties, among many others [26, 28, 112, 113].

BGCs are the physical grouping of the genes that encode all enzymes required to produce a SM, including pathway-specific regulatory genes [28, 114]. Two major biosynthetic systems containing multiple modules and enzymes are those of polyketide synthases (PKS) and nonribosomal peptide synthases (NRPS) [28, 115].

PKS and NRPS synthesize the two major classes of SMs, namely polyketides (PK) and nonribosomal peptides (NRP) [28]. PKS and NRPS are also popular targets in bioprospecting endeavors, given their reputation as producers of a broad range of SMs with important applications in healthcare and research [28]. PK and NRP, together with terpenoids and alkaloids, were regarded as the four major groups of SMs throughout the 20th century [116].

Yet, the NGS revolution at the turn of the 21st century unveiled another major class, that of ribosomally synthesized and post-translationally modified peptides (RiPPs), which have since attracted increasing interest [116, 117]. This interest stems from academic and industrial sectors alike, due to the structural variability and functional diversity shown by RiPPs [118].

The chemical space of RiPPs is determined by their nucleotide sequence, therefore linking the diversity of these small molecules with that of genes [118]. This genetically-encoded nature of RiPPs enables researchers to freely manipulate the scaffolds of the peptides by site-directed mutagenesis and efficiently screen the targets for those possessing characteristics of interest [118].

Even though several computational tools are able to accurately identify BGCs [113], these are not without limitations, as noted by Chen et al. [28]. For instance, these tools might rely on external databases and rules extracted from previous knowledge [28]. This reliance on pre-existing databases and algorithms implies that only known biosynthetic pathways whose rules are implemented in the software are detected [28]. Thus, biosynthetic pathways that make use of unidentified enzymes (i.e., those from the FDM) will be missed [28, 119]. Moreover, the bioprospection of BGC data from metagenomes is challenging, as the computational tools that do so commonly require high-quality genomes or those resolved from metagenomes as input [28, 119].

Another major challenge in natural product discovery is that a substantial portion of BGCs are transcriptionally silent or expressed at very low levels when in a standard laboratory setting [28, 120]. Thus, strategies designed to activate these silent BGCs are crucial for discovering new chemical scaffolds [29].

Nonetheless, there are successes arising from the systematic interrogation of BGCs from within the FDM [29]. For instance, a research team developed an algorithm that catalogs RiPP biosynthetic gene clusters [121]. Upon analyzing 65,000 prokaryotic genomes, they unearthed RiPP BGCs that coded for more than two-thousand novel natural products [29, 121].

Bull and Goodfellow have studied BGCs while focusing their bioprospecting efforts on the phylum Actinobacteria [26]. Their rationale for focusing on this taxon is fivefold: (i) the recurrent and foundational role of actinobacteria in soil ecosystems; (ii) the size and diversity of the taxon; (iii) the ceaseless discovery of new taxonomic radiations; (iv) the BGC-rich genomes of Actinobacteria; (v) and their unparalleled track record as producers of bioactive compounds

of notable ecological and economic value [26].

SMs discovered from Actinobacteria—especially *Streptomyces* strains—account for two thirds of known antibiotics [122], including those in clinical use today [29]. Actinobacteria are also known to produce roughly tenfold as many specialized metabolites as those known from laboratory experiments [29]. Consequently, this ability has renewed interest in these prokaryotes as producers of new chemical entities [29].

Indeed, Bull and Goodfellow emphasize that SMs have a significantly greater diversity and quantity of chemical scaffolds than those produced by combinatorial synthetic compounds, thus providing a compelling reason for prioritizing them in the search of novel drugs [26]. Their research into actinobacteria recovered from two extreme environments (i.e., deep sea and hyperarid desert) has uncovered a remarkable assortment of new members of chemical classes, and each of those products is either a new-in-a-class or first-in-a-class chemical entity [26]. The most widely distributed bioactivity of these compounds is that of antibacterial and anticancer activities [26]. Moreover, these compounds are also putative drug hits that could provide potential therapeutic targets for inflammatory diseases, Alzheimer's disease, and type II diabetes [26].

These discoveries further elevate the foregoing rationale, underlining the notion that the exploitation of biological know-how from within the FDM can offer an unprecedented range of biotechnological solutions that might not only be at the core of new markets and business models [123], but also at the bleeding edge of innovation in new therapeutics, industrial applications, and bioremediation strategies.

1.2.6 Additional insights into the functional dark matter

In the past few years, deep investigation of the FDM has prompted an exciting scientific revolution regarding not only the bioprospection of gene products of biotechnological interest, but also invaluable insights into the MDM in a broader sense.

For instance, Makarova et al., in a study of uncharacterized genomic "islands" from archaeal genomes, found that besides being highly abundant and comprising a heterogeneous gene pool of diverse putative functions, these islands also code for defense systems along with new variants of the CRISPR-Cas genome editing system [44].

Recent articles describe synthrophic networks in anaerobic methanogenic consortia of uncultured microorganisms [45, 48], as well as in benzene-degrading settings [124], revealing the potential applications for anaerobic bioreactors aimed at bioremediation and energy generation [45].

Another study has predicted metabolic roles for multifaceted chemoorganoheterotrophic bacterioplankton that would be involved with degradation of complex carbon compounds and the nitrogen cycle [49].

A different study revealed that uncultivated ultra small marine prokaryotes encoded for a wealth of gene homologs associated with diverse metabolic pathways such as carbon, methane, nitrogen, and sulfur, further highlighting that these little known prokaryotes presumably contribute to elemental cycling [125].

Wong et al. reconstructed 115 genomes assembled from hypersaline microbial mat metagenomes and uncovered putative dissimilatory sulfate reduction in surface hypersaline settings, novel eukaryotic signature proteins in the Asgard archaeal superphylum, organic carbon metabolism, many forms of RuBisCo (ribulose-1,5-bisphosphate carboxylase-oxygenase), high hydrogen production capacity, putative schizorhodopsins, and diversity-generating retroelements [126].

Around the same time, Wiegand and colleagues characterized and sequenced the genome of 79 bacterial strains from the enigmatic bacterial phylum Planctomycetes [34]. The authors identified previously unknown modes of bacterial cell division—such as lateral budding and binary fission—along with previously unidentified cell signaling processes and small-molecule production, asserting that planctomycetes are drastically different from model bacteria [34]. Their study also advanced that the vast majority of putative BGCs encoded by planctomycetes differ from known BGCs, hinting at an untapped potential for small molecule production [34].

Lackner and colleagues have described uncultured synthrophs of marine sponges as key producers of natural products with exceptional pharmacological potential, bearing a level of chemical richness similar to that of soil actinomycetes [25].

Similarly to the latter study, a different research group has identified extensive drug discovery potential amidst the microbiome of a different marine sponge, evidencing a wealth and breadth of untapped resources for novel chemistry [27].

Also worthy of mention are other studies that identified hundreds of metalloproteases with

signature catalytic motifs within ORFans, obtained from human gut, marine, and soil metagenomes [67]; probed uncharacterized groups of Acidobacteria displaying extensive carbon catabolic abilities, including polysaccharide breakdown and metabolism of lignin derivatives [46]; characterized new members of the *Oceanospirillales* order whose genomes code for enzymes capable of metabolizing crude oil [127]; and lastly, described exceptionally high diversity of actinobacteria in the highly arid Atacama desert, outlining its remarkable significance for future biodiscovery campaigns [128, 129].

1.3 Progress and pitfalls

There are numerous reasons why the function of many protein sequences has not been characterized yet. The problem starts with current sequencing technologies and nucleic acid extraction protocols, which have limitations, but also significant are most mainstream public databases and associated bioinformatics methodologies that in many cases are subject to unsupervised dissemination of information, and that includes the spread of misannotations.

In the following section we will discuss "how" and "why" these approaches might be fostering this predicament, and contribute to an ever-increasing inflation of the problem of protein sequences with no characterized function.

1.3.1 Metagenomics

Until recently, genome sequencing of microorganisms greatly relied on their culturability [5]. Indeed, until the turn of the century, researchers who aimed at exploring the richness of genomic and physiologic information shrouded amidst microbial communities often had to isolate the organisms from a given environmental sample and axenically cultivate them in order to sequence their genome. Even though full genome sequencing of microbial isolates yields high-quality data [130], it leaves out the unculturable realm of microbial communities.

However, recent progress in culture-independent DNA sequencing approaches such as metagenomics and single-cell genomics [5, 130, 131] along with new bioinformatics tools and development of annotation pipelines [132, 133] have improved our awareness of elusive microorganisms and our ability to identify them, as well as the ecological relationships among them and their environment. Each of these technologies and approaches also has their limitations, but when used together they provide discrete, yet highly complementary types of data [130, 131].

Direct DNA sequencing performed from an environmental sample, also known as shotgun metagenomics, is routinely used to survey the genomic content of microbial communities [43, 134]. By combining the resulting sequencing data with other types of data such as the ones on 16S rRNA gene amplicon sequencing, metatranscriptomics, and metaproteomics, a new systems approach for molecular microbial ecology was born and is now commonplace [134, 135].

Two of the major assets inherent to shotgun metagenomic approaches are its versatility, as it can be used with different types of samples and applied when other approaches, such as single-

cell genomics, have failed [131]; and its simplicity in terms of sample preparation, provided that a suitable amount of DNA has been previously extracted.

Due to its inherent characteristics, it would be expected that shotgun metagenomics would grant access to the genomic content of all representative members from any given microbial niche, even if these are organized in compact consortia such as aggregates or biofilms [131].

However, the taxonomic and phylogenetic assignment of sequencing reads can be challenging, keeping in mind that a metagenome represents a bulk of genomic data from a plenitude of diverse microorganisms within the sampled community, with reads of some groups eventually bearing more representatives than others [131]. Indeed, our current ability to classify reads of metagenomic origin is still impaired by undersampling of uncultured and unclassified microorganisms, as indicated by biased genome datasets of reference [43].

The procedure that assigns metagenomic reads to discrete operational taxonomic units, so taxon-specific gene inventories can be created, is commonly called binning, or metagenome classification [43, 134]. Even though the terms metagenome binning and metagenome classification have been used interchangeably, there are studies presenting a clear distinction between the two processes [134].

Metagenome binning consists of employing different clustering approaches that search for sequence compositional characteristics and does not rely on reference sequences; whereas metagenome classification assigns a given taxon to a metagenomic read based on its homology to sequences of known taxonomy [134]. Currently, metagenome binning can be achieved by clustering contigs according to their intrinsic nucleotide composition, namely, codon usage, GC ratio, tetranucleotide frequency, read depth, and even time series [131, 136]. Because the metagenome classification assigns a taxon to a given target sequence according to the homology it shares with known sequences, phylogenetic anchor genes are identified throughout the process [131].

Several examples of the application of these two procedures, whether alone or combined with one another, have provided the scientific community with the first complete or nearly complete genomes from various uncultured extremophiles [137–140], as outlined by Hedlund and colleagues [131].

Notwithstanding the significance and breadth of data gathered using these metagenomic approaches, the interpretation of the information is often complex, and insights obtained from

these microbial assortments are prone to be lost among the resulting metagenome data [141].

Additionally, annotation efforts targeting assembled metagenomic contigs usually rely on homology-estimation algorithms to predict protein function [141, 142], which does not take into account the role played by the organism in its natural environment nor its relationships with other members of the community [141].

Even when metagenomics is applied together with other techniques such as metatranscriptomics, single-cell genomics, and proteomics, in order to provide a "systems-wide" picture, the outcome reflects only the community's most representative reads, given that a substantial fraction of microorganisms' functional diversity is often lost because of confounding factors such as spatial microenvironments and/or phenotypic noise [141, 143].

Also, because metagenomics can only render a snapshot of a given microbial community at a given time, its application may generate further hurdles, from an analytical standpoint, such as difficulty to recover entire genomes and the possibility of masking population diversity [144], neglecting rare taxa [145], misconceiving different spatial assortments, and perhaps most importantly, limiting the real genomic context and functional imputation of unknown sequences [141].

1.3.2 Single-cell genomics

Single-cell genomics and its application in the field of molecular microbial ecology have offered a whole new view of microorganisms by looking at genomics through the basic unit of life: the cell [146]. It thus provides a highly complementary and often synergistic approach to metagenomics [43], given that it operates by isolating a single cell, following amplification of its whole genome [141]. This approach, in its full potential, can also generate complete noncomposite genomes [141, 147].

Given that this technique relies on the ability to amplify and sequence a single DNA molecule from an isolated cell in order to produce high-quality data, four challenging primary steps need to be taken, as remarked by Gawad et al. [146]. These steps are (i) to successfully isolate individual cells; (ii) to amplify their genomes and obtain enough genomic material for the following steps; (iii) to probe the genome in a way that efficiently addresses the scientific hypotheses one wants to test; and (iv) to analyze the sequencing data, whilst being mindful of inherent biases

and errors that could have been introduced during the preceding steps (for an in-depth review on this approach see [146]).

Accordingly, this approach offers a suitable alternative, or complementation [130, 131], to address some of metagenomics' shortcomings, while forgoing its greatest potential—the production of a magnitude of sequencing data [141].

Nonetheless, as reviewed in [141], single-cell genomics has shown to be a major asset for studying yet-uncultured microorganisms, providing both genomic insight and phylogenetic context of the uncultured majority [35, 43, 148]. Although both single-cell genomics and metagenomics can be used as complementary approaches, they both limit the interpretation of the output sequencing data. This approach allows that only the organisms under study have their environmental functional profiles estimated, as predicted by computational tools that impute protein function from sequence input data [141].

Hence, and as mentioned before, none of these techniques can perceive population-specific activity or dormancy, and hence can lead to potential magnification of the environmental role played by taxa with greater number of representatives, whilst neglecting meaningful contributions provided by rare organisms [141, 149], especially those that are uncultured [32, 43].

Indeed, no matter how much metagenomics and other culture-independent approaches improve our current knowledge regarding the uncultivated microbial majority, and the potential they may carry, they also inevitably raise issues, given that every methodological and analytical procedure has its fair share of advantages and drawbacks [36].

For instance, and as previously reviewed [36], steps such as (i) sample selection; (ii) the identification of genes of interest in targeted-sequencing projects; (iii) the filtering cutoffs and expectation thresholds applied to subsequent bioinformatic treatments, besides data standardization; (iv) and the cyclical nature of the methods and scientific questions posed by environmental genomics studies, may increase the probability that true biological novelty and metabolic uniqueness remain buried, and possibly even forgotten, amidst this bulk of raw data that may have been only barely analyzed [36].

Therefore, in spite of the ability of these approaches to unearth both taxonomic and genomic novelty, they fall short regarding determination of protein function, environmental contributions, and activity levels of the microorganisms under investigation [141].

1.3.3 Issues concerning DNA extraction protocols

Perhaps the step of DNA extraction may mitigate potential errors and biases arising from single-cell genomics and metagenomics approaches. In studies of molecular microbial ecology, it is reasonable to favor the enrichment of DNA of bacteria and archaea with a given morphological or physiological trait, prior to sequencing. However, and using metagenomic studies as an example, in many cases the interest may not be in favoring the DNA extraction of a particular type of prokaryote but rather to recover the bulk DNA of the whole community without discriminating for cell wall morphology, niche partitioning, aggregate-forming consortia, or functional activity.

Thus, the DNA extraction method must suit the goal of the study, without compromising downstream data interpretation. For instance, when extracting metagenomic DNA of environmental origin, two general approaches have been considered: (i) either the DNA is directly extracted in bulk from the sample; or (ii) cells are recovered from the sample, and only then are they lysed and the DNA extracted, which led to this latter procedure being called "indirect extraction" [150].

The first method, termed "direct extraction", has significant advantages; it is less time-consuming and yields great amounts of DNA while it also isolates a high yield of non-prokaryotic contaminant DNA [150]. The "indirect extraction", however, can overcome this limitation to a certain extent, since it has been shown to yield less non-prokaryotic DNA than the "direct extraction" method, while still collecting phylogenetically diverse nucleic acids [150, 151]. The latter is nonetheless more laborious and yields less DNA, as it is potentially biased towards microorganisms that are not assorted in an environmental matrix, biofilm, or in conjoined aggregates together with sample material [150].

From a single-cell sequencing perspective, the lysis step of a given DNA extraction protocol can be challenging, requiring the genomic DNA to be intact while dealing with an array of structurally assorted cells [141]. Thus, developing cell lysis protocols that efficiently release DNA from the cell, regardless of the type of cell structure, may increase the potential targets of single-cell sequencing methodologies [141].

The different fragment sizes generated by these approaches is also an important factor to consider. For instance, "direct extraction" protocols usually are more prone to produce shorter fragments of genomic DNA as a result of harsh enzymatic treatments and purification steps,

whereas "indirect extraction" approaches tend to preserve the genomic DNA in larger fragments that are further used to build metagenomic libraries and large prokaryotic DNA templates [150].

Environmental metagenomics has thus typically relied on the use of short reads from "direct extraction" approaches, followed by high-throughput sequencing, in order to characterize functional and taxonomic community groups. This is a robust methodology that nevertheless has shortcomings when factors such as sequencing depth is considered [20], as its potential to unveil the taxonomic and functional characteristics of a given community is restrained by read length [152].

In order to circumvent this hurdle, a strategy relies on assembling the short reads into larger contigs to next undergo routine bioinformatics analysis and annotation [152]. Nevertheless, given that metagenomic samples contain DNA from multiple organisms, these assemblies are sometimes inaccurate as they may include sequencing artifacts such as chimeric contigs resulting from erroneous combinations of reads from multiple species [152].

Recently, the development of fourth-generation sequencing by the nanopore-based sequencing platform [153] has indicated that these limitations may be resolved, since the platform can generate sequencing reads with no limit to its potential length [19]. This technology could indeed revolutionize environmental metagenomics, ultimately providing in-field, single-molecule, real-time sequencing of metagenomes produced by nanopore devices [154].

1.3.4 Rare taxa

Genome and metagenome sequencing techniques allow to probe the taxonomic assortment and coding potential enclosed by prokaryotic communities. Nonetheless, these methods are not only bounded by protocol heterogeneity (e.g., DNA extraction protocols, primers used for amplification), but also especially susceptible to depth bias [21]. The latter is mainly due to the difficulty in detecting prokaryotes that exist in low numbers, also known as the "microbial rare biosphere" [89].

Although scarcely represented, this minority is purported to perform many functions, namely that of a "seed bank" (dormant or metabolically inactive cells that behave as a genomic reservoir) or as "keystone species" (i.e., they have disproportional ecological roles in relation to their numbers), as reviewed by Pascoal et al. [89].

However, a significant proportion of this observed, yet unculturable, diversity might be resultant from stochastic phenomena (e.g., random dispersal), transient existence, they might require long incubation times to form visible colonies, require specific nutrients and physical conditions for growth [155], or quite simply be composed of dying or dead cells [89, 156].

As an illustration of the latter phenomenon, Bellali and colleagues set out to address the hypothesis that some species present in the human gut microbiota remain unamenable to culture because their cells are dead before reaching the end of the gastrointestinal tract, and not exclusively because of culture limitations [155]. They argued that even though metagenomics is able to detect uncultured prokaryotes, it fails at determining whether these organisms are alive or dead [155]. Therefore, these authors combined fluorescence-activated cell sorting (FACS) and culturomics in order to investigate this hypothesis [155]. As a result, they were able to discriminate between live, injured, and dead bacterial groups [155]. Among several conclusions, they outlined that: (i) minority species constituted a substantial portion of the live TDM in human fecal samples; (ii) 28% of bacterial operational taxonomic units in the total fecal samples were either dead or injured; and (iii) roughly two-thirds of the latter group were members of the TDM, consisting in part of anaerobic bacteria—thus potentially explaining why they were missing in culture [155]. Indeed, culturing obligate anaerobes is known to be challenging, given that these organisms often require specific laboratory equipment (e.g., anaerobic chamber) in order to provide an anoxic environment for them to thrive [21].

1.3.5 Culturomics

Despite the extensive breakthroughs provided by metagenomics and single-cell genomics, these techniques still do not allow for the effortless differentiation among strains of the same species, nor do they provide biological material for subsequent research [21]. As thus, axenic culture remains an indispensable step in the complete characterization of a given prokaryote and its genome [21]. A fairly recent approach that aims to reignite the pure culture of microorganisms, particularly those unamenable to conventional techniques, is that of culturomics [21].

Lagier et al. defined culturomics as a culturing approach that uses multiple culture conditions for the identification of prokaryotes that have eluded previous cultivation efforts [21]. Culturomics is thus a high-throughput culturing method that also couples MALDI-TOF mass spec-

trometry and 16S rRNA gene amplicon sequencing to expedite the identification of prokaryotic species [21]. Thus far it has enabled the identification of hundreds of new species (see references in [21]).

Besides expanding our knowledge of the repertoire of the prokaryotic biosphere, culturomics provides pure, viable cultures that can be further used for in vitro experiments [21, 155]. According to the description by Lagier et al., culturomics itself is a workflow composed of several steps, the first of which consisting of dividing and diversifying the environmental sample of interest into multiple culture conditions [21]. These conditions—complex media, broad range of temperatures and incubation times, addition of growth factors, supplements or metals, aerobic/anaerobic setting, etc—are designed to suppress the growth of the most represented populations and to promote the growth of fastidious or rare prokaryotes that exist at lower concentrations [21].

Nonetheless, targeted culture conditions can also be employed to promote the growth of specific taxa [21]. Secondly, swift taxa identification is achieved by MALDI-TOF mass spectrometry, i.e., comparing the protein mass spectra of the target isolate with that of a preexistent database [21]. If MALDI-TOF fails to identify the isolate, the latter is subjected to 16S rRNA gene amplicon sequencing [21]. As a final step, the discovery of novelty is confirmed by sequencing the isolate's genome [21].

Nevertheless, no technique is without drawbacks. In the case of culturomics these include: (i) the sheer workload posed by the multiplex culturing conditions; (ii) the inability to test a multitude of samples like in metagenomics; and (iii) the inability to directly provide data on gene expression and the functional potential enclosed within prokaryotic species by itself—as this always requires further genome sequencing of the newly isolated organisms [21].

1.3.6 Limitations of Heterologous Expression

Heterologous expression of gene products whose functional potential is of interest is a promising approach for the study of particular genomic elements from uncultivated taxa [157], which together with current developments in synthetic biology offer great promise as a strategy for re-coding entire operons for expression in model organisms [131, 158]. For instance, *in silico* analyses have predicted that *E. coli* can transcribe roughly 40% of genes from well-studied cultivated taxa if cloned in expression libraries [141, 159].

However, many obstacles inherent to this approach—particularly the presence of accessory proteins, ambiguous decoding or recoding, apoprotein activation, codon bias, codon reassignment, different promoter structures, inaccurate protein folding, low expression rates, lack of essential post-translational modifications, lack of essential protein secretion, rare codon utilization, shared metabolic pools, or even unforeseen genetic codes [40, 160]—indicate that only a minute percentage of the functional sequence space can be easily identified with this approach [141].

These limitations become even more acute when taking into account uncultured taxa, which may harbor extremely divergent and mostly uncharted DNA sequences and physiologies in a general sense. This is particularly true for extremophiles, given that the experimental characterization of their gene products is further encumbered by the lack of expression systems allowing for the production of recombinant proteins under the same extreme conditions required by their native hosts [41].

These particularities reduce the success rate of current methodologies in accessing the many functional novelties purportedly within the FDM, and leaving the majority of unknown functions and physiological potentials from uncultivated microorganisms concealed within their native expression hosts [141].

1.3.7 Public databases and canonical computational approaches

Most sequence submission portals have their own functional annotation pipelines. As an example, in 2015 the Sequence Read Archive (SRA) [161] amassed more than 3.6 petabases of raw sequence data from numerous sources [61], which akin to other sequence submission portals can often be redirected to integrated annotation pipelines [133, 162], and sometimes, manually curated thereafter [163].

After annotation, the information is added to or updated in different databases harboring distinct biological sequence data such as the NCBI's Nucleotide database [164] and the Translated EMBL Nucleotide Sequence Data Library [165]. Apart from expert-curated records, these databases often include genomic sequences whose annotation is computer-automated and unreviewed, with no explicit function being assigned to these sequences.

A previous report [166] on the Reference sequence database (RefSeq) at NCBI established that genomes and proteins of microbial origin form most of this dataset, even though only

a fraction of their reference bacterial genomes were actually manually curated. This lack of manual curation is understandable if one considers that several hundred million peptides and proteins have been unveiled so far [167].

Thus, it is clear that manual curation is not a viable option to deal with large amounts of sequencing data, least of all to characterize each protein individually in an experimental setting [77]. This circumstance leaves no choice but to employ unreviewed annotation methods, often using computational approaches as the sole alternative to predict the biological role of protein sequences [77].

Indubitably, the main problem concerning the annotation of genomic "unknowns" lies within the annotation pipelines themselves. More often, functional annotation pipelines integrate multiple types of experimental evidence or predictive inference supported by cross-referencing data [168]; one example is the establishment of inheritance of function through homology, as indicated by sequence alignment [169]. Yet regrettably, the functional annotation is frequently of poor quality [170] and requires resource-demanding manual curation to improve and validate it [171].

Strictly speaking, the quality of an annotation depends not only on the quality of the annotation pipelines that generated it, but also on the fast and comprehensive inclusion of new data on the preexistent protein functions that ship with the pipelines in the first place [63]. When function is inferred solely via annotation inheritance through homology, erroneous predictions might arise on many occasions, leading to misannotation [77]. According to Makarova et al., typical annotation errors include, but are not limited to, incorrect gene calling, prediction of genes in the wrong DNA strand, incorrect prediction of ORFs in non-coding sequences (e.g., CRISPR arrays), and erroneous start codon assignment [63].

Unfortunately, this phenomenon has become commonplace in public databases, with the percentage of incorrect functional assignments soaring from less than 5% in 1998 to as high as 40% in 2005 [77, 172]. A study from 2020 showed that up to 50% of protein sequences from public databases contain at least one error [173].

In reality, and in many cases, the function of a given genomic element cannot be simply inferred by common homology-estimation metrics. For instance, if there is no experiment-driven data to support the functional assignment of a given protein that has domains of unknown function, or no statistically robust computational evidence is available, the possibility space is

frequently too large to probe by standard estimation techniques.

Thus, bioinformatics procedures that rely solely on homology imputation based on sequence comparison are not applicable if homologous regions amongst targets of interest are undetectable [174]. Furthermore, the annotations assigned to putative catalytic proteins and predicted enzymes are generally more prone to error *in silico*, leading to several instances of contradictory information reported by different studies [169]. Thus, innovative methods of (re)classification and (re)annotation are greatly needed [175].

However, some challenges still remain, such as the difficulties in accessing the magnitude of unexplored sequences [69]. From a data-mining point of view, and aiming at the biotechnological potential as an example, these hindrances can be overcome. To this end, one needs: (i) to efficiently and effectively prospect and analyze the preexisting data, thus generating reliable information; (ii) to assign function to the gene products revealed by the data; (iii) to correlate these gene products with enzymes of industrial relevance, or a myriad of known secondary metabolites of biotechnological interest; and (iv) to accurately infer, with a certain degree of likelihood, the ones that encode for the most promising compounds [23], along with the ones that might encode for functional novelties from a mechanistic standpoint.

Theoretically, to achieve this outcome, a recurring workflow should exist between bioinformatic methodologies aimed at data issuing from culture-independent sequencing efforts and the corresponding benchwork where the genomic sequences are obtained and scrutinized, in order to guide cultivation efforts of yet-uncultivated, or even entirely new, microbial species [65]. This strategy can be described as a two-step cycle.

First, data produced by characterization of previously cultured microbial isolates should be used to populate public databases with reliable information in order to increment their accuracy [65]. Second, by making use of uncultured genome sequences from metagenomes, it is possible to predict the phenotype of a given species by unearthing its genomic makeup, as well as the putative ecological interactions and lifestyles of its community, rendering substantial information for isolating new microorganisms [65].

Fortunately, a number of initiatives have been undertaken focusing on the copious amounts of publicly available, but neglected, genomic data. These initiatives show that one can indeed shed light on the FDM and uncover latent genomic information [31, 87]. In fact, the vastness of data that can be mined is astounding, with some reports announcing that the percentage of

yet-unknown reads in metagenomic surveys can reach 99% [68, 69].

1.3.8 Logistical hindrances

Besides the foregoing reasons, there is a more logistical reason for this large quantity of metagenomic data to be left unannotated: only a few research projects aim at targeting the unknown sequences in these datasets [69]. This low number of projects is understandable, considering that searching for patterns amidst the data or endeavoring to discover new phenomena is essentially an exploratory science venture, following a "bottom-up" approach that goes from the data to the hypotheses and that not always produces publishable results [36, 176]. Since the purpose of exploratory science is not hypothesis-driven, it can either reveal new knowledge or simply fail in revealing any knowledge at all.

In addition to this, and similarly to what Dutilh stated in his opinion paper [69], researchers commonly find themselves under pressure to publish their results and be able to move to the next project, leaving numerous uncharacterized sequences drowned amidst the contigs that were able to be annotated by homology inference. This happens since using previously annotated sequences leads to straightforward conclusions [69], while overlooking the bioinformatics-to-benchwork working-cycle described above.

Thus, scientists contemplating the investment in purely exploratory studies take great risk and face a precarious situation [36], that keeps perpetuating the problem regarding sequences of unknown function, as explained in this review.

1.4 The path ahead

As reported in a previous review, there are two main dimensions that define the function of a given protein *sensu stricto*—the molecular function and the biological process [77]. Furthermore, expanding on the rationale set forth by the authors of the already mentioned review, in order to functionally annotate a protein one must first (i) gather sufficient evidence produced in the lab to support both aforesaid dimensions to the protein in question; and, more importantly (ii) be sure that these dimensions are reconcilable, otherwise the target protein remains uncharacterized [77]. We agree with this view and that one needs to have some level of skepticism when addressing unreviewed, pipeline-generated annotations, such as those found in public databases.

However, the rate at which proteins of unknown function are accumulating is nothing short of alarming, and experimental evidence—like the one generated by screening assays—cannot simply be produced for every single one of these proteins. The time and resources necessary to accomplish such a task are unimaginable. Additionally, to experimentally validate the molecular function and biological process of each of these proteins would take so long that it would be impossible to keep abreast with the pace at which these uncharacterized proteins are being discovered.

In our opinion, a compromise must be met, and to delineate a less strict definition of protein function annotation should be in order. We hypothesize that, as long as the underlying computational approach bears enough robustness, or makes use of several types of prediction sources, true *in silico* characterization of protein function is possible. This is especially true if the foregoing sources are orthogonal in nature while agreeing on the same function description—e.g., crossing sequence data with physicochemical metadata, predicted domains, protein-protein interaction (PPI) networks, Gene Ontology (GO) terms, Enzyme Commission (EC) numbers, genomic context, to name a few.

1.4.1 *In silico* protein function prediction

It is true that relying on sequence homology alone to annotate a yet unclassified target based on information from a known classified protein sequence is neither an adequate nor a robust strategy [77]. For this reason, several *in silico* approaches have been designed to partially

uncover or impute functional properties of genomic elements without explicitly using inferred information obtained with homology-estimation methods.

In the past these approaches have included, but have not been restricted to, predicting protein function by (i) detecting functional linkage between gene products [177]; (ii) fingerprinting protein domain signature motifs [178]; (iii) using protein association networks [179]; (iv) predicting metabolic activity based on enzyme structure [180]; (v) integration of multiple data sources [181]; (vi) analysis of the genomic context [182].

Of all these methods, the ones that rely on using the amino acid sequence as the only input feature are the most challenging to develop [183]. For these, machine-learning approaches [184, 185], as applied in the field of genomics [184], are reliable for inferring protein function from the sequence itself [183]. Nonetheless, common machine-learning approaches are faulty when used for the manipulation of raw data [186]. In fact, the development of the field of machine-learning has counted on a classification system built upon substantial knowledge of the subject at hand, along with meticulous fine-tuning to design a feature extractor that could not only transform the raw data into a fitting representation to be processed internally by the system, but which would also have the ability to identify unforeseen patterns within the data itself [186].

Nowadays, most machine-learning methods applied to protein function use features gathered from metadata associated with the primary protein sequences in order to train their models; and in due course they use those same models to predict the features of the protein sequence used as input [183], including its domains, putative protein-protein interactions, subcellular location, physicochemical properties (e.g., hydrophobicity), and secondary structure [183].

Presently, there are numerous protein function classifiers that make use of many different types of machine-learning techniques. As an illustration, these classifiers can be based on methods such as (i) association rule-based learning [187]; (ii) naïve Bayes classification [188]; (iii) k-nearest neighbor [181]; (iv) and support vector machines [189] (as summarized in [183]).

1.4.2 Deep Learning in a nutshell

The current state-of-the-art of protein function prediction makes use of a broad family of machine-learning methods called Deep Learning [186]. These methods are based on Representation Learning (or Feature Learning)—a set of techniques relying upon artificial neural net-

works—ultimately allowing for a given system to unearth the internal representations required for feature detection and classification all by itself, given raw data as input [186].

The family of methods that encompass Deep Learning, however, employ several levels of representation called layers, in which each one contributes to the transformation of the raw data and the increasing level of abstraction. The higher (i.e., deeper) the level, the more abstract the information is. This process of incremental abstraction throughout the layers of a neural network not only allows to uncover hidden patterns in the data, but also complex and intricate associations, while mitigating irrelevant fluctuations (i.e., noise), as long as sufficient transformations (i.e., abstractions) have been undergone on the original input data [186].

Classification approaches that make use of Deep Learning have revolutionized many scientific fields [190], drastically improving their prediction accuracy [183]. This revolution is particularly prominent in the realm of biological sciences [190], with numerous applications as broad as regulatory genomics [191], biomedicine [192], bioinformatics [193], computational biology [194], population genetic inference [195], gene expression [196], protein secondary-structure prediction [197], SNPs variant callers [198], and protein-protein interaction networks [199], amidst many others.

Still, and as pointed out by Ching and colleagues [190], it is critical to acknowledge that Deep Learning is a subfield of machine-learning, and as such it bears the same restraints as other approaches in this field [190]. Namely, one should be mindful that the outputs are nonetheless still inherently intertwined with the methodological design, and that the duality of correlation and causality are undeniably present [190].

Moreover, Deep Learning algorithms require huge amounts of data to train their models, which might be a deterring factor hindering its success in some instances of protein function prediction [200].

1.4.3 A summary of current protein function prediction tools

Notwithstanding these caveats, there are plenty of noteworthy computational approaches that rely on Deep Learning to predict the function of a given protein. For instance, the authors of DeepGO [201] developed this Deep Learning tool to unveil specific features inferred exclusively from the amino acid sequence, together with a multiple-species protein-protein interaction network. The results retrieved using the DeepGO approach show the associated GO

terms for each protein sequence in the user's query, and make use of the same hierarchical and relational-structured ontologies—in this case GO classes—as the background information needed to construct the model [201].

More recently, the same group has developed a new version of the foregoing tool, called DeepGOPlus [202]. They did so because of some limitations inherent to the original tool—namely a maximum upper bound for sequence length, inability to process proteins with ambiguous amino acids, and the inability to address the entire extent of GO terms present in the ontology [202]. The authors reported that DeepGOPlus not only overcomes the main limitations of its predecessor, but it also predicts any GO class that was ever used in an experimental annotation [202].

You et al. developed DeepText2GO [203], whose underlying framework is built upon the deep semantic text representation paradigm—i.e., "document to vector" (D2V) representation of text—along with several different types of information extracted from protein sequences such as protein families, domains, and motifs, as well as sequence homology. Both text-based features and sequence-derived ones are then integrated using a consensus approach. In their paper, they validated the tool through exhaustive testing on a large-scale protein dataset, aiming at protein function prediction through the use of text-based features [203]. The authors reported that DeepText2GO significantly outperforms text-based and sequence-based methods being employed separately [203]. They also reported to be the first study on protein function prediction that makes use of deep semantic representation [203].

Alternatively, the team behind ProLanGO [183] addressed the task of imputing the associated GO terms for a given protein in a wholly different way. The rationale behind the ProLanGO approach compares the effort of predicting the function of a protein as that applied to the translation of a language into another [183]. Hence, by using recurrent neural networks, they created a machine translation model that "translates" the amino acid sequence language into that of GO terms [183].

Another example is that of DeepFam [204], in which an alignment-free approach is used to retrieve functional categorization information along with protein family prediction, directly from the amino acid sequence data. The creators of DeepFam have shown that their method exhibits better performance both in accuracy, as well as in runtime, when compared to several alignment-free and alignment-based strategies. Besides that, it discriminates conserved regions,

and models these regions to their corresponding protein families without the need for multiple-alignments [204].

As for enzyme function, three methods should be mentioned. DEEPre [205], an approach that encompasses both end-to-end feature selection and classification model training, allows the user to simply input a given protein sequence while the model extracts its convolutional and sequential features, instead of having the user manually crafting these same features beforehand [205]. The authors also showed that DEEPre significantly improves the prediction performance for a given enzyme when compared with earlier methods [205].

The other method worth mentioning, mlDEEPre [206], is a novel and improved version of the previously described method, explicitly designed to infer the various functions of multi-functional enzymes. The team adopted a novel loss-function approach and associated it with the hierarchical relationship amongst enzyme labels, allowing for multi-functional enzyme prediction [206].

A different group created DeepEC [207], a deep learning-based computational framework that predicts the catalytic potential for a given protein, in addition to its 4th-digit Enzyme Commission (EC) number. This framework operates by using three distinct and independent convolutional neural networks (CNNs): the first predicts whether the protein sequence is an enzyme or not; the second predicts the EC number at the 3rd-digit, and the third predicts the EC number at the 4th-digit. The results provided by this tool are only pertaining to those proteins which have been predicted to be enzymes by the first CNN, and whose 3rd-digit EC is a prefix of the 4th-digit EC, as to mitigate false-positive predictions [207]. However, the tool also performs homology analysis if one of the CNNs fails to predict an EC number. Additionally, it also outputs to the user the log files corresponding to each individual CNN prediction, along with the confidence values for each prediction accordingly. Moreover, a given protein can be assigned more than one enzymatic function by DeepEC, with congruent predictions for all three CNNs.

Protein function prediction is undoubtedly a major challenge of the post-genomic era [208], particularly for the FDM from any domain of life. The progress made thus far has been substantial, and despite the fact that year upon year new protein function prediction tools keep surfacing, with ever-improving performance and accuracy, this biological conundrum is far from being resolved, as recently pinpointed by Makrodimitris and colleagues [208].

In this section we have described a number of computational approaches of reference that

can be used to aid in the unearthing of FDM. We share the opinion that Deep Learning approaches are the path to take, given that they offer unprecedented flexibility on how the raw data can be modeled, without the need to previously develop the system's internal data representation [190].

The flexibility achieved with Deep Learning has prompted the development of creative and truly groundbreaking modeling approaches, which would have been unachievable if using different machine-learning techniques [190]. There is the expectation that one day these computational methods will be able to condense huge collections of input data and transform the data into intelligible information, giving researchers the ability to query these data in new ways, namely those that have been overlooked thus far [190].

1.5 Concluding Remarks

Owing to the efforts and sequencing initiatives of numerous research groups worldwide, today hundreds of thousands of microbial genomes and metagenomes are available in public data repositories. This deluge of data can be readily used for data-mining endeavors and corresponding analytics aiming at the discovery of biological novelties. Secondary analyses of these sequence data can be further refined by new developments in bioinformatic and data mining tools expediting opportunities for data recycling, enabling to characterize and annotate the massive breadth of "known unknowns" that lie overlooked amidst public sequence data repositories.

An optimistic outlook from this endeavor is that by abiding through the recycling of (meta) genomic data, a breakthrough of new molecular functions and metabolites crucial for the biotechnology industry will be made, helping to achieve ahead of time the sustainable development goals for 2030 adopted by the United Nations.

References

[1] Blaser MJ, Cardon ZG, Cho MK, Dangl JL, Donohue TJ, Green JL, et al. Toward a Predictive Understanding of Earth's Microbiomes to Address 21st Century Challenges. MBio. 2016;7(3).

[2] Allwood AC, Walter MR, Kamber BS, Marshall CP, Burch IW. Stromatolite reef from the Early Archaean era of Australia. Nature. 2006;441(7094):714–718.

[3] McFall-Ngai M, Hadfield MG, Bosch TCG, Carey HV, Domazet-Lošo T, Douglas AE, et al. Animals in a bacterial world, a new imperative for the life sciences. Proceedings of the National Academy of Sciences. 2013;110(9):3229–3236.

[4] Locey KJ, Lennon JT. Scaling laws predict global microbial diversity. Proc Natl Acad Sci U S A. 2016;113(21):5970–5975.

[5] Solden L, Lloyd K, Wrighton K. The bright side of microbial dark matter: lessons learned from the uncultivated majority. Curr Opin Microbiol. 2016;31:217–226.

[6] Lagkouvardos I, Overmann J, Clavel T. Cultured microbes represent a substantial fraction of the human and mouse gut microbiota. Gut Microbes. 2017;8(5):493–503.

[7] Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer KH, et al. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat Rev Microbiol. 2014;12(9):635–645.

[8] Pedrós-Alió C, Manrubia S. The vast unknown microbial biosphere. Proc Natl Acad Sci U S A. 2016;113(24):6585–6587.

[9] Willis A. Extrapolating abundance curves has no predictive power for estimating microbial biodiversity. Proc Natl Acad Sci U S A. 2016;113(35):E5096.

[10] Lok C. Mining the microbial dark matter. Nature. 2015;522(7556):270–273.

[11] Małyska A, Markakis MN, Pereira CF, Cornelissen M. The Microbiome: A Life Science Opportunity for Our Society and Our Planet. Trends Biotechnol. 2019;37(12):1269–1272.

[12] Transforming our world: the 2030 Agenda for Sustainable Development. Sustainable Development Knowledge Platform;. https://sustainabledevelopment.un.org/post2015/transformingourworld.

[13] Jannasch HW, Jones GE. Bacterial Populations in Sea Water as Determined by Different Methods of Enumeration1. Limnol Oceanogr. 1959;4(2):128–139.

[14] Jones JG. Studies on Freshwater Bacteria: Effect of Medium Composition and Method on Estimates of Bacterial Population. J Appl Bacteriol. 1970;33(4):679–686.

[15] Rappé MS, Giovannoni SJ. The Uncultured Microbial Majority. Annu Rev Microbiol. 2003;57(1):369–394.

[16] Olsen GJ, Lane DJ, Giovannoni SJ, Pace NR, Stahl DA. Microbial Ecology and Evolution: A Ribosomal RNA Approach. Annu Rev Microbiol. 1986;40(1):337–365.

[17] Woese CR. Bacterial evolution. Microbiol Rev. 1987;51(2):221–271.

[18] Pace NR. Opening the door onto the natural microbial world: molecular microbial ecology. Harvey Lect. 1995;91:59–78.

[19] Karlsson E, Lärkeryd A, Sjödin A, Forsman M, Stenberg P. Scaffolding of a bacterial genome using MinION nanopore sequencing. Sci Rep. 2015;5:11996.

[20] Sims D, Sudbery I, Ilott NE, Heger A, Ponting CP. Sequencing depth and coverage: key considerations in genomic analyses. Nat Rev Genet. 2014;15(2):121–132.

[21] Lagier JC, Dubourg G, Million M, Cadoret F, Bilen M, Fenollar F, et al. Culturing the human microbiota and culturomics. Nat Rev Microbiol. 2018;16:540–550.

[22] Owen JG, Charlop-Powers Z, Smith AG, Ternei MA, Calle PY, Reddy BVB, et al. Multiplexed metagenome mining using short DNA sequence tags facilitates targeted discovery of epoxyketone proteasome inhibitors. Proc Natl Acad Sci U S A. 2015;112(14):4221–4226.

[23] Ziemert N, Alanjary M, Weber T. The evolution of genome mining in microbes – a review. Nat Prod Rep. 2016;33(8):988–1005.

[24] Ling LL, Schneider T, Peoples AJ, Spoering AL, Engels I, Conlon BP, et al. A new antibiotic kills pathogens without detectable resistance. Nature. 2015;517(7535):455–459.

[25] Lackner G, Peters EE, Helfrich EJN, Piel J. Insights into the lifestyle of uncultured bacterial natural product factories associated with marine sponges. Proc Natl Acad Sci U S A. 2017;114(3):E347–E356.

[26] Bull AT, Goodfellow M. Dark, rare and inspirational microbial matter in the extremobiosphere: 16 000 m of bioprospecting campaigns. Microbiology. 2019;165(12):1252–1264.

[27] Rust M, Helfrich EJN, Freeman MF, Nanudorn P, Field CM, Rückert C, et al. A multiproducer microbiome generates chemical diversity in the marine sponge. Proc Natl Acad Sci U S A. 2020;117(17):9508–9518.

[28] Chen R, Wong HL, Burns BP. New Approaches to Detect Biosynthetic Gene Clusters in the Environment. Medicines (Basel). 2019;6(1).

[29] Goodfellow M, Nouioui I, Sanderson R, Xie F, Bull AT. Rare taxa and dark microbial matter: novel bioactive actinobacteria abound in Atacama Desert soils. Antonie Van Leeuwenhoek. 2018;111(8):1315–1332.

[30] Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of the tree of life. Nat Microbiol. 2016;1:16048.

[31] Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. Nat Microbiol. 2017;2(11):1533–1542.

[32] Brown CT, Hug LA, Thomas BC, Sharon I, Castelle CJ, Singh A, et al. Unusual biology across a group comprising more than 15% of domain Bacteria. Nature. 2015;523(7559):208–211.

[33] Nasir A, Kim KM, Caetano-Anollés G. Lokiarchaeota: eukaryote-like missing links from microbial dark matter? Trends Microbiol. 2015;23(8):448–450.

[34] Wiegand S, Jogler M, Boedeker C, Pinto D, Vollmers J, Rivas-Marín E, et al. Cultivation and functional characterization of 79 planctomycetes uncovers their unique biology. Nat Microbiol. 2020;5(1):126–140.

[35] Marcy Y, Ouverney C, Bik EM, Lösekann T, Ivanova N, Martin HG, et al. Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. Proc Natl Acad Sci U S A. 2007;104(29):11889–11894.

[36] Bernard G, Pathmanathan JS, Lannes R, Lopez P, Bapteste E. Microbial dark matter investigations: how microbial studies transform biological knowledge and empirically sketch a logic of scientific discovery. Genome Biol Evol. 2018;.

[37] Krishnamurthy SR, Wang D. Origins and challenges of viral dark matter. Virus Res. 2017;239:136–142.

[38] Michalska K, Steen AD, Chhor G, Endres M, Webber AT, Bird J, et al. New aminopeptidase from "microbial dark matter" archaeon. FASEB J. 2015;29(9):4071–4079.

[39] Miller IJ, Weyna TR, Fong SS, Lim-Fong GE, Kwan JC. Single sample resolution of rare microbial dark matter in a marine invertebrate metagenome. Sci Rep. 2016;6(1).

[40] Sysoev M, Grötzinger SW, Renn D, Eppinger J, Rueping M, Karan R. Bioprospecting of Novel Extremozymes From Prokaryotes-The Advent of Culture-Independent Methods. Front Microbiol. 2021;12:630013.

[41] Grötzinger SW, Karan R, Strillinger E, Bader S, Frank A, Al Rowaihi IS, et al. Identification and Experimental Characterization of an Extremophilic Brine Pool Alcohol Dehydrogenase from Single Amplified Genomes. ACS Chem Biol. 2018;13(1):161–170.

[42] McLean JS, Lombardo MJ, Badger JH, Edlund A, Novotny M, Yee-Greenbaum J, et al. Candidate phylum TM6 genome recovered from a hospital sink biofilm provides genomic insights into this uncultivated phylum. Proc Natl Acad Sci U S A. 2013;110(26):E2390–9.

[43] Rinke C, Schwientek P, Sczyrba A, Ivanova NN, Anderson IJ, Cheng JF, et al. Insights into the phylogeny and coding potential of microbial dark matter. Nature. 2013;499(7459):431–437.

[44] Makarova KS, Wolf YI, Forterre P, Prangishvili D, Krupovic M, Koonin EV. Dark matter in archaeal genomes: a rich source of novel mobile elements, defense systems and secretory complexes. Extremophiles. 2014;18(5):877–893.

[45] Gies EA, Konwar KM, Beatty JT, Hallam SJ. Illuminating microbial dark matter in meromictic Sakinaw Lake. Appl Environ Microbiol. 2014;80(21):6807–6818.

[46] Wegner CE, Liesack W. Unexpected Dominance of Elusive Acidobacteria in Early Industrial Soft Coal Slags. Front Microbiol. 2017;8:1023.

[47] Mehrshad M, Rodriguez-Valera F, Amoozegar MA, López-García P, Ghai R. The enigmatic SAR202 cluster up close: shedding light on a globally distributed dark ocean lineage involved in sulfur cycling. ISME J. 2017;.

[48] Nobu MK, Narihiro T, Rinke C, Kamagata Y, Tringe SG, Woyke T, et al. Microbial dark matter ecogenomics reveals complex synergistic networks in a methanogenic bioreactor. ISME J. 2015;9(8):1710–1722.

[49] Thrash JC, Cameron Thrash J, Seitz KW, Baker BJ, Temperton B, Gillies LE, et al. Metabolic Roles of Uncultivated Bacterioplankton Lineages in the Northern Gulf of Mexico "Dead Zone". MBio. 2017;8(5):e01017–17.

[50] Momper L, Jungbluth SP, Lee MD, Amend JP. Energy and carbon metabolisms in a deep terrestrial subsurface fluid microbial community. ISME J. 2017;11(10):2319–2333.

[51] Hawley AK, Nobu MK, Wright JJ, Durno WE, Morgan-Lang C, Sage B, et al. Diverse Marinimicrobia bacteria may mediate coupled biogeochemical cycles along eco-thermodynamic gradients. Nat Commun. 2017;8(1):1507.

[52] Anantharaman K, Brown CT, Hug LA, Sharon I, Castelle CJ, Probst AJ, et al. Thousands of microbial genomes shed light on interconnected biogeochemical processes in an aquifer system. Nat Commun. 2016;7:13219.

[53] Hutchison CA 3rd, Chuang RY, Noskov VN, Assad-Garcia N, Deerinck TJ, Ellisman MH, et al. Design and synthesis of a minimal bacterial genome. Science. 2016;351(6280):aad6253.

[54] Piao H, Froula J, Du C, Kim TW, Hawley ER, Bauer S, et al. Identification of novel biomass-degrading enzymes from genomic dark matter: Populating genomic sequence space with functional annotation. Biotechnol Bioeng. 2014;111(8):1550–1565.

[55] Hadjithomas M, Chen IMA, Chu K, Ratner A, Palaniappan K, Szeto E, et al. IMG-ABC: A Knowledge Base To Fuel Discovery of Biosynthetic Gene Clusters and Novel Secondary Metabolites. MBio. 2015;6(4):e00932.

[56] Danso D, Chow J, Streit WR. Plastics: Environmental and Biotechnological Perspectives on Microbial Degradation. Appl Environ Microbiol. 2019;85(19).

[57] Akal AL, Karan R, Hohl A, Alam I, Vogler M, Grötzinger SW, et al. A polyextremophilic alcohol dehydrogenase from the Atlantis II Deep Red Sea brine pool. FEBS Open Bio. 2019;9(2):194–205.

[58] Verma S, Meghwanshi GK, Kumar R. Current perspectives for microbial lipases from extremophiles and metagenomics. Biochimie. 2021;182:23–36.

[59] Rashid M, Stingl U. Contemporary molecular tools in microbial ecology and their application to advancing biotechnology. Biotechnol Adv. 2015;33(8):1755–1773.

[60] Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics. 2011;12:385.

[61] Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big Data: Astronomical or Genomical? PLoS Biol. 2015;13(7):e1002195.

[62] Erdin S, Lisewski AM, Lichtarge O. Protein function prediction: towards integration of similarity metrics. Curr Opin Struct Biol. 2011;21(2):180–188.

[63] Makarova KS, Wolf YI, Koonin EV. Towards functional characterization of archaeal genomic dark matter. Biochem Soc Trans. 2019;47(1):389–398.

[64] Al-Shahib A, Breitling R, Gilbert DR. Predicting protein function by machine learning on amino acid sequences – a critical evaluation. BMC Genomics. 2007;8(1):78.

[65] Garza DR, Dutilh BE. From cultured to uncultured genome sequences: metagenomics and modeling microbial ecosystems. Cell Mol Life Sci. 2015;72(22):4287–4308.

[66] Becraft ED, Dodsworth JA, Murugapiran SK, Ohlsson JI, Briggs BR, Kanbar J, et al. Single-Cell-Genomics-Facilitated Read Binning of Candidate Phylum EM19 Genomes from Geothermal Spring Metagenomes. Appl Environ Microbiol. 2015;82(4):992–1003.

[67] Lobb B, Kurtz DA, Moreno-Hagelsieb G, Doxey AC. Remote homology and the functions of metagenomic dark matter. Front Genet. 2015;6:234.

[68] Mokili JL, Rohwer F, Dutilh BE. Metagenomics and future perspectives in virus discovery. Curr Opin Virol. 2012;2(1):63–77.

[69] Dutilh BE. Metagenomic ventures into outer sequence space. Bacteriophage. 2014;4(4):e979664.

[70] Siew N, Fischer D. Analysis of singleton ORFans in fully sequenced microbial genomes. Proteins. 2003;53(2):241–251.

[71] van Dijk EL, Auger H, Jaszczyszyn Y, Thermes C. Ten years of next-generation sequencing technology. Trends Genet. 2014;30(9):418–426.

[72] Andersson JO, Andersson SG. Pseudogenes, junk DNA, and the dynamics of Rickettsia genomes. Mol Biol Evol. 2001;18(5):829–839.

[73] Mira A. Microbial genome evolution: sources of variability. Curr Opin Microbiol. 2002;5(5):506–512.

[74] Schmid KJ, Aquadro CF. The evolutionary analysis of "orphans" from the Drosophila genome identifies rapidly diverging and incorrectly annotated genes. Genetics. 2001;159(2):589–598.

[75] Hu P, Janga SC, Babu M, Díaz-Mejía JJ, Butland G, Yang W, et al. Global functional atlas of Escherichia coli encompassing previously uncharacterized proteins. PLoS Biol. 2009;7(4):e96.

[76] Hanson AD, Henry CS, Fiehn O, de Crécy-Lagard V. Metabolite Damage and Metabolite Damage Control in Plants. Annu Rev Plant Biol. 2016;67(1):131–152.

[77] Ellens KW, Christian N, Singh C, Satagopam VP, May P, Linster CL. Confronting the catalytic dark matter encoded by sequenced genomes. Nucleic Acids Res. 2017;45(20):11495–11514.

[78] Choudhary C, Weinert BT, Nishida Y, Verdin E, Mann M. The growing landscape of lysine acetylation links metabolism and cell signalling. Nat Rev Mol Cell Biol. 2014;15(8):536–550.

[79] Van Schaftingen E, Veiga-da Cunha M, Linster CL. Enzyme complexity in intermediary metabolism. J Inherit Metab Dis. 2015;38(4):721–727.

[80] Jeffery CJ. Protein moonlighting: what is it, and why is it important? Philos Trans R Soc Lond B Biol Sci. 2018;373(1738).

[81] Mani M, Chen C, Amblee V, Liu H, Mathur T, Zwicke G, et al. MoonProt: a database for proteins that are known to moonlight. Nucleic Acids Res. 2015;43(Database issue):D277–82.

[82] Wilson GA, Bertrand N, Patel Y, Hughes JB, Feil EJ, Field D. Orphans as taxonomically restricted and ecologically important genes. Microbiology. 2005;151(Pt 8):2499–2501.

[83] Kaessmann H. Origins, evolution, and phenotypic impact of new genes. Genome Res. 2010;20(10):1313–1326.

[84] Tautz D, Domazet-Lošo T. The evolutionary origin of orphan genes. Nat Rev Genet. 2011;12(10):692–702.

[85] Yin Y, Fischer D. On the origin of microbial ORFans: quantifying the strength of the evidence for viral lateral transfer. BMC Evol Biol. 2006;6:63.

[86] Cortez D, Forterre P, Gribaldo S. A hidden reservoir of integrative elements is the major source of recently acquired foreign genes and ORFans in archaeal and bacterial genomes. Genome Biol. 2009;10(6):R65.

[87] Dutilh BE, Cassman N, McNair K, Sanchez SE, Silva GGZ, Boling L, et al. A highly abundant bacteriophage discovered in the unknown sequences of human faecal metagenomes. Nat Commun. 2014;5:4498.

[88] Sberro H, Fremin BJ, Zlitni S, Edfors F, Greenfield N, Snyder MP, et al. Large-Scale Analyses of Human Microbiomes Reveal Thousands of Small, Novel Genes. Cell. 2019;178(5):1245–1259.e14.

[89] Pascoal F, Magalhães C, Costa R. The Link Between the Ecology of the Prokaryotic Rare Biosphere and Its Biotechnological Potential. Front Microbiol. 2020;11:231.

[90] Bioremediation 3.0: Engineering pollutant-removing bacteria in the times of systemic biology. Biotechnol Adv. 2017;35(7):845–866.

[91] Tournier V, Topham CM, Gilles A, David B, Folgoas C, Moya-Leclair E, et al. An engineered PET depolymerase to break down and recycle plastic bottles. Nature. 2020;580(7802):216–219.

[92] Ragaert K, Delva L, Van Geem K. Mechanical and chemical recycling of solid plastic waste. Waste Manag. 2017;69:24–58.

[93] Yoshida S, Hiraga K, Takehana T, Taniguchi I, Yamaji H, Maeda Y, et al. A bacterium that degrades and assimilates poly(ethylene terephthalate). Science. 2016;351(6278):1196–1199.

[94] Franden MA, Jayakody LN, Li WJ, Wagner NJ, Cleveland NS, Michener WE, et al. Engineering Pseudomonas putida KT2440 for efficient ethylene glycol utilization. Metab Eng. 2018;48:197–207.

[95] Austin HP, Allen MD, Donohoe BS, Rorrer NA, Kearns FL, Silveira RL, et al. Characterization and engineering of a plastic-degrading aromatic polyesterase. Proc Natl Acad Sci U S A. 2018;115(19):E4350–E4357.

[96] Palm GJ, Reisky L, Böttcher D, Müller H, Michels EAP, Walczak MC, et al. Structure of the plastic-degrading Ideonella sakaiensis MHETase bound to a substrate. Nat Commun. 2019;10(1):1717.

[97] Nichols D, Cahoon N, Trakhtenberg EM, Pham L, Mehta A, Belanger A, et al. Use of ichip for high-throughput in situ cultivation of "uncultivable" microbial species. Appl Environ Microbiol. 2010;76(8):2445–2450.

[98] Berdy B, Spoering AL, Ling LL, Epstein SS. In situ cultivation of previously uncultivable microorganisms using the ichip. Nat Protoc. 2017;12(10):2232–2242.

[99] Li S, Yang X, Yang S, Zhu M, Wang X. Technology prospecting on enzymes: application, marketing and engineering. Comput Struct Biotechnol J. 2012;2:e201209017.

[100] Bruno S, Coppola D, di Prisco G, Giordano D, Verde C. Enzymes from Marine Polar Regions and Their Biotechnological Applications. Mar Drugs. 2019;17(10).

[101] Singh R, Kumar M, Mittal A, Mehta PK. Microbial enzymes: industrial progress in 21st century. 3 Biotech. 2016;6(2):174.

[102] Gurung N, Ray S, Bose S, Rai V. A broader view: microbial enzymes and their relevance in industries, medicine, and beyond. Biomed Res Int. 2013;2013:329121.

[103] Robinson PK. Enzymes: principles and biotechnological applications. Essays Biochem. 2015;59:1–41.

[104] Cabrera MÁ, Blamey JM. Biotechnological applications of archaeal enzymes from extreme environments. Biol Res. 2018;51(1):37.

[105] Ramesh A, Harani Devi P, Chattopadhyay S, Kavitha M. Commercial Applications of Microbial Enzymes. In: Microorganisms for Sustainability. Singapore: Springer Singapore; 2020. p. 137–184.

[106] Meghwanshi GK, Kaur N, Verma S, Dabi NK, Vashishtha A, Charan PD, et al. Enzymes for pharmaceutical and therapeutic applications. Biotechnol Appl Biochem. 2020;67(4):586–601.

[107] Singh BK. Organophosphorus-degrading bacteria: ecology and industrial applications. Nat Rev Microbiol. 2009;7(2):156–164.

[108] Smith HO, Wilcox KW. A restriction enzyme from Hemophilus influenzae. I. Purification and general properties. J Mol Biol. 1970;51(2):379–391.

[109] Brock TD, Freeze H. Thermus aquaticus gen. n. and sp. n., a nonsporulating extreme thermophile. J Bacteriol. 1969;98(1):289–297.

[110] Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, et al. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science. 1988;239(4839):487–491.

[111] Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science. 2012;337(6096):816–821.

[112] Newman DJ, Cragg GM. Natural Products as Sources of New Drugs from 1981 to 2014. J Nat Prod. 2016;79(3):629–661.

[113] Chavali AK, Rhee SY. Bioinformatics tools for the identification of gene clusters that biosynthesize specialized metabolites. Brief Bioinform. 2018;19(5):1022–1034.

[114] Medema MH, Kottmann R, Yilmaz P, Cummings M, Biggins JB, Blin K, et al. Minimum Information about a Biosynthetic Gene cluster. Nat Chem Biol. 2015;11(9):625–631.

[115] Weber T, Kim HU. The secondary metabolite bioinformatics portal: Computational tools to facilitate synthetic biology of secondary metabolite production. Synth Syst Biotechnol. 2016;1(2):69–79.

[116] Arnison PG, Bibb MJ, Bierbaum G, Bowers AA, Bugni TS, Bulaj G, et al. Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. Nat Prod Rep. 2013;30(1):108–160.

[117] Hetrick KJ, van der Donk WA. Ribosomally synthesized and post-translationally modified peptide natural product discovery in the genomic era. Curr Opin Chem Biol. 2017;38:36–44.

[118] Zhong Z, He B, Li J, Li YX. Challenges and advances in genome mining of ribosomally synthesized and post-translationally modified peptides (RiPPs). Synth Syst Biotechnol. 2020;5(3):155–172.

[119] Blin K, Kim HU, Medema MH, Weber T. Recent development of antiSMASH and other computational approaches to mine secondary metabolite biosynthetic gene clusters. Brief Bioinform. 2019;20(4):1103–1113.

[120] Ren H, Wang B, Zhao H. Breaking the silence: new strategies for discovering novel natural products. Curr Opin Biotechnol. 2017;48:21–27.

[121] Skinnider MA, Johnston CW, Edgar RE, Dejong CA, Merwin NJ, Rees PN, et al. Genomic charting of ribosomally synthesized natural product chemical space facilitates targeted mining. Proc Natl Acad Sci U S A. 2016;113(42):E6343–E6351.

[122] Bérdy J. Thoughts and facts about antibiotics: where we are now and where we are heading. J Antibiot. 2012;65(8):385–395.

[123] Cornelissen M, Małyska A, Nanda AK, Lankhorst RK, Parry MAJ, Saltenis VR, et al. Biotechnology for Tomorrow's World: Scenarios to Guide Directions for Future Innovation. Trends Biotechnol. 2021;39(5):438–444.

[124] Luo F, Devine CE, Edwards EA. Cultivating microbial dark matter in benzene-degrading methanogenic consortia. Environ Microbiol. 2016;18(9):2923–2936.

[125] Lannes R, Cavaud L, Lopez P, Bapteste E. Marine Ultrasmall Prokaryotes Likely Affect the Cycling of Carbon, Methane, Nitrogen, and Sulfur. Genome Biol Evol. 2021;13(1).

[126] Wong HL, MacLeod FI, White RA 3rd, Visscher PT, Burns BP. Microbial dark matter filling the niche in hypersaline microbial mats. Microbiome. 2020;8(1):135.

[127] Mason OU, Hazen TC, Borglin S, Chain PSG, Dubinsky EA, Fortney JL, et al. Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to Deepwater Horizon oil spill. ISME J. 2012;6(9):1715–1727.

[128] Idris H, Goodfellow M, Sanderson R, Asenjo JA, Bull AT. Actinobacterial Rare Biospheres and Dark Matter Revealed in Habitats of the Chilean Atacama Desert. Sci Rep. 2017;7(1):8373.

[129] Bull AT, Idris H, Sanderson R, Asenjo J, Andrews B, Goodfellow M. High altitude, hyper-arid soils of the Central-Andes harbor mega-diverse communities of actinobacteria. Extremophiles. 2018;22(1):47–57.

[130] Hedlund BP, Dodsworth JA, Staley JT. The changing landscape of microbial biodiversity exploration and its implications for systematics. Syst Appl Microbiol. 2015;38(4):231–236.

[131] Hedlund BP, Dodsworth JA, Murugapiran SK, Rinke C, Woyke T. Impact of single-cell genomics and metagenomics on the emerging view of extremophile "microbial dark matter". Extremophiles. 2014;18(5):865–875.

[132] Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, et al. The SEED and the Rapid Annotation of microbial genomes using Subsystems Technology (RAST). Nucleic Acids Res. 2014;42(Database issue):D206–14.

[133] Tatusova T, DiCuccio M, Badretdin A, Chetvernin V, Nawrocki EP, Zaslavsky L, et al. NCBI prokaryotic genome annotation pipeline. Nucleic Acids Res. 2016;44(14):6614–6624.

[134] Teeling H, Glöckner FO. Current opportunities and challenges in microbial metagenome analysis–a bioinformatic perspective. Brief Bioinform. 2012;13(6):728–742.

[135] Simon C, Daniel R. Metagenomic analyses: past and future trends. Appl Environ Microbiol. 2011;77(4):1153–1161.

[136] Scholz MB, Lo CC, Chain PSG. Next generation sequencing and bioinformatic bottlenecks: the current state of metagenomic data analysis. Curr Opin Biotechnol. 2012;23(1):9–15.

[137] Baker BJ, Comolli LR, Dick GJ, Hauser LJ, Hyatt D, Dill BD, et al. Enigmatic, ultrasmall, uncultivated Archaea. Proc Natl Acad Sci U S A. 2010;107(19):8806–8811.

[138] Narasingarao P, Podell S, Ugalde JA, Brochier-Armanet C, Emerson JB, Brocks JJ, et al. De novo metagenomic assembly reveals abundant novel major lineage of Archaea in hypersaline microbial communities. ISME J. 2012;6(1):81–93.

[139] Nunoura T, Takaki Y, Kakuta J, Nishi S, Sugahara J, Kazama H, et al. Insights into the evolution of Archaea and eukaryotic protein modifier systems revealed by the genome of a novel archaeal group. Nucleic Acids Res. 2011;39(8):3204–3223.

[140] Takami H, Noguchi H, Takaki Y, Uchiyama I, Toyoda A, Nishi S, et al. A deeply branching thermophilic bacterium with an ancient acetyl-CoA pathway dominates a subsurface ecosystem. PLoS One. 2012;7(1):e30559.

[141] Doud DFR, Woyke T. Novel approaches in function-driven single-cell genomics. FEMS Microbiol Rev. 2017;41(4):538–548.

[142] Woyke T, Jarett J. Function-driven single-cell genomics. Microb Biotechnol. 2015;8(1):38–39.

[143] Ackermann M. Microbial individuality in the natural environment. ISME J. 2013;7(3):465–467.

[144] Engel P, Stepanauskas R, Moran NA. Hidden diversity in honey bee gut symbionts detected by single-cell genomics. PLoS Genet. 2014;10(9):e1004596.

[145] D Ainsworth T, Krause L, Bridge T, Torda G, Raina JB, Zakrzewski M, et al. The coral core microbiome identifies rare bacterial taxa as ubiquitous endosymbionts. ISME J. 2015;9(10):2261–2274.

[146] Gawad C, Koh W, Quake SR. Single-cell genome sequencing: current state of the science. Nat Rev Genet. 2016;17(3):175–188.

[147] Woyke T, Tighe D, Mavromatis K, Clum A, Copeland A, Schackwitz W, et al. One Bacterial Cell, One Complete Genome. PLoS One. 2010;5(4):e10314.

[148] Martijn J, Schulz F, Zaremba-Niedzwiedzka K, Viklund J, Stepanauskas R, Andersson SGE, et al. Single-cell genomics of a rare environmental alphaproteobacterium provides unique insights into Rickettsiaceae evolution. ISME J. 2015;9(11):2373–2385.

[149] Martinez-Garcia M, Brazel DM, Swan BK, Arnosti C, Chain PSG, Reitenga KG, et al. Capturing single cell genomes of active polysaccharide degraders: an unexpected contribution of Verrucomicrobia. PLoS One. 2012;7(4):e35314.

[150] Kakirde KS, Parsley LC, Liles MR. Size Does Matter: Application-driven Approaches for Soil Metagenomics. Soil Biol Biochem. 2010;42(11):1911–1923.

[151] Gabor EM, de Vries EJ, Janssen DB. Efficient recovery of environmental DNA for expression cloning by indirect extraction methods. FEMS Microbiol Ecol. 2003;44(2):153–163.

[152] Brown BL, Watson M, Minot SS, Rivera MC, Franklin RB. MinION^{TM} nanopore sequencing of environmental metagenomes: a synthetic approach. Gigascience. 2017;6(3):1–10.

[153] Eisenstein M. Oxford Nanopore announcement sets sequencing sector abuzz. Nat Biotechnol. 2012;30(4):295–296.

[154] Edwards A, Debbonaire AR, Sattler B, Mur LAJ, Hodson AJ. Extreme metagenomics using nanopore DNA sequencing: a field report from Svalbard, 78 N; 2016.

[155] Bellali S, Lagier JC, Million M, Anani H, Haddad G, Francis R, et al. Running after ghosts: are dead bacteria the dark matter of the human gut microbiota? Gut Microbes. 2021;13(1):1–12.

[156] Pedrós-Alió C. The Rare Bacterial Biosphere; 2012.

[157] Lloyd KG, Schreiber L, Petersen DG, Kjeldsen KU, Lever MA, Steen AD, et al. Predominant archaea in marine sediments degrade detrital proteins. Nature. 2013;496(7444):215–218.

[158] Temme K, Zhao D, Voigt CA. Refactoring the nitrogen fixation gene cluster from Klebsiella oxytoca. Proc Natl Acad Sci U S A. 2012;109(18):7085–7090.

[159] Gabor EM, Alkema WBL, Janssen DB. Quantifying the accessibility of the metagenome by random expression cloning techniques. Environ Microbiol. 2004;6(9):879–886.

[160] Ling J, O'Donoghue P, Söll D. Genetic code flexibility in microorganisms: novel mechanisms and impact on physiology. Nat Rev Microbiol. 2015;13(11):707–721.

[161] Leinonen R, Sugawara H, Shumway M, on behalf of the International Nucleotide Sequence Database Collaboration. The Sequence Read Archive. Nucleic Acids Res. 2010;39(Database):D19–D21.

[162] Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, et al. The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC Bioinformatics. 2008;9:386.

[163] Famiglietti ML, Estreicher A, Gos A, Bolleman J, Géhant S, Breuza L, et al. Genetic variations and diseases in UniProtKB/Swiss-Prot: the ins and outs of expert manual curation. Hum Mutat. 2014;35(8):927–935.

[164] NCBI Resource Coordinators. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res. 2018;46(D1):D8–D13.

[165] Boeckmann B. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 2003;31(1):365–370.

[166] O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44(D1):D733–45.

[167] Index of all protein sequence records from NCBI;. https://www.ncbi.nlm.nih.gov/protein/?term=all%5Bfilter%5D.

[168] Zallot R, Harrison K, Kolaczkowski B, de Crécy-Lagard V. Functional Annotations of Paralogs: A Blessing and a Curse. Life. 2016;6(4):39.

[169] Lee D, Redfern O, Orengo C. Predicting protein function from sequence and structure. Nat Rev Mol Cell Biol. 2007;8(12):995–1005.

[170] Koonin EV, Galperin MY. Sequence - Evolution - Function: Computational Approaches in Comparative Genomics. Boston: Kluwer Academic; 2010.

[171] Poux S, Magrane M, Arighi CN, Bridge A, O'Donovan C, Laiho K, et al. Expert curation in UniProtKB: a case study on dealing with conflicting and erroneous data. Database. 2014;2014:bau016.

[172] Schnoes AM, Brown SD, Dodevski I, Babbitt PC. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol. 2009;5(12):e1000605.

[173] Meyer C, Scalzitti N, Jeannin-Girardon A, Collet P, Poch O, Thompson JD. Understanding the causes of errors in eukaryotic protein-coding gene prediction: a case study of primate proteomes. BMC Bioinformatics. 2020;21(1):513.

[174] Coutinho TJD, Franco GR, Lobo FP. Homology-independent metrics for comparative genomics. Comput Struct Biotechnol J. 2015;13:352–357.

[175] Barona-Gómez F. Re-annotation of the sequence $>$ annotation: opportunities for the functional microbiologist. Microb Biotechnol. 2015;8(1):2–4.

[176] Burian RM. Exploratory Experimentation. In: Encyclopedia of Systems Biology; 2013. p. 720–723.

[177] Date SV. The Rosetta Stone Method. In: Methods in Molecular Biology^{TM}; 2008. p. 169–180.

[178] Koehorst JJ, Saccenti E, Schaap PJ, Martins Dos Santos VAP, Suarez-Diez M. Protein domain architectures provide a fast, efficient and scalable alternative to sequence-based methods for comparative functional genomics. F1000Res. 2016;5:1987.

[179] Wu Q, Ye Y, Ng MK, Ho SS, Shi R. Collective prediction of protein functions from protein-protein interaction networks. BMC Bioinformatics. 2014;15 Suppl 2:S9.

[180] Hermann JC, Marti-Arbona R, Fedorov AA, Fedorov E, Almo SC, Shoichet BK, et al. Structure-based activity prediction for an enzyme of unknown function. Nature. 2007;448(7155):775–779.

[181] Lan L, Djuric N, Guo Y, Vucetic S. MS-kNN: protein function prediction by integrating multiple data sources. BMC Bioinformatics. 2013;14 Suppl 3:S8.

[182] Enault F, Suhre K, Claverie JM. Phydbac "Gene Function Predictor": a gene annotation tool based on genomic context analysis. BMC Bioinformatics. 2005;6:247.

[183] Cao R, Freitas C, Chan L, Sun M, Jiang H, Chen Z. ProLanGO: Protein Function Prediction Using Neural Machine Translation Based on a Recurrent Neural Network. Molecules. 2017;22(10).

[184] Libbrecht MW, Noble WS. Machine learning applications in genetics and genomics. Nat Rev Genet. 2015;16(6):321–332.

[185] Bernardes JS, Pedreira CE. A review of protein function prediction under machine learning perspective. Recent Pat Biotechnol. 2013;7(2):122–141.

[186] LeCun Y, Bengio Y, Hinton G. Deep learning. Nature. 2015;521(7553):436–444.

[187] Cao R, Cheng J. Integrated protein function prediction by mining function associations, sequences, and protein-protein and gene-gene interaction networks. Methods. 2016;93:84–91.

[188] Halperin I, Glazer DS, Wu S, Altman RB. The FEATURE framework for protein function annotation: modeling new functions, improving performance, and extending to novel applications. BMC Genomics. 2008;9 Suppl 2:S2.

[189] Cai CZ, Han LY, Ji ZL, Chen X, Chen YZ. SVM-Prot: Web-based support vector machine software for functional classification of a protein from its primary sequence. Nucleic Acids Res. 2003;31(13):3692–3697.

[190] Ching T, Himmelstein DS, Beaulieu-Jones BK, Kalinin AA, Do BT, Way GP, et al. Opportunities and obstacles for deep learning in biology and medicine. J R Soc Interface. 2018;15(141).

[191] Park Y, Kellis M. Deep learning for regulatory genomics. Nat Biotechnol. 2015;33(8):825–826.

[192] Mamoshina P, Vieira A, Putin E, Zhavoronkov A. Applications of Deep Learning in Biomedicine. Mol Pharm. 2016;13(5):1445–1454.

[193] Min S, Lee B, Yoon S. Deep learning in bioinformatics. Brief Bioinform. 2016; p. bbw068.

[194] Angermueller C, Pärnamaa T, Parts L, Stegle O. Deep learning for computational biology. Mol Syst Biol. 2016;12(7):878.

[195] Sheehan S, Song YS. Deep Learning for Population Genetic Inference. PLoS Comput Biol. 2016;12(3):e1004845.

[196] Singh R, Lanchantin J, Robins G, Qi Y. DeepChrome: deep-learning for predicting gene expression from histone modifications. Bioinformatics. 2016;32(17):i639–i648.

[197] Wang S, Peng J, Ma J, Xu J. Protein Secondary Structure Prediction Using Deep Convolutional Neural Fields. Sci Rep. 2016;6:18962.

[198] Poplin R, Chang PC, Alexander D, Schwartz S, Colthurst T, Ku A, et al. A universal SNP and small-indel variant caller using deep neural networks. Nat Biotechnol. 2018;36(10):983–987.

[199] Du X, Sun S, Hu C, Yao Y, Yan Y, Zhang Y. DeepPPI: Boosting Prediction of Protein-Protein Interactions with Deep Neural Networks. J Chem Inf Model. 2017;57(6):1499–1510.

[200] Bonetta R, Valentino G. Machine learning techniques for protein function prediction. Proteins. 2020;88(3):397–413.

[201] Kulmanov M, Khan MA, Hoehndorf R, Wren J. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. Bioinformatics. 2018;34(4):660–668.

[202] Kulmanov M, Hoehndorf R. DeepGOPlus: improved protein function prediction from sequence. Bioinformatics. 2020;36(2):422–429.

[203] You R, Huang X, Zhu S. DeepText2GO: Improving large-scale protein function prediction with deep semantic text representation. Methods. 2018;145:82–90.

[204] Seo S, Oh M, Park Y, Kim S. DeepFam: deep learning based alignment-free method for protein family modeling and prediction. Bioinformatics. 2018;34(13):i254–i262.

[205] Li Y, Wang S, Umarov R, Xie B, Fan M, Li L, et al. DEEPre: sequence-based enzyme EC number prediction by deep learning. Bioinformatics. 2018;34(5):760–769.

[206] Zou Z, Tian S, Gao X, Li Y. mlDEEPre: Multi-Functional Enzyme Function Prediction With Hierarchical Multi-Label Deep Learning. Front Genet. 2018;9:714.

[207] Ryu JY, Kim HU, Lee SY. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. Proc Natl Acad Sci U S A. 2019;116(28):13996–14001.

[208] Makrodimitris S, van Ham RCHJ, Reinders MJT. Automatic Gene Function Prediction in the 2020's. Genes. 2020;11(11).

Chapter 2

Characterization of Functional Dark Matter from public databases

This chapter partially transcribes the contents of the following manuscript:

> **Escudeiro P, Henry CS & Dias R** (2022). *Quantification and description of prokaryotic Functional Dark Matter from public databases* (in preparation).

Abstract

The untapped portion of microbial diversity and its undisclosed coding potential is known as the "microbial dark matter" (MDM). Uncharacterized prokaryotic genes were assumed at first to be "junk" genomic elements, pseudogenes, or simply misannotations. Nonetheless, increasing evidence indicates that these elements encode for specific functions.

These genetic assortments have been shown to be an asset for identifying novel enzymes or new physiologic mechanisms in a broader sense. This realization could prompt an increase in biotechnological solutions, in a world with ever-increasing societal demands.

Science has profited from numerous breakthroughs regarding enzymes of microbial origin, as substantiated by copious Nobel Prize laureates. Novel discoveries concerning enzymes of biotechnological interest are also of utmost importance for the development of the industrial enzymes business. This emphasizes the pivotal role of microbial gene discovery.

In this work we gathered and reconciled Bacterial and Archaeal protein sequence information from 5 major public databases. We then applied a clustering routine through different sequence identity thresholds, grouping these protein sequences into hierarchically assorted putative families. We then proceeded to infer the physico-chemical features; Gene Ontologies; Enzyme Commission numbers; as well as the Conserved Domains, for each of these families. We managed to aggregate these families into distinct groupings, according to each of the metadata types inferred early on. This has allowed us to create a "portrait" for the possible protein functions enclosed in these families, as well as the establishment of a new global resource for prokaryotic functional dark matter.

2.1 Introduction

Over the past decade, the growth rate of genomic data acquisition has been staggering. Estimates indicate that the total amount of sequencing data being produced doubles approximately every seven months [1]. This feat is mainly owed to the breakthrough of next-generation sequencing technologies.

In spite of the massive data harnessed by current sequencing technologies, the actual function of more than 35% of genomic elements enclosed in a given microbial genome still remain unknown [2]. Known and extensively studied microorganisms possess a substantial fraction of genes of unknown function encoded in their genomes [3, 4]. Which nonetheless have been deemed to code for key physiological properties essential to life [4]. The fraction of unfunctionalized genes can range up to 50% in some cases, such as newly sequenced genomes [5] and those of uncultured microbial candidate taxa [6–10]. In more extreme scenarios, as indicated by metagenomic studies, these percentages can span from 85% [11] up to 99% of total gene content [12, 13].

Most public sequence data repositories and databases possess an incredible amount of unannotated sequence data ready to be prospected and analysed. A plethora of gene products as imputed from these sequences are often classified as “uncharacterized”, “putative”, “predicted”, “hypothetical”, or just simply “unknown” [14]. Fact that led this untapped portion of microbial functional diversity to be frequently referred to as “genomic dark matter”, or “functional dark matter”.

The rate at which these proteins of unknown function are accumulating is nothing short of alarming. Experimental evidence cannot simply be performed for every single one of these proteins. The amount of resources to engage in such an endeavor would be unimaginable, and would never keep abreast with the pace at which these elements are being discovered.

For these reasons, manual curation cannot be taken into consideration as a viable option, least of all to characterize each protein individually in an experimental setting. This circumstance leaves no choice but to employ unreviewed annotation methods, customarily via computational approaches, as the sole alternative as to predict the molecular function these proteins might portray. Moreover, many of these proteins fail to be annotated via conventional homology-based methods, often lacking even remote recognizable homologs in public databases [11]. Thusly questioning the extent to which they encode for actual gene products, as well as

the extent to which they might be annotated and their functions assigned, whilst still remaining absent of meaningful biological interpretation [11].

These genetic assortments of unknown function can prove themselves to be an outstanding asset for identifying novel metabolite-producing proteins, enzymes, or new physiologic mechanisms in a broader sense [10, 15]. This could in due course prompt an increase in innovative biotechnological solutions in a world of ever-increasing societal demands. It is of paramount importance to bring this uncharted knowledge to light, as to disclose novel biotic solutions and answers to environmental and climate changes; overcoming societal challenges as in the health, agronomical and food industry; and devising cutting-edge tools for the advancement of fields such as metabolic engineering and synthetic biology.

We reckon that even though there is a dire need to unravel the "unknown unknowns" subsumed within the "black box" of microbial communities, there is perhaps a greater imperative that should be addressed: to effectively characterize and annotate the massive breadth of "known unknowns" that lie overlooked amidst public sequence data repositories.

To this end, we have gathered all proteins of unknown function for Archaea and Bacteria from 5 major public sequence data repositories, and aimed at reducing their redundancy via a well-established protein sequence clustering algorithm. The resulting cluster representatives were regarded as orthologous protein families, and all subsequent work was performed on these same families.

These protein families were ultimately characterized by consolidating the metadata as generated by several types of function prediction sources, namely their Conserved Domains (CD), predicted Gene Ontology terms (GO terms) and Enzyme Commission numbers (EC), as well as physico-chemical metadata. We also refer to the most over-represented putative functions in comparison to a golden standard dataset

Our work allowed us to aggregate these protein families in distinct groups, either according to their protein domain composition; their physico-chemical properties; their GO term predictions; or their putative EC class, respectively. This has allowed us to create a "portrait" for the possible protein functions as enclosed in the prokaryotic "functional dark matter".

All the source data and metadata produced in this study will be publicly available to the scientific community, in order to promote the characterization of the "known unknowns" that populate sequence data repositories to this day.

2.2 Methods

2.2.1 Datasources

We downloaded all non-redundant GenPept Flat-Files from both "bacteria" and "archaea" sub-directories inside the Reference sequence (RefSeq) database [16] FTP server directory (ftp://ftp.ncbi.nih.gov/refseq/release/), on October the 23rd 2018 (RefSeq version 91). We also downloaded all "BCT" division files from GenBank's [17] FTP server directory (ftp://ftp.ncbi.nih.gov/genbank/) on October the 20th 2018 (GenBank version 228). We listed all public Archaeal and Bacterial genomes from the Pathosystems Resource Integration Center (PATRIC) [18], by accessing its FTP server (ftp://ftp.patricbrc.org/genomes). From a total of 200,795 genomes, only 200,779 (99,99%) had the required FASTA and GFF3 files. These 200,779 genomes were downloaded on October the 27th 2018.

We queried an API at the Joint Genome Institute (JGI) Genome Portal homepage (https://genome.jgi.doe.gov/portal/ext-api/), for all the genomes of Archaeal or Bacterial origin. This query retrieved the identifiers of 39,066 JGI genomes for Bacteria, and 1,066 for Archaea. From these numbers only 39,033 (99.99%) and 1,063 (99.71%) were from the Integrated Microbial Genomes (IMG) [19], respectively. Next, we requested JGI's data-retrieval API for a download link for each of the 40,096 genome identifiers belonging to IMG, and downloaded them on October the 23rd 2018. From the combined number of Bacterial and Archaeal genomes, 5,991 (14,94%) either did not have an operational download link, or did not have both required FASTA and GFF3 files. These 5,991 genomes were discarded from further downstream processing.

We also retrieved protein sequence data from the Translated EMBL Nucleotide Sequence Data Library (TrEMBL) [20] by downloading both "uniprot_trembl_archaea.dat.gz", and "uniprot_trembl_bacteria.dat.gz" files from UniProt's FTP server (ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/taxonomic_divisions/), on December the 4th 2018. We proceeded the same way in order to retrieve the complete SwissProt database [20] ("uniprot_sprot.dat.gz"), from UniProt's FTP server (ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/), on November the 30th 2018. A schematic representation of the steps undertaken in this subsection is shown in Figure 2.1.

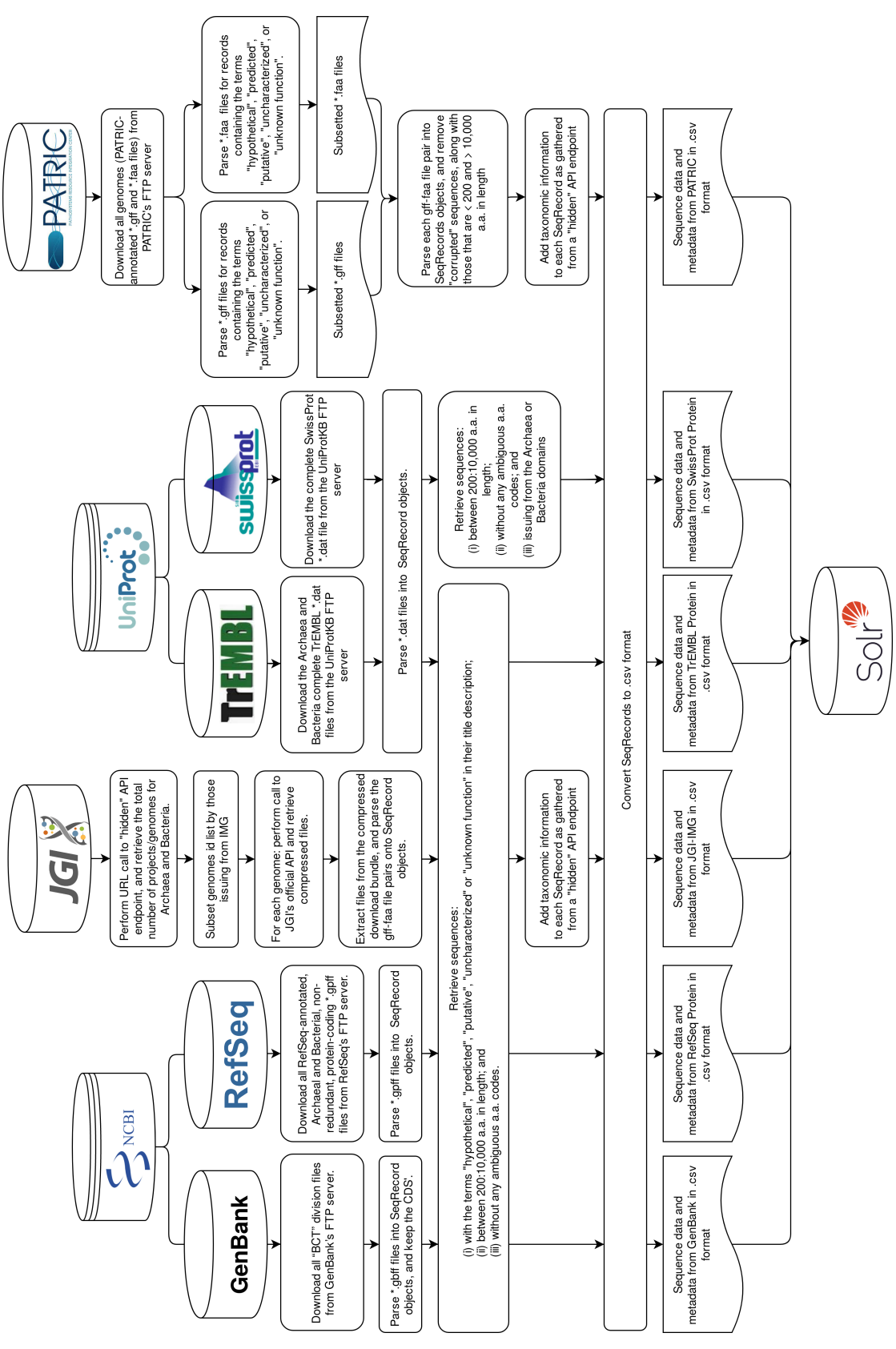


Figure 2.1: Overview of the steps undertaken in subsection 2.2.1 and 2.2.2. Each process is represented by a solid rectangle with rounded corners. The disks at the top represent public databases. The disks pertaining to the NCBI and UniProt Knowledgebases show additional lines. The disk at the bottom represents in-house data storage.

2.2.2 Inclusion Criteria and Quality Control

For each file retrieved in the previous steps, except the one issuing from SwissProt, we parsed its contents and only retrieved protein sequences that met the following criteria: (i) issuing from the Archaea or Bacteria domains; (ii) sequence length ranging between 200 and 10,000 amino-acids; (iii) not enclosing any ambiguous or rare amino-acid codes—B (Asparagine or aspartic acid), J (Leucine or Isoleucine), O (Pyrrolysine), U (Selenocysteine), X (Unspecified or unknown amino acid), Z (Glutamine or glutamic acid)—and (iv) whose description enclosed at least one of the following semantic terms associated with proteins of unknown function: "hypothetical", "predicted", "putative", "uncharacterized" or "unknown function". Every protein sequence that met these criteria was considered a protein of unknown function of prokaryotic origin. The total number of sequences retrieved this way was 134,894,520.

We parsed the contents of the SwissProt database but only retrieved protein sequences that met points (i) through (iii) of the criteria enunciated above. We did not apply any filter to the description of these sequences. The total number of sequences retrieved this way was 235,544 out of the 558,681 originally present in the "`uniprot_sprot.dat.gz`" file. These 235,544 sequences will be henceforth referred to as the "SwissProt dataset". We also created an Apache Solr data-framework, to enable data-storage, indexing, and querying. All protein sequence data and metadata was uploaded to this framework. A schematic representation of the steps undertaken in this subsection is shown in Figure 2.1.

2.2.3 Clustering Approach

2.2.3.1 Redundancy Reduction

We combined all 134,894,520 protein sequences of unknown function into a single FASTA file. Next, we downloaded and installed CD-HIT [21] (version `4.6.8-2017-1208`) according to the instructions at `https://github.com/weizhongli/cdhit/wiki/2.-Installation`, on May the 28th 2018. CD-HIT was run with this FASTA file as input, and the following parameters: minimum global sequence identity threshold "`-c`" set to "`0.95`" (i.e., 95%); "`-d`" set to "`0`" (use sequence name in FASTA header till the first white space); "`-M`" set to "`0`" (use all available RAM); and "`-T`" set to "`150`" (use 150 threads). All other parameters were set to default. The output FASTA file enclosed 27,633,277 cluster representative sequences. A schematic representation of the steps undertaken in this subsection is shown in Figure 2.2.

2.2.3.2 Hierarchical Routine

We performed chained runs of CD-HIT to create a hierarchical structure for the protein sequences of unknown function in a neighbor-joining way [1]. These chained runs consisted in using the output FASTA file from the previous run as input for the next run (i.e., seeding, as previously demonstrated by [22]). We used the output FASTA file from the clustering step described in subsection 2.2.3.1 as input for the first of 7 runs. We achieved these chained runs by automating each CD-HIT iteration. The minimum global sequence identity threshold parameter ("`-c`") was set in decremental steps of 5%, from 90% to 60%. The word size parameter [2] ("`-n`") was set to "`5`" for thresholds 70% to 90%, and "`4`" for thresholds 60% to 70%. For each of these iterations, we used the additional parameter settings: "`-d`" set to "`0`", "`-M`" set to "`0`", and "`-T`" set to "`150`". All other parameters were set to default.

Next, we used the "`clstr_rev.pl`" Perl script from the CD-HIT suite to combine the "offspring" output cluster membership data files (i.e., "`*.clstr`" files) with their "parent" cluster membership data files. We did this to propagate the cluster members from one global sequence identity threshold to the one below, without the need to re-cluster all original cluster representative sequences from the first run. At the end of this hierarchical routine, we gathered a FASTA file comprising the 12,654,843 protein sequence representatives at 60% global sequence identity resolution. These 12,654,843 protein sequences will be henceforth referred to as the "Hypotheticals" dataset. A schematic representation of the steps undertaken in this subsection is shown in Figure 2.2.

2.2.4 Molecular Function Characterization

2.2.4.1 Imputed Physico-Chemical Metadata

We used BioPython's [23] `ProtParam` module to calculate the physico-chemical properties for each protein sequence in the Hypotheticals dataset, using the FASTA file as input. The physico-chemical properties calculated were: sequence length; molecular weight; grand average of hydropathicity [24]; aromaticity [25]; isoelectric point, and instability index [26]. A schematic representation of the steps undertaken in this subsection is shown in Figure 2.3.

[1] https://github.com/weizhongli/cdhit/wiki/3.-User's-Guide#Hierarchically_clustering
[2] https://github.com/weizhongli/cdhit/wiki/3.-User's-Guide#CDHIT

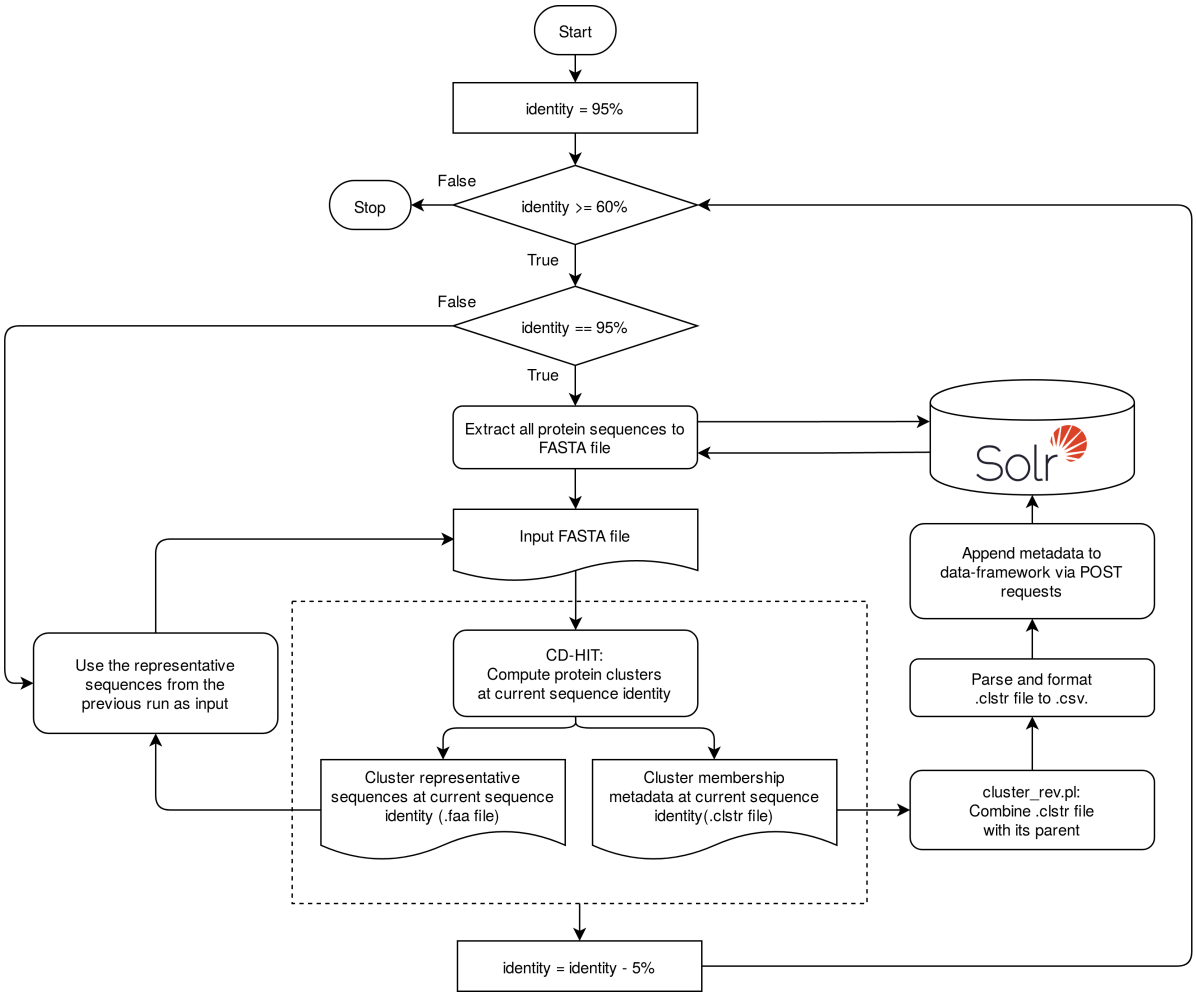


Figure 2.2: Overview of the clustering approach described in subsection 2.2.3. Each process is represented by a solid rectangle with rounded corners. The solid rectangles with sharp corners represent variables. The oval shapes portray the beginning and the end of the overall clustering routine. The diamonds illustrate conditional choices. The dashed rectangle symbolizes a group of processes that behave as one. The disk represents in-house data storage.

2.2.4.2 Gene Ontology estimation

To predict the GO terms for the Hypotheticals dataset, and that of SwissProt, we used DeepGO [27]. We downloaded and installed DeepGO along with its dependencies, according to the instructions in its GitHub repository (`github.com/bio-ontology-research-group/deepgo`), on December the 27th 2018. Next, we split the FASTA file described in subsection 2.2.3.2 into smaller files containing 1,000 proteins each. We did the same for the SwissProt dataset. Then, DeepGO was run in a multiprocessing setting, where one DeepGO instance would process one of the files originated this way. After each process had completed, another FASTA file would be assigned to it, until all individual FASTA files had been processed. A schematic representation of the steps undertaken in this subsection is shown in Figure 2.3.

2.2.4.3 Enzyme/Non-enzyme Classification and EC Number prediction

To predict the catalytic potential, along with the putative 3rd and 4th-digit EC numbers, for the Hypotheticals dataset, and that of SwissProt, we used DeepEC [28]. DeepEC was downloaded along with its dependencies, and installed on July the 15th 2019, according to the instructions in its Bitbucket repository (`bitbucket.org/kaistsystemsbiology/deepec/src/master/`). Next, DeepEC was run using the same multiprocessing setting, and the same input files, as those described in subsection 2.2.4.2. A schematic representation of the steps undertaken in this subsection is shown in Figure 2.3.

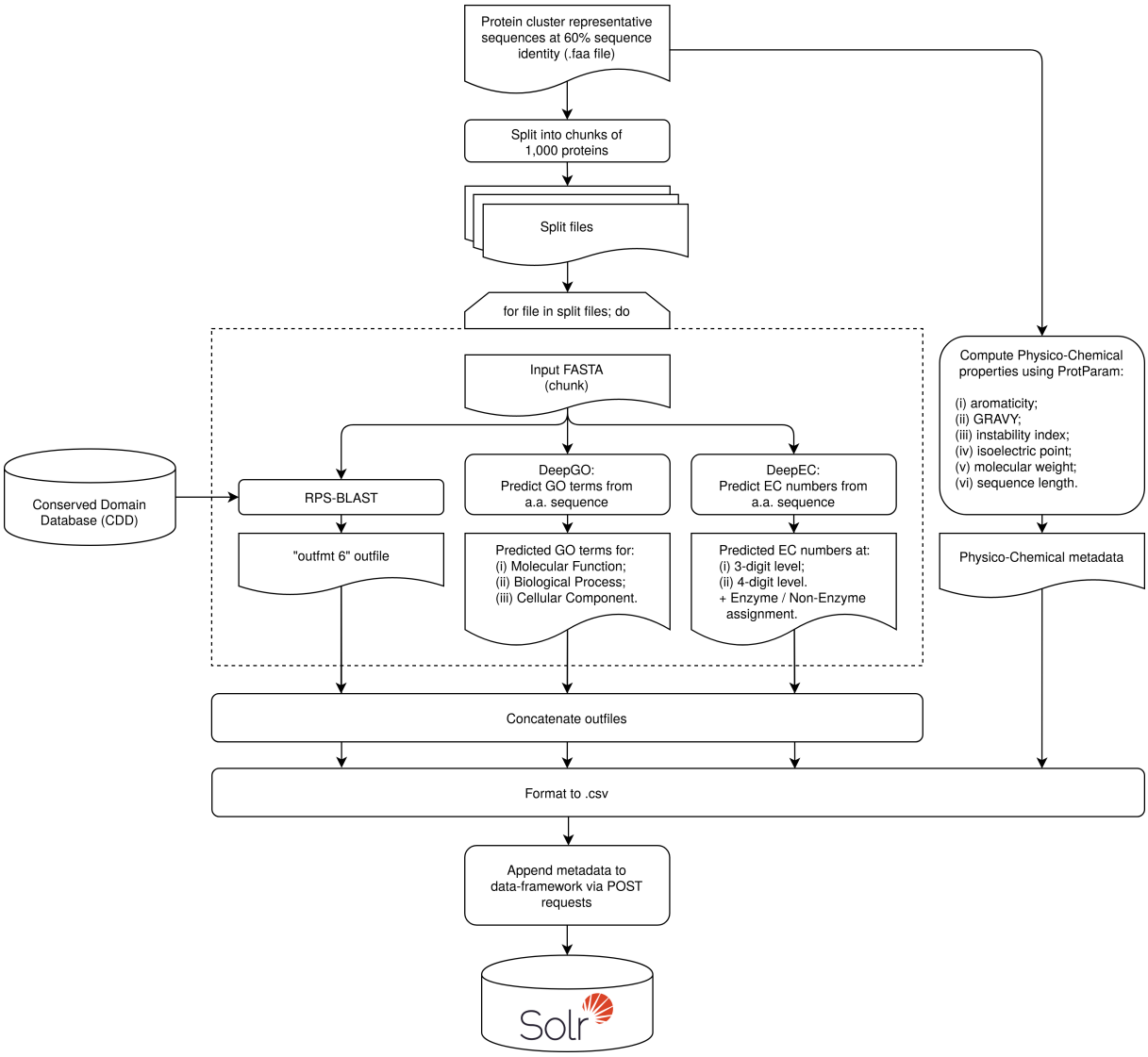


Figure 2.3: Overview of the molecular function characterization step described throughout subsection 2.2.4. Each process is represented by a solid rectangle with rounded corners. The solid rectangle with the top corners cut off represents the beginning of a for loop. The dashed rectangle symbolizes a group of processes that act on each of the individual split files inside the loop. The disk on the left represents a data resource, and the one at the bottom depicts in-house data storage.

2.2.4.4 Protein Domain Imputation

To search for domains present in the proteins from the Hypotheticals dataset, and that of SwissProt, we used the Reversed Position Specific Blast (RPS-BLAST) [29]. First, we downloaded the BLAST+ standalone executables from its FTP server directory at `ftp://ftp.ncbi.nlm.nih.gov/blast/executables/blast+/LATEST/` (version `2.7.1`). We also downloaded the pre-formatted Conserved Domain Database (CDD) [30] from its FTP server directory at `ftp://ftp.ncbi.nih.gov/pub/mmdb/cdd/little_endian/` ("`Cdd_LE.tar.gz`" file), on December the 13th 2018. This pre-formatted CDD search database contains all domain models present in the CD-Search tool [31] default database. Next, RPS-BLAST was run against CDD with the "`-outfmt`" parameter set to "6", using the same multiprocessing setting, and the same input files, as described in subsection 2.2.4.2.

To ensure the quality of the resulting domain annotations, we applied a *post hoc* bit score cutoff of $\geqslant 80$ to all hits retrieved by RPS-BLAST. We chose to filter our results by bit scores for 2 reasons. First, bit scores can be used to compare alignment scores from different searches, even by executing RPS-BLAST in a multiprocessing setting, because they are normalized with respect to the scoring system [32]. Second, bit scores may provide a better standard for inferring homology than percent identity (See [33]). A schematic representation of the steps undertaken in this subsection is shown in Figure 2.3.

2.2.5 Bit-score Normalization

We sought to normalize the bit-scores gathered for each domain imputation made by RPS-BLAST for two reasons. First, so that these are not as influenced by outliers as the original value distribution. Second, to rescale them from 0 to 1. To this end, we first performed a quantile transformation on the bit-score cumulative distribution function (which we generalized below as F_X for the random variable X), by using the quantile function of the standard normal distribution, also known as probit function (Φ^{-1}). Afterwards, we rescaled the resulting distribution with min-max feature scaling. We named the equation that subsumes both of these steps as Q, and define it as:

$$Q(x) = \frac{x' - x'_{min}}{x'_{max} - x'_{min}} \mid x' \in (\Phi^{-1} \circ F_X)(x) \tag{2.1}$$

Where the white circle denotes the function composition operation, between the quantile function of the standard normal distribution, and the cumulative distribution function of the random variable X, such that given two arbitrary functions f and g, we write $g(f(x)) = (g \circ f)(x)$. After these bit-score transformations, and as a quality control measure, only the transformed bit-scores whose original value was $\geqslant 80$ were considered for downstream processing.

2.2.6 Over-Representation Analysis

Consider that for a given classification system—i.e., EC, GO, or CD—the total population of proteins is the sum of those from the Hypotheticals dataset, plus those from SwissProt, that have at least one annotation from that classification system. The purpose of this analysis was to understand if the Hypotheticals dataset was enriched for a given annotation t (e.g., an EC number), in relation to the total protein population.

Let H_0 denote the null hypothesis, where the property of a protein to be annotated with t, and that of belonging to the Hypotheticals dataset are independent. Or equivalently, that $p_1 = p_2$, where p_1 is the probability of proteins from the Hypotheticals dataset to be annotated with t, and p_2 the probability of proteins from the SwissProt dataset to be annotated with t [34]. We can formulate this problem as a contingency table (Table 2.1), where: N is the total protein population; K is the number of proteins from the Hypotheticals dataset; n is the total number of proteins annotated with t; and k is the number of proteins from the Hypotheticals dataset annotated with t.

Table 2.1: Classification of proteins per dataset according to a given semantic term.

	$\in t$	$\notin t$	**Total**
Hypotheticals	k	$K-k$	K
SwissProt	$n-k$	$(N-K)-(n-k)$	$N-K$
Total	n	$N-n$	N

The exact distribution of the random variable X whose realization is the observed value k, is the hypergeometric distribution, with the following probability mass function:

$$p_X(k) = \Pr(X = k) = \frac{\binom{K}{k}\binom{N-K}{n-k}}{\binom{N}{n}} \tag{2.2}$$

Whose cumulative distribution function (CDF) F_X, is:

$$F_X(k) = \Pr(X \leq k) = \sum_{i=0}^{\lfloor k \rfloor} \frac{\binom{K}{i}\binom{N-K}{n-i}}{\binom{N}{n}} \tag{2.3}$$

Choosing a critical region to maximize the power of the test—and therefore the choice of the corresponding p-value—is dependent on the alternative hypothesis H_1 [34]. In the present case we define H_1 such that there is an over-representation of proteins from the Hypotheticals dataset that are annotated with t, in relation to those from SwissProt (i.e., $p_1 > p_2$).

Consequently, the test to be performed is one-sided—i.e., the critical region is the upper tail of the hypergeometric distribution. As such, we used the survival function of the hypergeometric distribution, in order to calculate the p-value associated with each annotation t, which is given by:

$$\Pr(X > k) = 1 - F_X(k) = 1 - \sum_{i=0}^{\lfloor k \rfloor} \frac{\binom{K}{i}\binom{N-K}{n-i}}{\binom{N}{n}} \tag{2.4}$$

For this analysis, and for each protein, we only considered the GO terms at maximum depth of its sub-DAG, the 4th-digit EC numbers, and the most informative CD(s). The most informative CD of a protein is the CD annotation, whose probability of occurrence is the lowest, among all CD annotations for that protein.

In order to control the false positive rate, we applied the Benjamini-Hochberg (BH) correction [35] to all p-values calculated this way. For each term that underwent this analysis we calculated the proportion of proteins from the Hypotheticals dataset annotated with t (denoted by k), divided by the total number of proteins annotated with t (denoted by n)—i.e., the "Rich factor".

We only considered the results for terms whose "Rich factor" $\geqslant 0.95$, and whose BH-corrected p-value $\leqslant 1e-5$. The reason for this stringency was to ensure we retrieved the most statistically significant term over-representations.

2.3 Results and Discussion

2.3.1 Putative Unfunctionalized Protein Families

First, we sought to create a data repository comprising protein sequences of unknown function, from prokaryotic origin. To do so, we gathered protein sequence data from 5 distinct public databases (Table 2.2). We chose these databases because they employ different annotation pipelines to characterize their sequence data. Thus, they arguably produce non-overlapping sequence annotations. The number of retrieved sequences was 134,894,520.

Table 2.2: Number of protein sequences of unknown function gathered from each datasource.

Datasource	No. of proteins	Percentage
PATRIC	95,324,268	70.67%
TrEMBL	15,979,932	11.85%
IMG	14,392,089	10.67%
RefSeq	9,157,884	6.79%
GenBank	40,347	0.03%
Total	134,894,520	100%

To mitigate the computational load required to analyse $>$ 134 million sequences, we applied a multi-step clustering approach using CD-HIT [21], as it had been done before [22, 36, 37]. This approach spanned 8 clustering iterations. In the first iteration we wanted to reduce the redundancy of our dataset. To this end, we clustered all sequences using a 95% minimum global sequence identity threshold. From this first iteration we gathered 27,633,277 protein sequence cluster representatives. This way, our initial dataset of $>$ 134 million protein sequences can be represented by a smaller, non-redundant one, where a single representative is assigned to each cluster.

Next, we assorted this non-redundant dataset into putative protein families. We aimed to do so for two reasons. First, we wanted to understand the extent to which these proteins differed from one another. Second, we sought to reduce the size of our dataset as much as possible, while maintaining reasonably conserved protein clusters. Thus, the $\sim$27 million representative sequences were clustered further through a 7-step hierarchical routine (Figure 2.2). We achieved this by (i) setting CD-HIT's minimum global sequence identity threshold in decremental steps of 5%, from 90% to 60%; and by (ii) using the cluster representatives from the previous run as input for the next run. By doing so, we were able to map our $\sim$27 million sequences throughout

7 global sequence identity thresholds. We chose not to set a minimum alignment length cutoff, allowing for shorter sequences to be clustered with larger ones. Our reasoning is that these shorter sequences could be fragments of the same larger protein.

Afterwards, we wanted to understand how these representative sequences were arranged throughout the thresholds. Namely the number of clusters and that of singletons. We noticed that the majority of representative sequences from each clustering run were singletons, regardless of threshold (Figure 2.4.A). This suggests that these sequences of unknown function are very dissimilar to one another. The number of clusters increases from threshold 90% throughout 65%, and decreases from 65% to 60%. This could be due to the fact that at the 60% threshold, the representative sequences bear enough sequence identity to each-other, so that the rate at which clusters combine between themselves, surpasses the rate at which singletons aggregate to form new clusters.

We observed that both the number and proportion of singletons decrease throughout the thresholds (Figure 2.4.B). By the 70% threshold, the number of sequences belonging to a cluster (54.9%) outweighs that of singletons (45.1%). At the 60% threshold, the ratio of sequences belonging to a cluster is 67.6%, whereas that of singletons is 32.4%. This implies that, by decreasing the global sequence identity threshold, more singletons aggregate to form new clusters, in addition to being assigned to pre-existent ones.

A

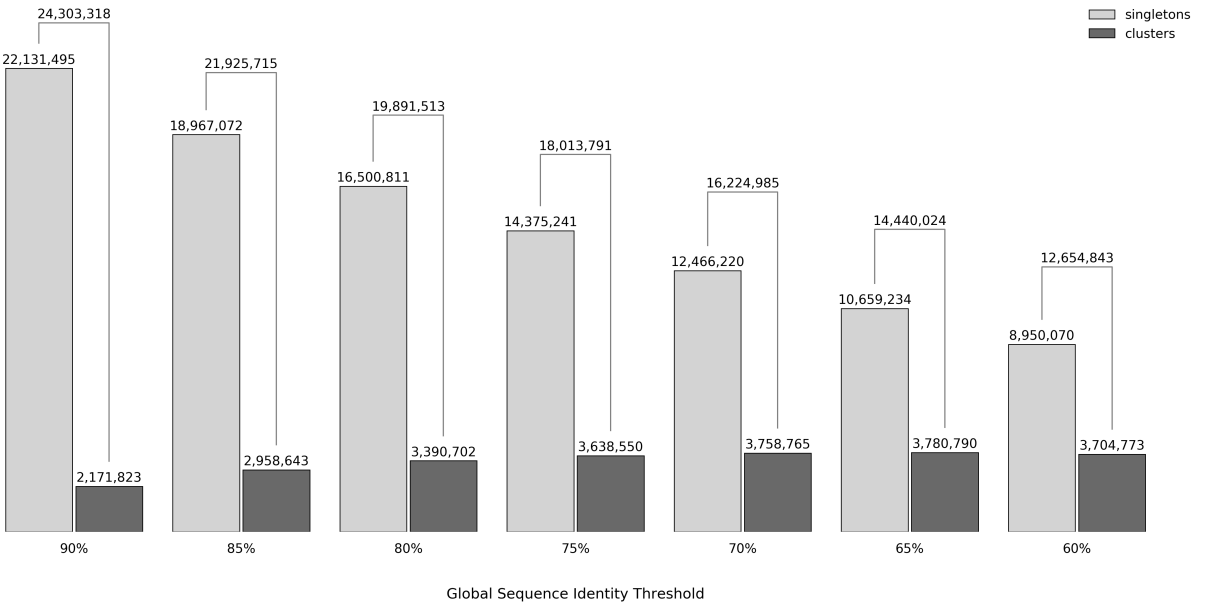


Figure 2.4. (Continued on next page.)

B

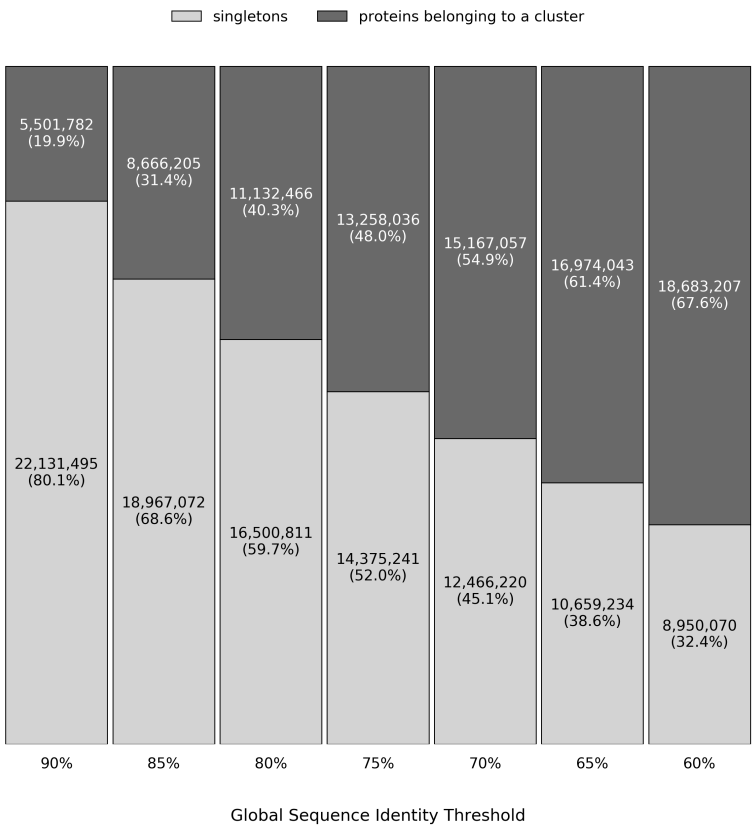


Figure 2.4: (**A**) Number of protein sequence representatives belonging to clusters (dark grey), and that of singletons (light grey), per global sequence identity threshold. The protein sequence counts regarding each group are shown on top of its respective bar, and the total number of representatives for each threshold is shown on top of brackets. (**B**) Cluster (dark grey) and singleton (light grey) membership for the 27,633,277 protein representative sequences gathered from the first clustering run with a 95% sequence identity threshold, and throughout the succeeding sequence identity thresholds. Each subgroup is shown on top of one another, and their sum amounts to 100%.

We also inquired about the extent to which the clusters were conserved. To this end, we assessed the distributions for the mean and standard deviation of global sequence identity values, between the cluster members and their representatives (Figure 2.5), per threshold. We have also analyzed the distributions of cluster sizes (Figure 2.6).

The distribution of mean values range from the threshold of global sequence identity set by the clustering run, up to 100%. This is observable for all thresholds (Figure 2.5). At the 60% threshold, the interquartile range for the mean values spans from ∼78% to ∼90%. The density of these values is also greater from the value of 80% global sequence identity upwards. The values for standard deviation of global sequence identity to the cluster representative range from 0% to $<$ 30% at the 60% threshold, and the interquartile range of this distribution spans from 10% to 20%. These observations suggest that even at this threshold, there are highly conserved clusters.

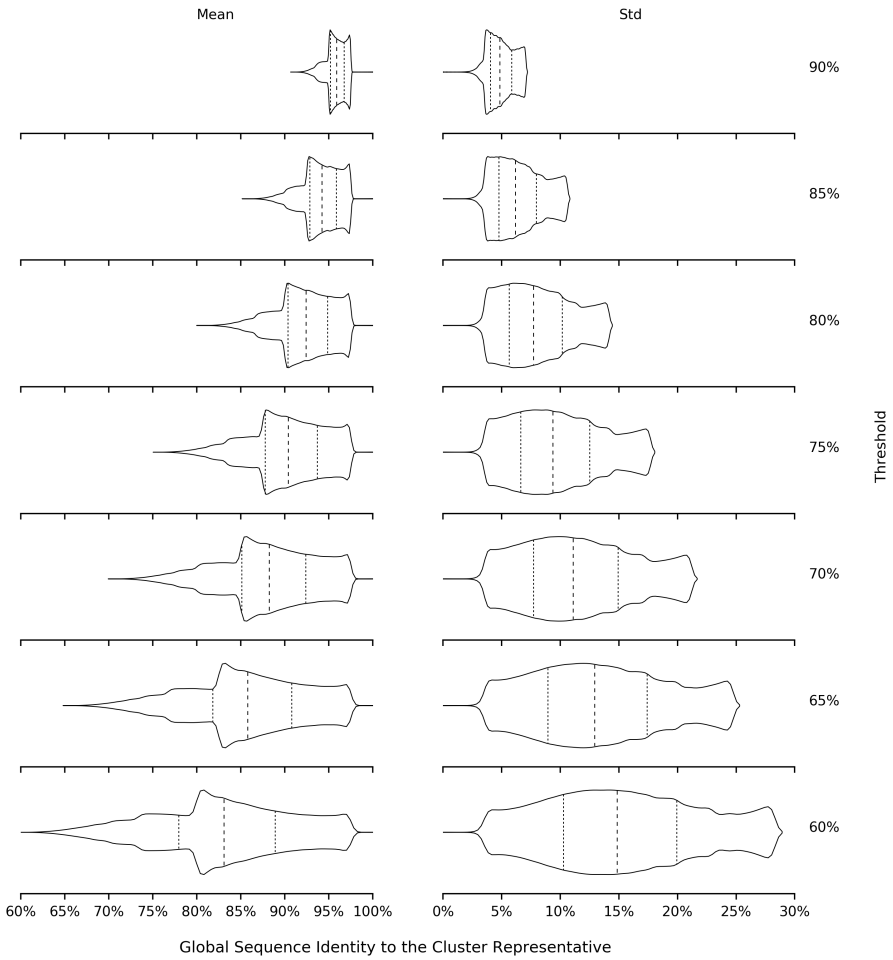


Figure 2.5: Distributions of the mean, and standard deviation values of global sequence identity, between the cluster members and their cluster representatives, per global sequence identity threshold. The x-axis denotes the values of global sequence identity between a given cluster member and its representative, and the y-axis denotes the global sequence identity threshold of the respective CD-HIT clustering run. The dotted lines represent the 1st and 3rd quartiles, and the dashed lines represent the median. The singleton representative sequences were excluded from this analysis.

We observed that the majority of clusters comprise < 100 members, regardless of threshold (Figure 2.6). The size of the largest cluster increases from $\sim$250 at the 90% threshold, to $\sim$1,700 at 60%. These results indicate that the clusters are conserved to the extent that even at the 60% threshold, the majority of clusters is still less than 100 members in size.

We chose to use the 12,654,843 protein sequence representatives at the 60% global sequence identity threshold, for all succeeding metadata imputation tasks. We chose to do so for two reasons. First, the 60% threshold offered the smallest amount of protein sequences to work with, allowing for faster computation. And second, at this threshold, we postulate that these proteins are grouped into well-conserved families, given the hierarchical nature of the implemented clustering routine. Therefore we posit that they share the same hypothetical molecular function. We will henceforth refer to these $\sim$12 million proteins as the "Hypotheticals dataset".

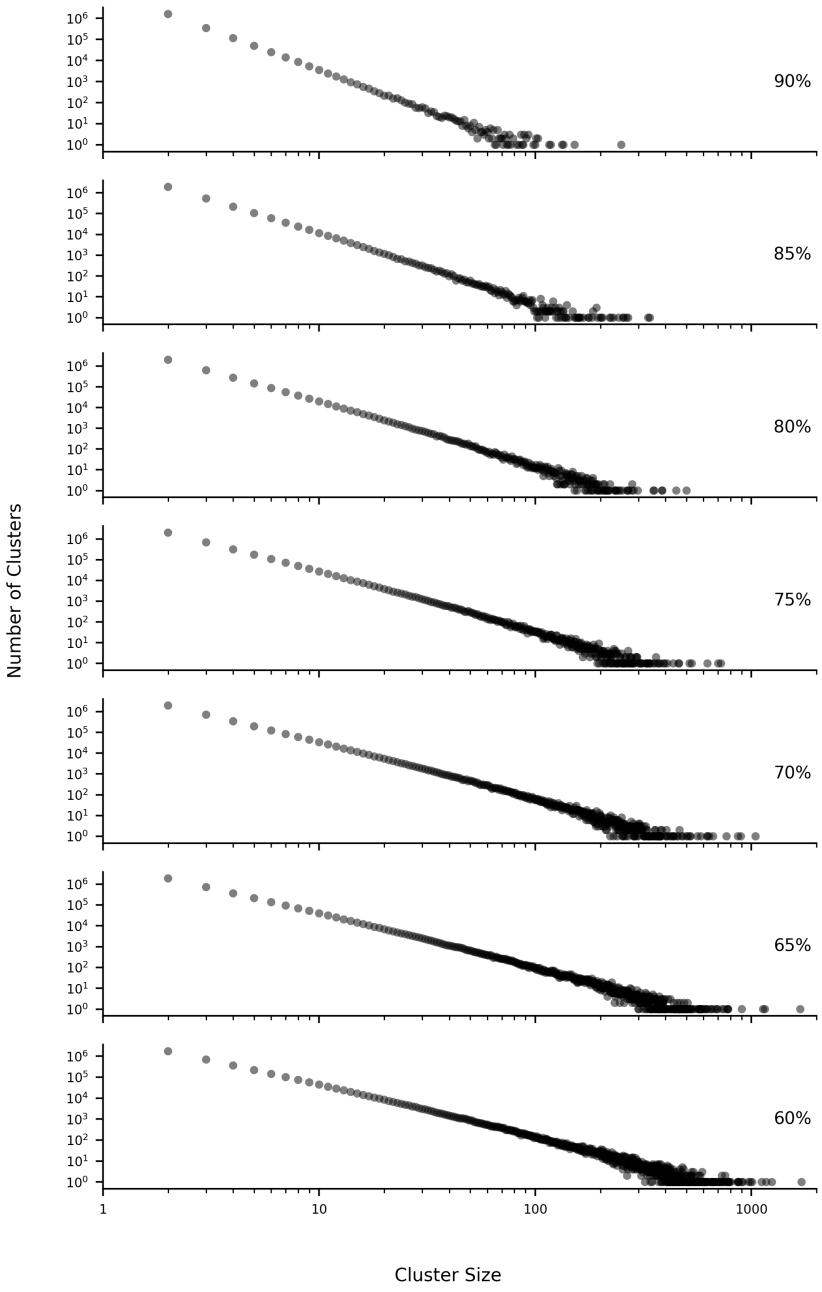


Figure 2.6: Cluster number distribution (y-axis) in function of cluster size (x-axis), per global sequence identity threshold. Both axes are in logarithmic scale. The singleton representative sequences were excluded from this analysis.

2.3.2 Metadata imputation and data source representativity

Our next goal was to predict the catalytic potential and the molecular function for all proteins in the Hypotheticals dataset. To that end, we used three *in silico* tools. The first was DeepGO [27], an ontology-aware classifier that predicts protein function from sequence, and outputs its results in the Gene Ontology (GO) format. The second was DeepEC [28], a deep learning-based framework that predicts the catalytic potential of a given protein, and its 4th-digit Enzyme

Commision (EC) number. The third was the Reversed Position Specific Blast (RPS-BLAST) [29], that queries a given protein sequence against a database of pre-calculated Position-Specific Score Matrices (PSSMs), and outputs the matching domains. In this work we chose to use NCBI's Conserved Domain Database (CDD) [30] as our PSSM database. After using all ∼12 million proteins from the the Hypotheticals dataset as input, we wanted to know how many were annotated, and how many intersections arose between these three classification systems (Figure 2.7.A). In this work we exclusively address the GO terms for the Molecular Function (MF) category.

We observed that 99.91% of the proteins from the Hypotheticals dataset have at least one GO term. This annotation coverage could be due to the fact that the GO hierarchy is very unspecific at the first levels of depth. Therefore, an unspecific annotation might be easier to infer given an uncharacterized protein sequence. For this reason, we posit that DeepGO might be the most permissive out of the three tools that we used.

A

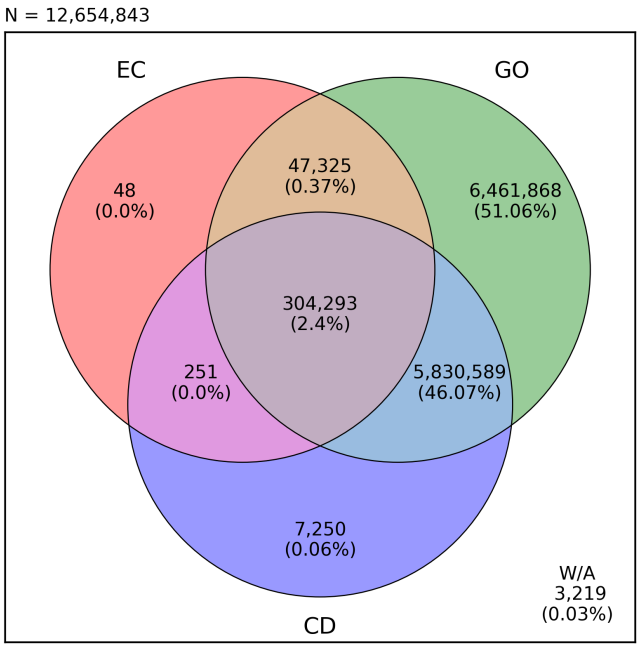


Figure 2.7: (**A**) Classification system membership of the proteins from the Hypotheticals dataset. The number of proteins that failed to be annotated by both DeepGO, DeepEC, and RPS-BLAST are shown outside of the Venn diagram, and are titled with "W/A": without annotation. (**B**) Representativity of the 5 different data sources, among cluster members pertaining to each of the representative proteins from the Hypotheticals dataset. Each row of the plot contains one Venn diagram (left), and one stacked bar chart (right). The Venn diagram depicts which slice from **A** is being addressed (color-filled). The counts to the right of each stacked bar represent the total number of representative sequences, and the y-axis denotes the number of distinct data sources being represented. The percentages enclosed in parenthesis relate to the total number of sequences in each slice of the Venn diagram to the left. For each bar, the subgroup pertaining to singletons is shown in light grey; the one whose representatives belong to clusters with size between 2 and 4 sequences, in dark grey; and the one whose representatives belong to clusters with a size $\geqslant$ 5 sequences, in black.

Figure 2.7. (Continued from last page.)

On the other hand, only 2.78% of proteins have an assigned EC number. The reason for this might be that 97.22% of these representatives—i.e., those devoid of an EC number—may not be enzymes at all. Out of ∼12 million, 48.53% have at least one Conserved Domain (CD), and 46.07% have at least one GO term in addition to one CD. We reckon that this amount falls short of the expected, given that RPS-BLAST can detect even remote homologies [29]. However, this percentage might be explained by our stringent bit-score cutoff (see subsection 2.2.4.4).

Moreover, 2.4% of these ∼12 million proteins were annotated by all three classification systems. We consider this to be reasonable, given that only 2.78% have a predicted EC number in the first place. This means that 86.33% of the representatives with at least one EC number also have at least one GO term, and one CD. Conversely, the percentage of representatives that could not be annotated by any tool is 0.03%.

We also inquired about the data source representation within each cluster, per classification system and intersections thereof (Figure 2.7.B). The most abundant elements are singletons from a single data source, irrespectively of the Venn diagram slice being addressed. This was expected, as singleton sequences represent the majority of representatives (see Figure 2.4.A).

The majority of cluster members belong to either a single data source or two, also regardless of the Venn diagram slice. This could be an effect from the cluster size, given that at the 60% threshold most clusters have less than 5 members (see Figure 2.6). It is also possible that some clusters lack multi-data source representativity. This is suggested by the presence of clusters with 5 or more elements spanning only one to two data sources (Figure 2.7.B). The reason for this might be that 70.67% of our initial cohort belong to a single datasource (see Table 2.2).

2.3.3 Gene Ontology terms

Intrigued that 99.91% of proteins from the Hypotheticals dataset had been annotated by DeepGO, we wondered what their putative functions were. To address all 12,644,075 proteins we first needed to group them by the broadest level of the GO DAG—i.e. the first (Figure 2.8).

We noticed that there were proteins that had more than one GO term at the first level of depth. This might indicate two different scenarios. On the one hand, it suggests that these proteins have multiple molecular functions. On the other hand, it implies ambiguity regarding DeepGO's predictions. This appears to be reflected to some extent in the prediction scores, where the groups of proteins with a single GO term at this depth seem to rank higher. This

might also be due to the fact that the prediction scores shown for the proteins with multiple GO assignments correspond to the average between the individual values.

Within each group, only a small fraction of proteins possess a GO term annotation deeper than the 3rd level of the GO DAG. This emphasizes on how vaguely annotated these proteins are. It also justifies the extensive annotation coverage provided by DeepGO, and arguably, its permissiveness. Alternatively, this might also indicate that these representatives are very divergent from those used by DeepGO as a training set. This supposition might explain the broadness of GO annotations at this depth.

Nearly all proteins (99.39%) were annotated with the GO term "binding" (41.86%), "catalytic activity" (4.7%), or both (52.82%). This implies that, according to DeepGO, 57.52% of our representatives might be enzymes. It also contrasts with the number of proteins annotated by DeepEC—from all 12,654,843 representatives, only 2.78% were annotated as possible enzymes (see Figure 2.7.A).

The group of proteins with the highest prediction scores are those with "catalytic activity" as their sole GO term at the first level of depth. This is curious because it suggests that DeepGO's predictions for putative enzymes ranked best.

2.3.4 Enzyme Commission numbers

We were interested in identifying putative enzymes in the Hypotheticals dataset. For this reason we used DeepEC [28], as previously stated. DeepEC uses three convolutional neural networks (CNNs) to predict the enzymatic function of a given protein sequence. The first CNN predicts whether the protein is an enzyme or not. The second CNN predicts the 3rd-digit EC number. And the third CNN predicts the 4th-digit EC number.

To mitigate false-positives, DeepEC only provides results for proteins that were predicted to be enzymes by the first CNN, and whose predicted 3rd-digit EC number is a prefix of the predicted 4th-digit one.

DeepEC also performs homology analysis if one of the CNNs fails to predict an EC number [28]. Additionally, it outputs to the user the log files corresponding to each CNN prediction, and its prediction scores. The predictions from homology analysis lack prediction scores. For that reason they were excluded from any downstream analysis. There were 15,333 proteins that were excluded from our results as a consequence.

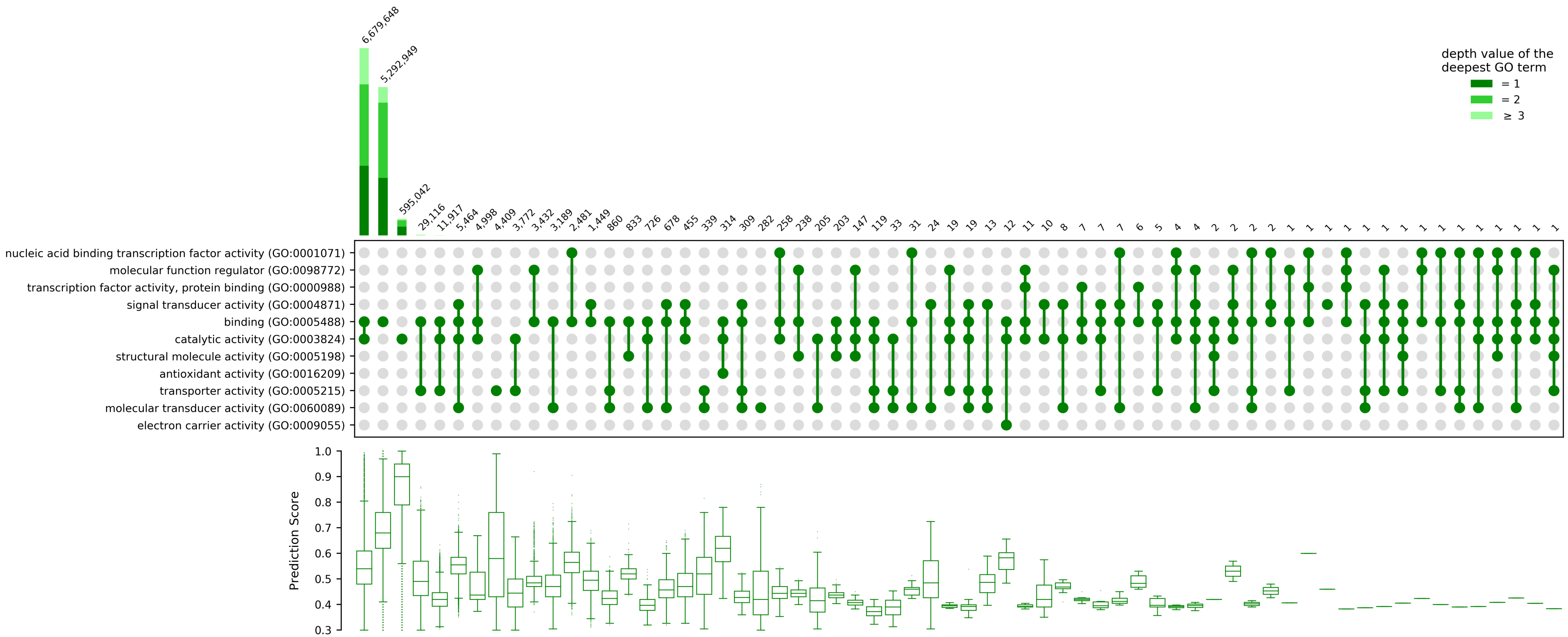


Figure 2.8: GO term assignment of the proteins from the Hypotheticals dataset. Only the GO terms at the 1st level of depth of the GO-DAG are shown. The proteins were grouped by their GO term assignments at the 1st level of depth, and each group of proteins can have more than one GO term at this level. Each column represents a group of proteins sharing the same GO term, or alternatively, the same GO term combination. Each group is represented vertically by a stacked bar chart (top panel), a dot matrix (middle panel), and a boxplot (bottom panel). Each bar is subsetted into 3 subgroups according to the depth value of the deepermost GO term of a given protein. The numbers on top of each bar represent the total number of protein sequences for that group. The y-axis of the boxplots refers to the "DNN" values associated with each GO term assignment as outputted by DeepGO. If a protein shares more than one GO term at the 1st level of depth, the prediction score for that combination corresponds to the mean of the individual values.

Moreover, DeepEC can assign more than one EC number to a given protein. We were curious about how many representatives had multiple EC assignments, and what types of assignments these were. Therefore, we sought to address this subset of proteins separately.

Out of 336,584 representatives with predicted EC numbers, 9,622 (2.86%) had multiple assignments (Figure 2.9.A). We identified four classes of proteins with multiple EC assignments, according to the EC digit they shared (Figure 2.9.A, Table S2.1–Table S2.4). The majority (90.95%) of these proteins either share the same 3rd-digit EC (87.16%), or the same 2nd-digit EC (3.76%) in their assignments. The 2nd and 3rd-digit of an EC number characterize the reaction that particular enzyme catalyzes in terms of several factors, like the reactive species, the type of bond being acted upon, and group or product involved. With this in mind, we posit that the representatives sharing either the 2nd or 3rd EC digit might be promiscuous enzymes. This catalytic promiscuity might explain why they failed to be annotated in the first place.

The remaining proteins with multiple EC assignments either share the 1st-digit EC (1.26%), or no EC digit at all (7.78%) in their assignments. We argue that these groups might enclose "moonlighting" [38] enzymes, or proteins that underwent gene fusion events. We make this assumption based on the fact that their multiple assignments either share the same enzymatic group (i.e., 1st-digit EC), or span multiple enzymatic groups altogether.

A

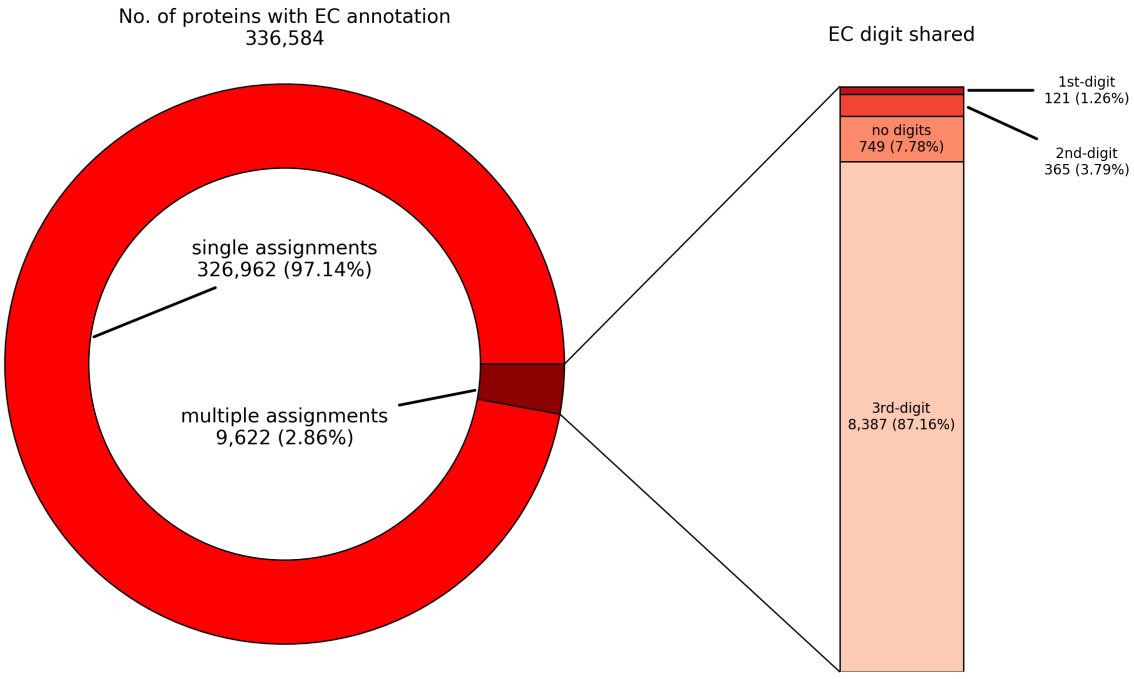


Figure 2.9. (Continued on next page.)

B

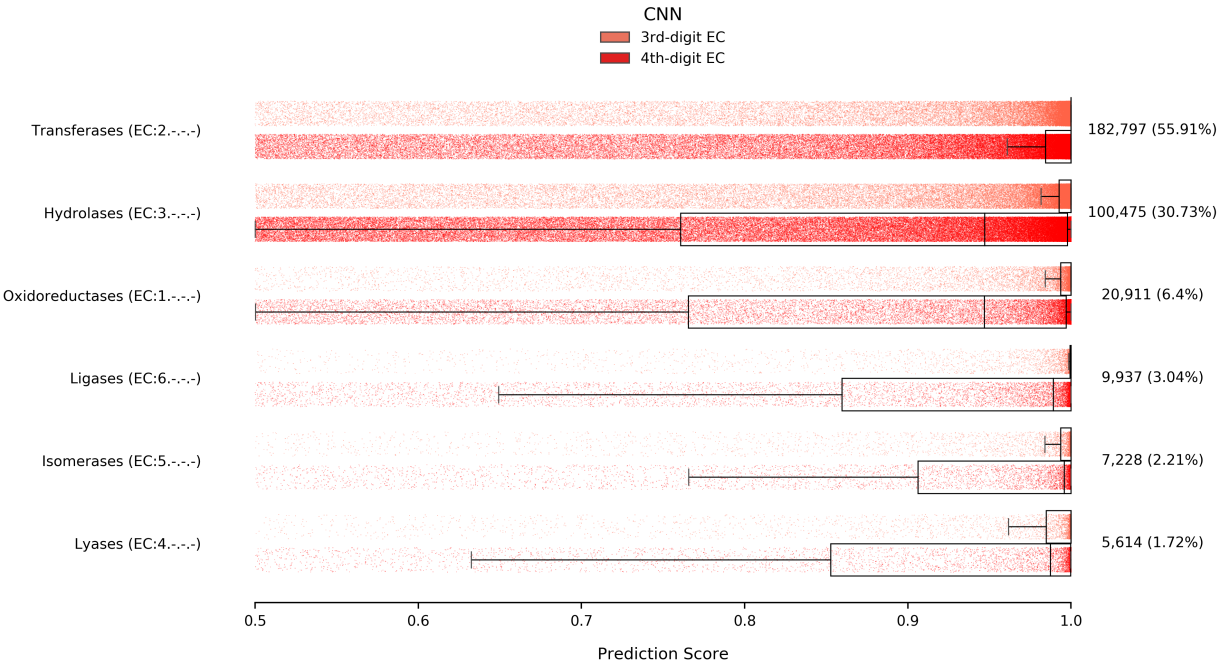


Figure 2.9: (**A**) Amount of proteins from the Hypotheticals dataset with one or more EC assignments (donut chart). The number of proteins refers to those that were annotated by DeepEC's CNNs (i.e., not by homology analysis). There are 4 classes of representative proteins with multiple EC assignments (barchart). These are assorted according to the EC digit they share: The 3rd-digit (e.g., EC:X.X.X.-); the 2nd-digit (e.g., EC:X.X.-.-); the 1st-digit (e.g., EC:X.-.-.-); or no digits (e.g., EC:X.-.-.- and EC:Y.-.-.-). (**B**) Proteins from the Hypotheticals dataset with single EC assignments, grouped by the 1st-digit EC number. The x-axis of the boxplots refers to the "DNN" values for each EC number prediction. The prediction scores refer to two out of three DeepEC's CNNs. The x-axis lower limit starts at DeepEC's default prediction score threshold. The total counts for each group are shown to the right. The percentages are in relation to the total number of representative proteins with single EC assignments.

Next, we analysed all 326,962 representatives that were assigned a single EC number. We also wanted to study the underlying CNN prediction scores for these assignments. To avoid redundancy, we only assessed the scores for the CNNs that predict the 3rd-digit and the 4th-digit EC. This way, we grouped the 326,962 representatives by their 1st-digit EC, and displayed the prediction score distributions for each of these two CNNs (Figure 2.9.B).

The two most abundant groups of proteins with a single EC assignment are those of Transferases (55.91%) and Hydrolases (30.73%). The majority of prediction scores for the 3rd-digit EC CNN are higher than 0.95. This suggests high confidence in their 3rd-digit EC predictions, regardless of enzymatic group being addressed. This is particularly evident for the Transferase and Ligase groups.

The prediction scores for the 4th-digit EC CNN have a larger spread. The interquartiles for these scores range from ∼0.75 to 1. This might suggest that predicting a 4th-digit EC number

from protein sequence alone is more challenging than predicting its respective 3rd-digit EC. It is possible that the score of a given prediction decreases with the specificity of the predicted annotation. This makes sense given that the 4th-digit EC classifies the overall reaction of an enzymatic sub-subclass, and is therefore the most specific level of EC annotation.

2.3.5 Conserved Domains

We wanted to know what CDs were present in the 6,142,383 representative proteins that had at least one CD assignment. However, given both the size of this cohort, and the fact that each protein might have multiple CDs, we had to find a way to summarize this data.

To do so, and for each protein, we mapped its CD PSSM ids with the site types they are associated with. These site types are generic annotations of the high-level function played by a specific CD. Each CD might have multiple site types annotations, or none.

Afterwards, we grouped these proteins by their distinct CD site types—or combinations thereof—and counted how many proteins were in each group (Figure 2.10). More than half (66.11%) of these proteins have no CD site type annotation. The reason for this is that these site types only exist for NCBI-curated domain models [3], whereas CDD encloses domains from multiple sources besides those curated at NCBI. Moreover, 3.13% of these proteins have domains that are solely annotated by CDD as 'unassigned or type "other"'.

Additionally, we wanted to understand the distribution of the number of CDs per protein, and the relation between the number of CDs and sequence length. The majority of representatives has less than 100 CDs per protein sequence (Figure 2.11.A). This is also observed when assessing the number of distinct CDs per protein sequence instead (Figure 2.11.B).

However, either addressing the total number of CDs per protein sequence, or the number of distinct ones, a trend can be witnessed in both distributions, between 100 and ∼280 CDs (Figure 2.11). This "plateau" suggests that there is a continuity in the frequency of proteins that have 100-280 CDs. Moreover, the maximum number of distinct CDs per protein peaks near 500 (Figure 2.11.B). We observe two linear trends upon addressing the relationship between the number of CDs and protein sequence length (Figure 2.12.A). These trends appear to behave asymptotically upon addressing the distinct number of CDs instead (Figure 2.12.B).

[3] ftp://ftp.ncbi.nih.gov/pub/mmdb/cdd/README

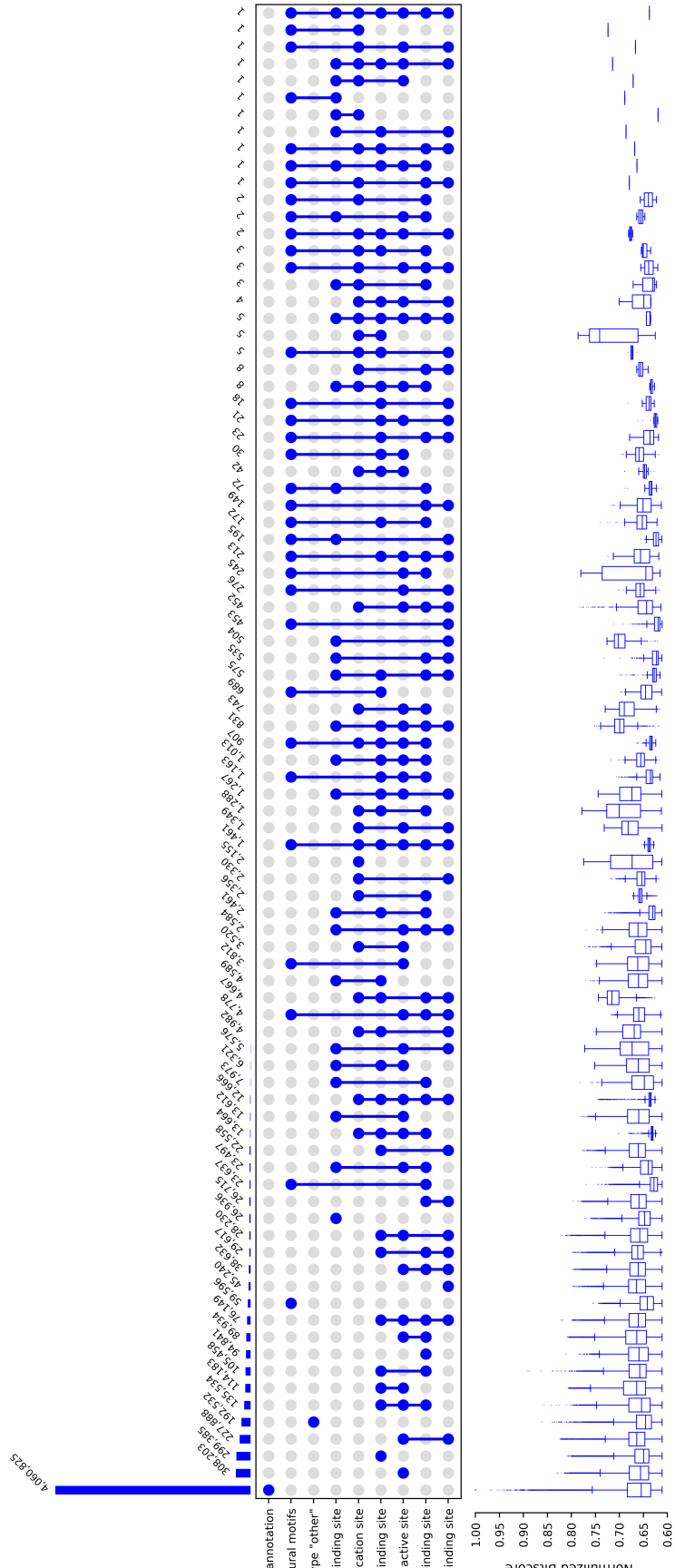
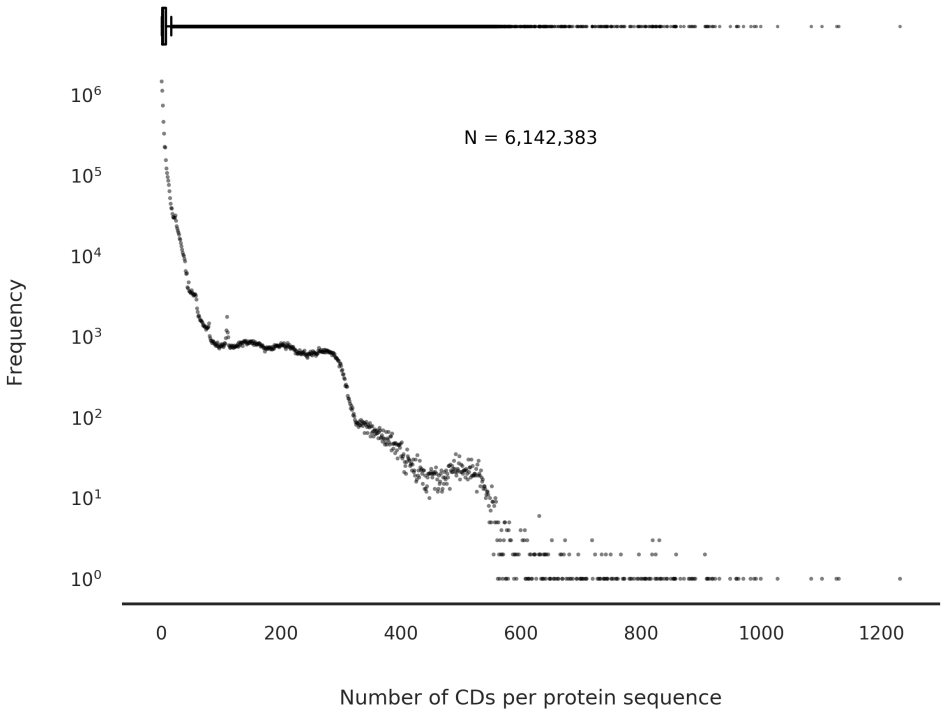


Figure 2.10: CD assignment for 6,142,383 proteins from the Hypotheticals dataset. The proteins were grouped by their distinct CD site type assignments. Each column represents a group of proteins sharing the same distinct CD site type assignment, or alternatively, the same distinct CD site type assignment combination. Site type combinations might either originate from a single CD that has multiple sites; multiple CDs with a single site; or multiple CDs with multiple sites. To avoid redundancy, we removed "No annotation type" and "unassigned or type other" from the groups of proteins whose CDs had other site types besides these. Each group is represented vertically by a bar chart (top panel), a dot matrix (middle panel), and a boxplot (bottom panel). The numbers on top of each bar represent the total number of protein sequences for that group. The y-axis of the boxplots refers to the mean values of the normalized bit scores associated with each CD assignment as outputted by RPS-BLAST, per protein.

A



B

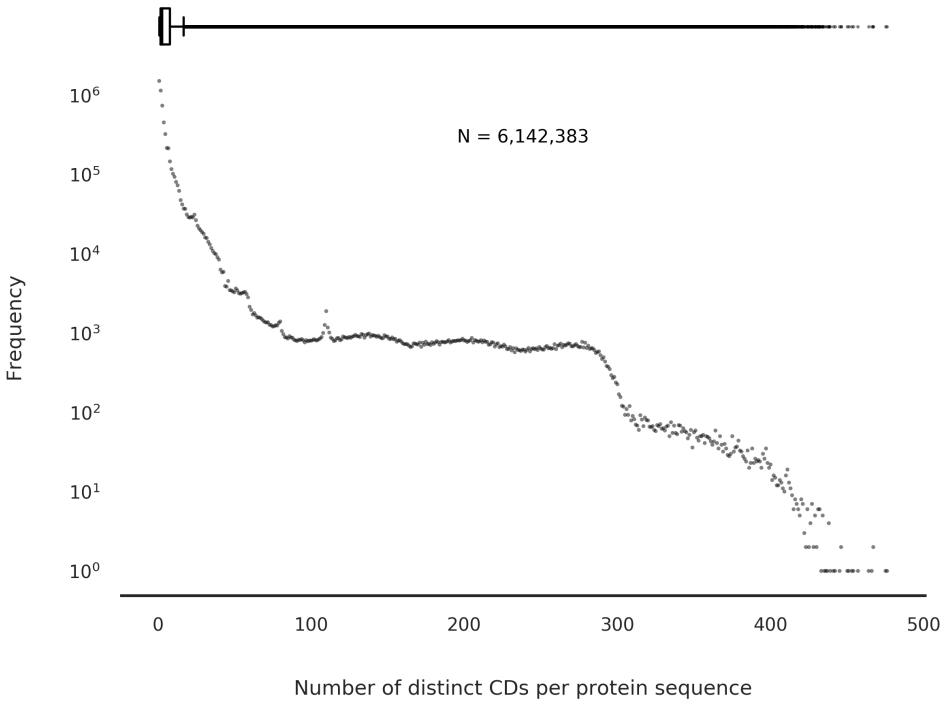
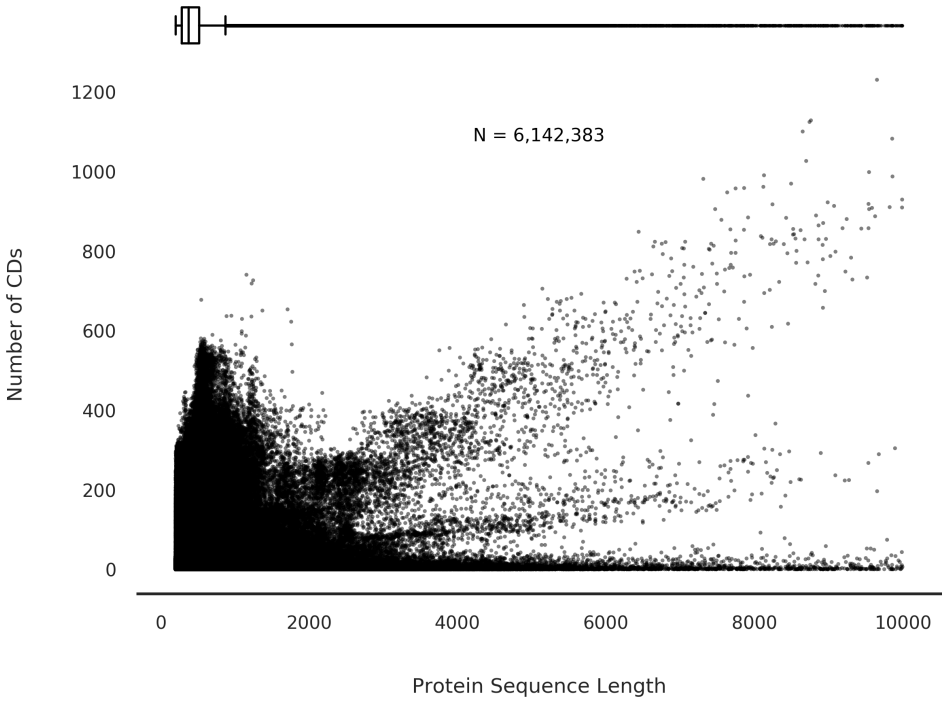


Figure 2.11: Frequency distributions of the total number of CDs per protein sequence (**A**), and the number of distinct CDs per protein sequence (**B**). Both distributions are also represented as boxplots at the top of each figure. The y-axis is in logarithmic scale. The number in the top-center of the plot refers to the total number of proteins from the Hypotheticals dataset that have at least one CD, with a bit score $\geqslant 80$.

A



B

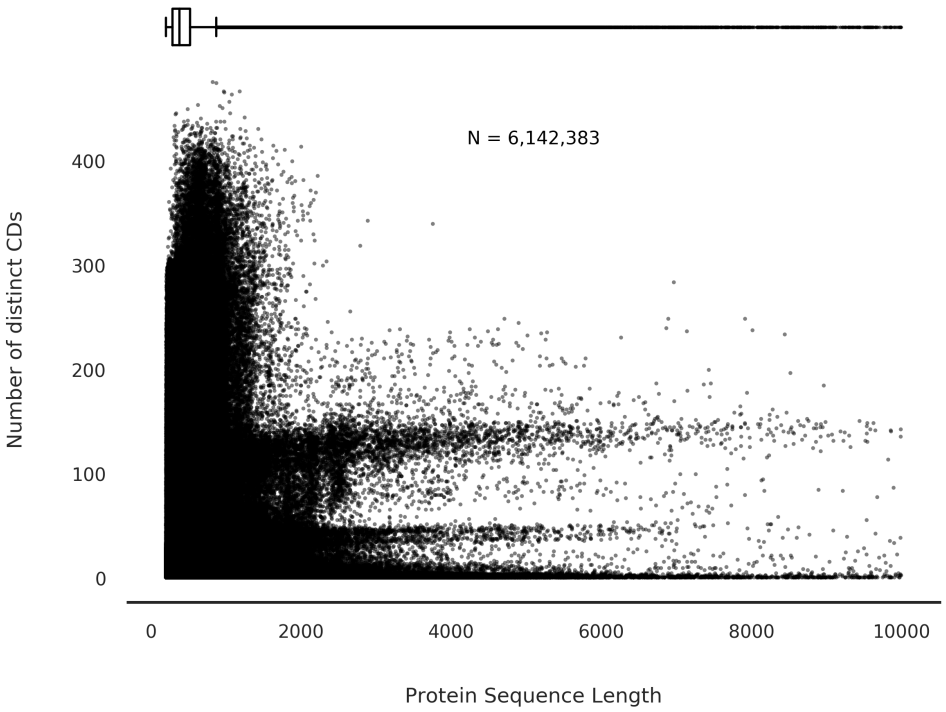


Figure 2.12: Distributions of the total number of CDs per protein sequence (**A**), and the number of distinct CDs per protein sequence (**B**), in function of protein sequence length. Both distributions are also represented as boxplots at the top of each figure. The number in the top-center of the plot refers to the total number of proteins from the Hypotheticals dataset that have at least one CD, with a bit score $\geqslant 80$.

2.3.6 Physico-Chemical Metadata

We sought to survey the physico-chemical properties of these representative sequences. We were also interested in understanding possible relations between these properties.

To this end, and for each protein, we computed the following physico-chemical properties: (i) aromaticity coefficient; (ii) grand average of hydropathicity; (iii) instability index; (iv) iso-electric point (pI); (v) molecular weight; and (vi) sequence length.

The aromaticity coefficient for most proteins ranges between 0 and $\sim$0.25, while only a few show values above 0.5 (Figure 2.13). Aromaticity has no correlation with the remaining physico-chemical properties, albeit a weak negative correlation with the instability index ($\rho = -0.22$, p-value $= 0.0$).

The grand average of hydropathicity ranges between -2.5 and 2.5, with the majority of these proteins centering near the value of 0. Similarly to the aromaticity coefficient, the grand average of hydropathicity also has no correlation with the other physico-chemical properties, with the exception of the instability index, for which a weak negative correlation is shown ($\rho = -0.24$, p-value $= 0.0$). Both the grand average of hydropathicity and the coefficient of aromaticity tend to 0 in relation to increasing molecular weight, and sequence length values.

The instability index for these proteins ranges from 0 to 100, with a few reaching up to 300. This property has no correlation to any of the remainder counterparts.

These sequences appear to be arranged into two major density curves in relation to their pI. These density curves coincide with the pH of 5 and 10. A similar result was reported by Schwartz et al [39] on a dataset of predicted ORFs from complete genomes of organisms issuing from Bacteria, Archaea, and Eukarya. These authors concluded that the bimodality shown by the pI distributions for Bacteria and Archaea protein sequence data are likely associated with different protein subcellular localizations; presumably cytosolic (pI $\approx$ 5) and integral membrane proteins (pI $\approx$ 9) [39]. Moreover, there is a smaller density peak between the pI's of 5 and 10, near the the value of 8 (Figure 2.13). Additionally, the pI tends to a value of 5 for increasing molecular weight, and sequence length values.

Unsurprisingly, the molecular weight and the sequence length have a near perfect positive correlation ($\rho = 0.99$, p-value $= 0.0$). This is also shown by the near identical distributions between these two properties and the others. Most of these proteins have a molecular weight of less than 250kDa, and a length of less than 2,500 amino-acids.

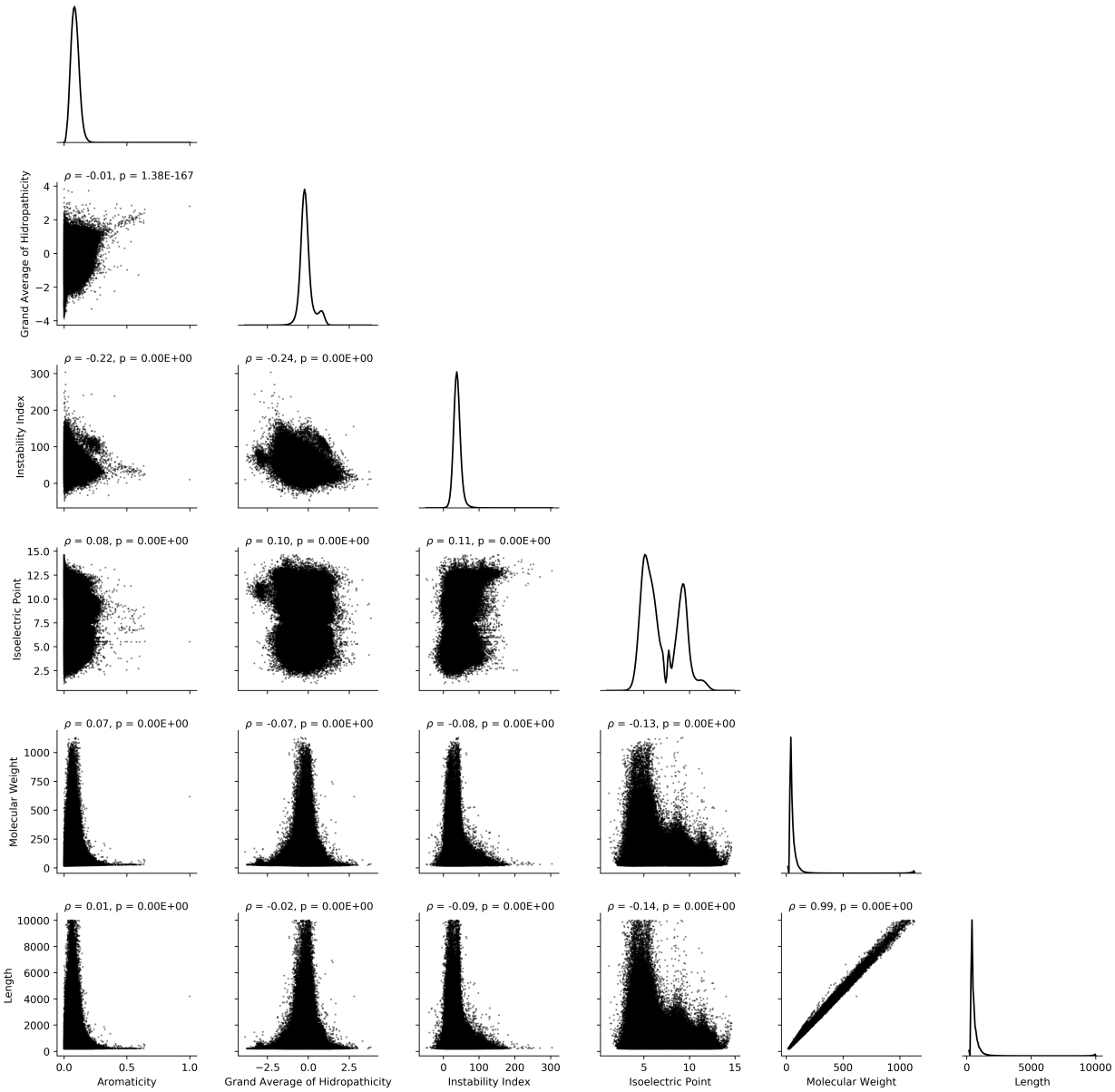


Figure 2.13: Distributions of physico-chemical properties generated *in silico* for the proteins from the Hypotheticals dataset. Each property is plotted against one-another in a pairwise arrangement. The value for Spearman's rank correlation coefficient (ρ), and the associated p-value (p) are shown for each correlation. The diagonal kernel-density estimation shows the distribution of a given property, and the y-axis represents the value density (not shown).

2.3.7 Over Representation Analysis of semantic terms

We wondered if our representative proteins had over-represented terms in comparison to a golden standard dataset. To answer this question we proceeded as follows. We downloaded the SwissProt knowledgebase [20]. Then we parsed the proteins in SwissProt and only retrieved those of Archaeal or Bacterial origin. This amounted to a total of 235,544 proteins. We presumed that the proteins from SwissProt already had GO, EC and domain annotations, as a

product of UniProt's internal annotation tools. Nonetheless, we re-annotated them using the same tools we used for the proteins from the Hypotheticals dataset—i.e., DeepGO, DeepEC, and RPS-BLAST. We did this because otherwise we could not compare these datasets without introducing biases that could compromise the interpretation of the results.

Afterwards, for each dataset, and for each protein, we gathered (i) the GO term(s) at maximum depth of its sub-DAG; (ii) the 4th-digit EC number(s); and (iii) the most informative CD(s). Next, we performed an Over Representation Analysis (ORA) on these terms, between the two datasets, and for each annotation type. We also applied a *post hoc* Benjamini-Hochberg (BH) correction [35] to all p-values provided by the hypothesis test, because this analysis constituted a multiple-testing scenario.

To ensure stringency and statistical significance, we only present the results for terms whose: (i) "Rich factor" $\geqslant 0.95$; (ii) BH-corrected p-value $\leqslant 1e-5$; and (iii) with at least one SwissProt representative (i.e., $n-k>0$).

The two over-represented GO terms with the largest number of proteins are "binding", and "protein binding" (Figure 2.14.A). Although we only considered the GO term(s) at maximum depth of the sub-DAG for a given protein, these two terms are located at a depth level of 1, and 2 respectively. This indicates that the most specific GO term for these proteins is highly unspecific. This is not surprising, as we had shown that only a small fraction of proteins possess a GO term annotation deeper than the 3rd level of the GO DAG (see Figure 2.8).

A

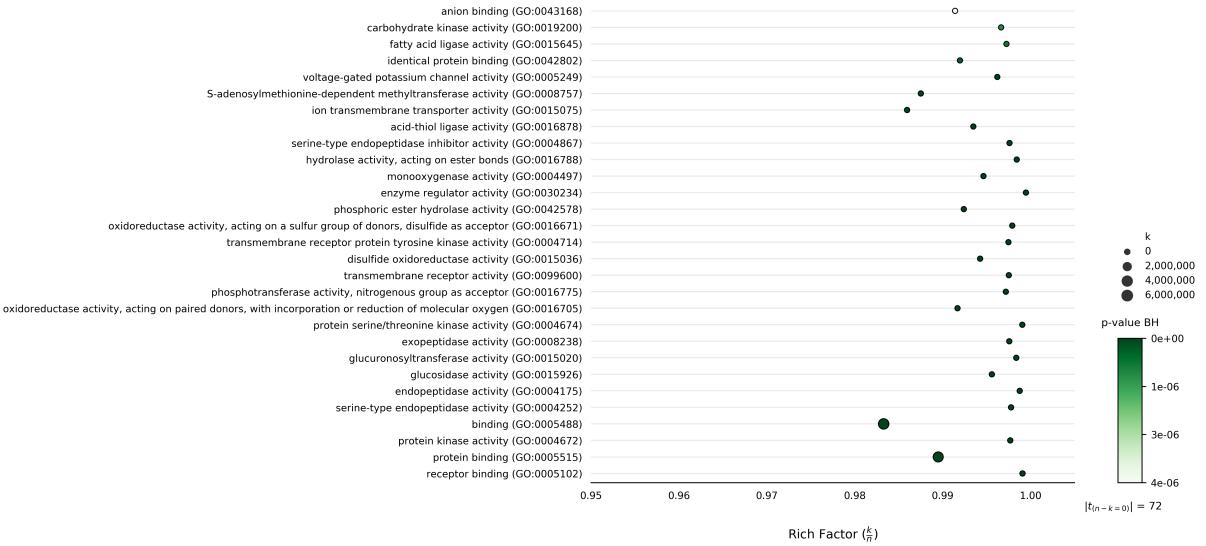


Figure 2.14. (Continued on next page.)

B

Figure 2.14. (Continued on next page.)

There are several over-represented GO terms associated with transferase activity. There are also those related with a type of phosphotransferases, the kinases. Additionally, there are terms concerning hydrolytic functions in a broader sense, as well as more specific ones, namely: “exopeptidase activity”; “endopeptidase activity”; “glucosidase activity”; and “serine-type endopeptidase activity”. The last term is also accompanied by its negative-regulator counterpart “serine-type endopeptidase inhibitor activity”.

C

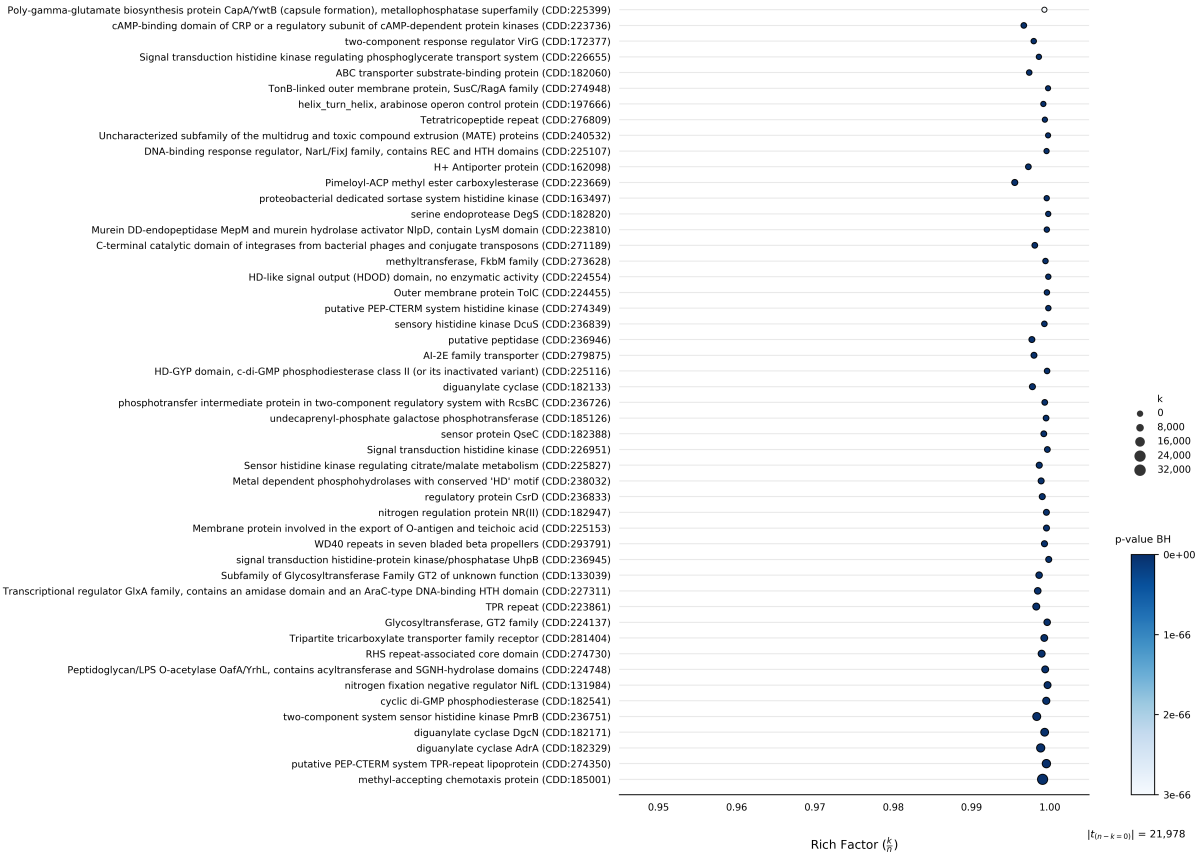


Figure 2.14: Most significant ORA results per annotation type: GO (**A**), EC (**B**), and CD (**C**). Each marker is colored according to its p-value. The size of a marker is proportional to the number of proteins that have that annotation. Each colorbar and "size-bar" has a different scale, depending on the annotation type. Rich factor: number of proteins from the Hypotheticals dataset annotated with t (k), divided by the total number of proteins annotated with t (n); p-value BH: Benjamini-Hochberg-corrected p-value; $|t_{(n-k=0)}|$: number of statistically significant (p-value $\leqslant 1e-5$), over-represented terms from the Hypotheticals dataset, without counterparts in the SwissProt dataset (i.e., Rich Factor = 1).

The remaining over-represented GO terms belong to four major groups. Those whose common ancestor is that of "oxidoreductase activity"; those associated with transmembrane transport; terms whose common ancestor is the term "binding"; and two terms related to ligase activity. There are also 72 statistically significant and over-represented GO terms from our dataset, whose counterparts are absent from SwissProt (denoted as $|t_{(n-k=0)}|$).

There are 90 over-represented 4th-digit ECs (Figure 2.14.B). From these, 50 refer to Hydrolases; 24 to Transferases; 12 to Oxidoreductases; 2 to Lyases; and 2 to Ligases. There are no over-represented Isomerases.

The most abundant over-represented EC class is that of Hydrolases. Out of these 50 ECs, 24 belong to Glycosylases (EC:3.2.-.-). From the Glycosylase class, the 4th-digit ECs with the

highest number of proteins annotated to, are Alpha-, Beta-, Glucan 1,4-alpha-, and Glucan endo-1,3-beta-D- Glucosidases (EC:3.2.1.20, 21, 3, and 39, respectively); Cellulase (EC:3.2.1.4); Endo-1,4-beta-xylanase (EC:3.2.1.8); Beta-galactosidase (EC:3.2.1.23); and mannosyl-glycoprotein endo-beta-N-acetylglucosaminidase (EC:3.2.1.96).

Other interesting over-represented 4th-digit ECs from the Glycosylase class include: Chitinase (EC:3.2.1.14), Lysozyme (EC:3.2.1.17), Exo-alpha-sialidase (EC:3.2.1.18), Levanase (EC:3.2.1.65), Licheninase (EC:3.2.1.73), and rRNA N-glycosylase (EC:3.2.2.22).

The second most diverse sub-class of Hydrolases is that of those acting on ester bonds (EC:3.1.-.-), from which 12 4th-digit ECs are present. Those with the highest number of proteins annotated to are: Triacylglycerol lipase (EC:3.1.1.3), Type II site-specific deoxyribonuclease (EC:3.1.21.4), Ribonuclease P (EC:3.1.26.5), and Protein-serine/threonine phosphatase (EC:3.1.3.16). Additionally, there are 222 over-represented and statistically-significant 4th-digit ECs whose representative counterparts are absent from SwissProt.

We only show the top 50 over-represented CDs (Figure 2.14.C) because there are a total of 1,541 CDs that met our selection criteria. This number is 21,978 for the CDs present in our database but lacking from SwissProt (i.e., $n - k = 0$, and Rich factor = 1).

There is a prevalence of signal transduction mechanisms, and regulatory systems among the putative functions for these CDs. The second foremost group of putative functions is that of cell wall, membrane, and envelope-related activity and metabolism. Other noteworthy functions include: catalytic domains of integrases from phages and transposons; structural function; carbohydrate transport and metabolism; and chemotaxis activity.

2.3.8 Taxonomic Representation

We wanted to understand the taxonomic provenance of these proteins of unknown function. To this end, we first gathered the taxonomic metadata from all 134,894,520 initial proteins. Next, we grouped this data by the domain and phyla of the organism from which each protein issued. We then calculated the relative abundance of each phylum within each domain (Figure 2.15).

The majority (75%) of proteins of unknown function from Archaea belong to the phylum Euryarchaeota (Figure 2.15.A), whereas only 12% are from either Candidate phyla, or unknown phyla. This suggests that most Archaeal proteins of unknown function issue from known phyla.

This distinction is more evident in Bacteria (Figure 2.15.B), where 90% of these proteins of unknown function issue from one of the four major phyla—i.e., Proteobacteria, Firmicutes, Actinobacteria and Bacteroidetes—and only 2.8% issue from either Candidate phyla, or unknown phyla.

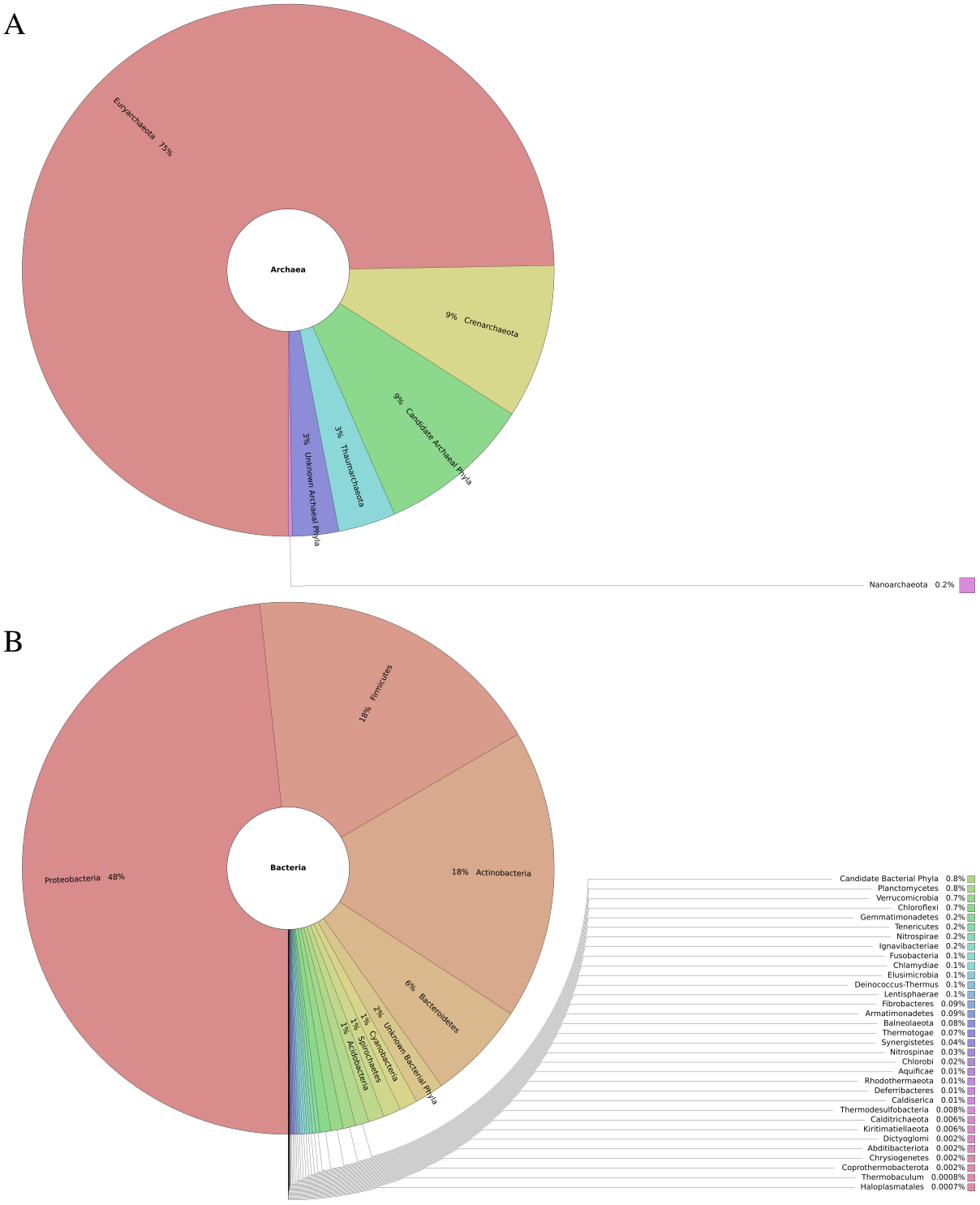


Figure 2.15: Taxonomic provenance of 134,894,520 proteins of unknown function. These figures were grouped by domain: (**A**) Archaea, and (**B**) Bacteria. The percentages represent the relative abundance of each phylum per domain. These plots were generated by Krona [40].

2.4 Conclusions

Apart from quantifying the amount of protein sequences of unknown function, we created a repository containing $>$ 134 million protein sequences whose molecular function could not be inferred by public databases. The vast majority of these proteins issue from known phyla, regardless of Domain. To our knowledge, this constitutes the first worldwide centralized collection of Functional Dark Matter of prokaryotic origin. These sequences may be further scrutinized by the scientific community, upon the public release of this data in a near future.

We have clustered our repository throughout 8 global sequence identity thresholds, akin to previous work [22, 36, 37]. The first clustering step provided a 79.51% reduction in size. The last clustering step allowed us to achieve an additional 54.2% size reduction from the non-redundant set; implying a $\sim$90.62% size reduction relative to the initial number of $>$ 134 million sequences. These results have led us to conclude two things. First, that it is possible to render our repository by a non-redundant set of representative proteins. Second, that there might be a high level of redundancy among (or within) independent data-sources, for protein sequences of unknown function. We also noticed that the majority of representative sequences from each clustering run were singletons, regardless of threshold. Moreover, most clusters comprise $<$ 100 members, also irrespective of threshold. These results have lead us to conclude that these uncharacterized sequences are very dissimilar to one another.

We have established a dataset containing 12,654,843 protein sequence representatives issuing from clusters at a 60% global sequence identity resolution. We managed to annotate 99.97% (12,651,624) of this dataset with at least one term, from at least one classification system (i.e., EC, GO or CD). This was a significant development, given that we started this endeavor with unannotated sequence data in the first place. Only a minute fraction of proteins (0.03%) was unamenable to annotation by any of the classification systems. Moreover, 99.91% (12,644,075) of this dataset was annotated with at least one GO term. The two most over-represented GO terms were “binding”, and “protein binding”. There were also over-represented GO terms associated with transferase activity, and those concerning hydrolytic functions in a broader sense.

We were able to assign at least one CD to 48.53% (6,142,383) of the foregoing sequences. Although this coverage fell short of our expectations, we posit that it might have been the product of a bit-score cutoff that was too stringent. It is also possible that the sequences lacking CD annotation might not be proteins at all. Another hypothesis is that these sequences cor-

respond to false positives from the ORF-calling step, presumably undertaken upstream from our work. There were a total of 1,541 over-represented CDs that met our selection criteria. From the 50 most represented, we highlight several putative activities such as: signal transduction mechanisms; regulatory systems; cell wall; membrane and envelope-related activity and metabolism; catalytic domains of integrases from phages and transposons; carbohydrate transport and metabolism; and chemotaxis activity.

DeepEC assigned at least one 4th-digit EC to 2.78% (351,917) of these proteins, thus suggesting that they are potentially enzymes. The most abundant enzymatic classes were Transferases (182,797) and Hydrolases (100,475). We emphasize that 304,293 (86%) of these putative enzymes also have GO and CD annotation. We found 90 over-represented 4-digit ECs, from which 50 refer to Hydrolases. We also found 9,622 putative enzymes that had more than one predicted 4th-digit EC. Most of these enzymes have EC numbers that share the first 3 digits (8,387); or the first 2 (365). We hypothesize that these are catalytically promiscuous enzymes. The remaining putative enzymes have multiple ECs that either share the 1st-digit (121), or span multiple enzymatic groups altogether (749). These might be enzymes that perform multiple functions (i.e., "moonlighting" enzymes). In theory, these might have attained multiple active sites via gene fusion events. Additional work is necessary to determine the promiscuity, or multifunctionality, of these enzymes. Addressing this particular subset of enzymes is of great significance for future biotechnological developments, as it might offer innovative solutions for the enzymes industry. By making use of enzymes capable of catalysing several key reactions in a given industrial process, the need for multiple enzymes, and thus multiple physico-chemical requirements, might be disregarded altogether.

This work has offered a small glimpse into the potential functions of these elusive proteins. We also tentatively suggest that true *in silico* protein function prediction might be achievable, provided further efforts are put forward by forthcoming researchers into this fascinating field of research.

Supplementary Information

Table S2.1: Number of protein sequences with multiple EC annotations sharing the 3rd-digit EC. "No.": number of proteins; "%": percentage of the total number of proteins in this table.

EC number	Description	No.	%
EC:3.2.1.-	Glycosidases	3,062	36.51%
EC:3.6.3.-	Hydrolases, acting on acid anhydrides; catalyzing transmembrane movement of substances	1,276	15.21%
EC:3.6.4.-	Hydrolases, acting on ATP; involved in cellular and subcellular movement	1,272	15.17%
EC:2.1.1.-	Methyltransferases	812	9.68%
EC:2.7.11.-	Protein-serine/threonine kinases	294	3.51%
EC:6.2.1.-	Acid–thiol ligases	244	2.91%
EC:2.7.7.-	Nucleotidyltransferases	222	2.65%
EC:1.1.1.-	Oxidoreductases, acting on the CH-OH group of donors, with NAD(+) or NADP(+) as acceptor	192	2.29%
EC:3.4.11.-	Aminopeptidases	141	1.68%
EC:3.1.1.-	Carboxylic ester hydrolases	107	1.28%
EC:3.1.3.-	Phosphoric monoester hydrolases	93	1.11%
EC:3.4.21.-	Serine endopeptidases	87	1.04%
EC:3.5.4.-	Hydrolases, acting on carbon-nitrogen bonds, in cyclic amidines	77	0.92%
EC:2.3.1.-	Acyltransferases, transferring groups other than amino-acyl groups	67	0.8%
EC:6.3.2.-	Acid–amino-acid ligases (peptide synthases)	57	0.68%
EC:2.7.1.-	Phosphotransferases with an alcohol group as acceptor	44	0.52%
EC:1.2.1.-	Oxidoreductases, acting on the aldehyde or oxo group of donors, with NAD(+) or NADP(+) as acceptor	30	0.36%
EC:1.6.5.-	Oxidoreductases, acting on NADH or NADPH, with a quinone or similar compound as acceptor	28	0.33%

Continued on next page

Table S2.1 – Continued from previous page

EC number	Description	No.	%
EC:2.5.1.-	Transferases, transferring alkyl or aryl groups, transferring alkyl or aryl groups	27	0.32%
EC:5.4.2.-	Phosphotransferases (phosphomutases)	26	0.31%
EC:2.6.1.-	Transaminases	24	0.29%
EC:2.4.1.-	Hexosyltransferases	18	0.21%
EC:5.1.3.-	Isomerases, racemases and epimerases, acting on carbohydrates and derivatives	13	0.16%
EC:1.14.13.-	Oxidoreductases, acting on paired donors, with NADH or NADPH as one donor	13	0.16%
EC:2.8.4.-	Transferases, transferring sulfur-containing groups, transferring alkylthio groups	12	0.14%
EC:1.18.1.-	Oxidoreductases, acting on iron-sulfur proteins as donors, with NAD(+) or NADP(+) as acceptor	11	0.13%
EC:2.8.1.-	Sulfurtransferases	10	0.12%
EC:2.7.4.-	Phosphotransferases with a phosphate group as acceptor	10	0.12%
EC:2.7.8.-	Transferases for other substituted phosphate groups	6	0.07%
EC:2.3.3.-	Acyltransferases, acyl groups converted into alkyl groups on transfer	6	0.07%
EC:4.2.1.-	Hydro-lyases	5	0.06%
EC:3.1.21.-	Endodeoxyribonucleases producing 5'-phosphomonoesters	5	0.06%
EC:4.99.1.-	Other lyases	5	0.06%
EC:3.5.5.-	Hydrolases, acting on carbon-nitrogen bonds, in nitriles	5	0.06%
EC:1.3.8.-	Oxidoreductases, acting on the CH-CH group of donors, with a flavin as acceptor	4	0.05%
EC:5.4.99.-	Intramolecular transferases, transferring other groups	4	0.05%
EC:4.4.1.-	Carbon-sulfur lyases	4	0.05%
EC:4.3.1.-	Ammonia-lyases	4	0.05%
EC:1.11.1.-	Peroxidases	4	0.05%

Continued on next page

Table S2.1 – Continued from previous page

EC number	Description	No.	%
EC:1.8.4.-	Oxidoreductases, acting on a sulfur group of donors, with a disulfide as acceptor	4	0.05%
EC:3.3.2.-	Ether hydrolases	4	0.05%
EC:1.13.11.-	Oxidoreductases, acting on single donors with incorporation of molecular oxygen (oxygenases). The oxygen incorporated need not be derived from O(2), with incorporation of two atoms of oxygen	4	0.05%
EC:5.3.1.-	Intramolecular oxidoreductases, interconverting aldoses and ketoses	3	0.04%
EC:3.5.1.-	Hydrolases, acting on carbon-nitrogen bonds, in linear amides	3	0.04%
EC:4.1.2.-	Aldehyde-lyases	3	0.04%
EC:4.2.2.-	Carbon-oxygen lyases, acting on polysaccharides	3	0.04%
EC:3.6.1.-	Hydrolases, acting on acid anhydrides, in phosphorus-containing anhydrides	3	0.04%
EC:4.6.1.-	Phosphorus-oxygen lyases	3	0.04%
EC:3.4.14.-	Dipeptidyl-peptidases and tripeptidyl-peptidases	3	0.04%
EC:3.1.4.-	Phosphoric diester hydrolases	3	0.04%
EC:1.3.1.-	Oxidoreductases, acting on the CH-CH group of donors, with NAD(+) or NADP(+) as acceptor	2	0.02%
EC:3.5.3.-	Hydrolases, acting on carbon-nitrogen bonds, in linear amidines	2	0.02%
EC:3.5.2.-	Hydrolases, acting on carbon-nitrogen bonds, in cyclic amides	2	0.02%
EC:1.2.4.-	Oxidoreductases, acting on the aldehyde or oxo group of donors, with a disulfide as acceptor	2	0.02%
EC:3.4.24.-	Metalloendopeptidases	2	0.02%
EC:1.8.1.-	Oxidoreductases, acting on a sulfur group of donors, with NAD(+) or NADP(+) as acceptor	2	0.02%

Continued on next page

Table S2.1 – Continued from previous page

EC number	Description	No.	%
EC:2.4.2.-	Pentosyltransferases	2	0.02%
EC:2.7.10.-	Protein-tyrosine kinases	2	0.02%
EC:6.3.5.-	Carbon–nitrogen ligases with glutamine as amido-N-donor	1	0.01%
EC:6.1.1.-	Ligases forming aminoacyl-tRNA and related compounds	1	0.01%
EC:5.99.1.-	Other isomerases	1	0.01%
EC:3.1.26.-	Endoribonucleases producing 5'-phosphomonoesters	1	0.01%
EC:4.1.99.-	Other carbon-carbon lyases	1	0.01%
EC:4.1.1.-	Carboxy-lyases	1	0.01%
EC:2.8.3.-	CoA-transferases	1	0.01%
EC:2.4.99.-	Glycosyltransferases, transferring other glycosyl groups	1	0.01%
EC:2.3.2.-	Aminoacyltransferases	1	0.01%
EC:2.2.1.-	Transketolases and transaldolases	1	0.01%
EC:1.4.3.-	Oxidoreductases, acting on the CH-NH(2) group of donors, with oxygen as acceptor	1	0.01%
EC:1.3.5.-	Oxidoreductases, acting on the CH-CH group of donors, with a quinone or related compound as acceptor	1	0.01%
EC:1.2.7.-	Oxidoreductases, acting on the aldehyde or oxo group of donors, with an iron-sulfur protein as acceptor	1	0.01%
EC:6.4.1.-	Ligases, forming carbon-carbon bonds	1	0.01%

Table S2.2: Number of protein sequences with multiple EC annotations sharing the 2nd-digit EC. "No.": number of proteins; "%": percentage of the total number of proteins in this table.

EC number	Description	No.	%
EC:2.7.-.-	Transferases, transferring phosphorus-containing groups	344	94.25%
EC:4.1.-.-	Carbon-carbon lyases	9	2.47%
EC:3.4.-.-	Hydrolases, acting on peptide bonds (peptidases)	4	1.1%
EC:3.1.-.-	Hydrolases, acting on ester bonds	3	0.82%
EC:3.6.-.-	Hydrolases, acting on acid anhydrides	3	0.82%
EC:3.2.-.-	Glycosylases	2	0.55%

Table S2.3: Number of protein sequences with multiple EC annotations sharing the 1st-digit EC. "No.": number of proteins; "%": percentage of the total number of proteins in this table.

EC number	Description	No.	%
EC:2.-.-.-	Transferases	61	50.41%
EC:3.-.-.-	Hydrolases	53	43.8%
EC:1.-.-.-	Oxidoreductases	6	4.96%
EC:4.-.-.-	Lyases	1	0.83%

Table S2.4: Number of protein sequences with multiple EC annotations that do not share an EC digit. "No.": number of proteins; "%": percentage of the total number of proteins in this table.

EC number	Description	No.	%
EC:3.-.-.- + EC:4.-.-.-	Hydrolases + Lyases	207	27.64%
EC:2.-.-.- + EC:3.-.-.-	Transferases + Hydrolases	174	23.23%
EC:1.-.-.- + EC:3.-.-.-	Oxidoreductases + Hydrolases	123	16.42%
EC:4.-.-.- + EC:5.-.-.-	Lyases + Isomerases	76	10.15%
EC:2.-.-.- + EC:4.-.-.-	Transferases + Lyases	36	4.81%
EC:2.-.-.- + EC:6.-.-.-	Transferases + Ligases	30	4.01%
EC:1.-.-.- + EC:4.-.-.-	Oxidoreductases + Lyases	27	3.6%
EC:3.-.-.- + EC:6.-.-.-	Hydrolases + Ligases	22	2.94%
EC:2.-.-.- + EC:5.-.-.-	Transferases + Isomerases	13	1.74%
EC:1.-.-.- + EC:2.-.-.-	Oxidoreductases + Transferases	11	1.47%
EC:1.-.-.- + EC:2.-.-.- + EC:4.-.-.-	Oxidoreductases + Transferases + Lyases	9	1.2%
EC:1.-.-.- + EC:4.-.-.- + EC:5.-.-.-	Oxidoreductases + Lyases + Isomerases	8	1.07%
EC:1.-.-.- + EC:5.-.-.-	Oxidoreductases + Isomerases	7	0.93%
EC:3.-.-.- + EC:5.-.-.-	Hydrolases + Isomerases	4	0.53%
EC:4.-.-.- + EC:6.-.-.-	Lyases + Ligases	1	0.13%
EC:5.-.-.- + EC:6.-.-.-	Isomerases + Ligases	1	0.13%

References

[1] Stephens ZD, Lee SY, Faghri F, Campbell RH, Zhai C, Efron MJ, et al. Big Data: Astronomical or Genomical? PLoS Biol. 2015;13(7):e1002195.

[2] Piao H, Froula J, Du C, Kim TW, Hawley ER, Bauer S, et al. Identification of novel biomass-degrading enzymes from genomic dark matter: Populating genomic sequence space with functional annotation. Biotechnol Bioeng. 2014;111(8):1550–1565.

[3] Hu P, Janga SC, Babu M, Díaz-Mejía JJ, Butland G, Yang W, et al. Global functional atlas of Escherichia coli encompassing previously uncharacterized proteins. PLoS Biol. 2009;7(4):e96.

[4] Hutchison CA 3rd, Chuang RY, Noskov VN, Assad-Garcia N, Deerinck TJ, Ellisman MH, et al. Design and synthesis of a minimal bacterial genome. Science. 2016;351(6280):aad6253.

[5] Al-Shahib A, Breitling R, Gilbert DR. Predicting protein function by machine learning on amino acid sequences – a critical evaluation. BMC Genomics. 2007;8(1):78.

[6] McLean JS, Lombardo MJ, Badger JH, Edlund A, Novotny M, Yee-Greenbaum J, et al. Candidate phylum TM6 genome recovered from a hospital sink biofilm provides genomic insights into this uncultivated phylum. Proc Natl Acad Sci U S A. 2013;110(26):E2390–9.

[7] Marcy Y, Ouverney C, Bik EM, Lösekann T, Ivanova N, Martin HG, et al. Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. Proc Natl Acad Sci U S A. 2007;104(29):11889–11894.

[8] Garza DR, Dutilh BE. From cultured to uncultured genome sequences: metagenomics and modeling microbial ecosystems. Cell Mol Life Sci. 2015;72(22):4287–4308.

[9] Becraft ED, Dodsworth JA, Murugapiran SK, Ohlsson JI, Briggs BR, Kanbar J, et al. Single-Cell-Genomics-Facilitated Read Binning of Candidate Phylum EM19 Genomes from Geothermal Spring Metagenomes. Appl Environ Microbiol. 2015;82(4):992–1003.

[10] Bernard G, Pathmanathan JS, Lannes R, Lopez P, Bapteste E. Microbial dark matter investigations: how microbial studies transform biological knowledge and empirically sketch a logic of scientific discovery. Genome Biol Evol. 2018;.

[11] Lobb B, Kurtz DA, Moreno-Hagelsieb G, Doxey AC. Remote homology and the functions of metagenomic dark matter. Front Genet. 2015;6:234.

[12] Mokili JL, Rohwer F, Dutilh BE. Metagenomics and future perspectives in virus discovery. Curr Opin Virol. 2012;2(1):63–77.

[13] Dutilh BE. Metagenomic ventures into outer sequence space. Bacteriophage. 2014;4(4):e979664.

[14] Erdin S, Lisewski AM, Lichtarge O. Protein function prediction: towards integration of similarity metrics. Curr Opin Struct Biol. 2011;21(2):180–188.

[15] Rashid M, Stingl U. Contemporary molecular tools in microbial ecology and their application to advancing biotechnology. Biotechnol Adv. 2015;33(8):1755–1773.

[16] O'Leary NA, Wright MW, Brister JR, Ciufo S, Haddad D, McVeigh R, et al. Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. Nucleic Acids Res. 2016;44(D1):D733–45.

[17] Clark K, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW. GenBank. Nucleic Acids Res. 2016;44(D1):D67–72.

[18] Wattam AR, Davis JJ, Assaf R, Boisvert S, Brettin T, Bun C, et al. Improvements to PATRIC, the all-bacterial Bioinformatics Database and Analysis Resource Center. Nucleic Acids Res. 2017;45(D1):D535–D542.

[19] Markowitz VM, Chen IMA, Palaniappan K, Chu K, Szeto E, Grechkin Y, et al. The integrated microbial genomes system: an expanding comparative analysis resource. Nucleic Acids Res. 2010;38(Database issue):D382–90.

[20] Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 2003;31(1):365–370.

[21] Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28(23):3150–3152.

[22] Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics. 2007;23(10):1282–1288.

[23] Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25(11):1422–1423.

[24] Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol. 1982;157(1):105–132.

[25] Lobry JR, Gautier C. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 Escherichia coli chromosome-encoded genes. Nucleic Acids Res. 1994;22(15):3174–3180.

[26] Guruprasad K, Reddy BV, Pandit MW. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. Protein Eng. 1990;4(2):155–161.

[27] Kulmanov M, Khan MA, Hoehndorf R, Wren J. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. Bioinformatics. 2018;34(4):660–668.

[28] Ryu JY, Kim HU, Lee SY. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. Proc Natl Acad Sci U S A. 2019;116(28):13996–14001.

[29] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–3402.

[30] Marchler-Bauer A, Bo Y, Han L, He J, Lanczycki CJ, Lu S, et al. CDD/SPARCLE: functional classification of proteins via subfamily domain architectures. Nucleic Acids Res. 2017;45(D1):D200–D203.

[31] Marchler-Bauer A, Bryant SH. CD-Search: protein domain annotations on the fly. Nucleic Acids Res. 2004;32(Web Server issue):W327–31.

[32] Fassler J, Cooper P. BLAST Glossary. In: BLAST® Help [Internet]. National Center for Biotechnology Information (US); 2011.

[33] Pearson WR. An introduction to sequence similarity (“homology”) searching. Curr Protoc Bioinformatics. 2013;Chapter 3:Unit3.1.

[34] Rivals I, Personnaz L, Taing L, Potier MC. Enrichment or depletion of a GO category within a class of genes: which test?; 2007.

[35] Benjamini Y, Hochberg Y. Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing; 1995.

[36] Li W, Wooley JC, Godzik A. Probing metagenomics by rapid cluster analysis of very large datasets. PLoS One. 2008;3(10):e3375.

[37] Li W. Analysis and comparison of very large metagenomes with fast clustering and functional annotation. BMC Bioinformatics. 2009;10:359.

[38] Jeffery CJ. Protein moonlighting: what is it, and why is it important? Philos Trans R Soc Lond B Biol Sci. 2018;373(1738).

[39] Schwartz R, Ting CS, King J. Whole proteome pI values correlate with subcellular localizations of proteins for organisms within the three domains of life. Genome Res. 2001;11(5):703–709.

[40] Ondov BD, Bergman NH, Phillippy AM. Interactive metagenomic visualization in a Web browser. BMC Bioinformatics. 2011;12:385.

Chapter 3

An Information-Theoretic Approach to Systemize Functional Dark Matter

This chapter partially transcribes the contents of the following manuscript:

> **Escudeiro P, Couto FM, Henry CS & Dias R** (2022). *An Information-Theoretic Approach to Systemize Biological Sequence Data Annotation* (in preparation).

Abstract

Most sequence databases contain large amounts of data requiring further analysis. The gene products imputed from those data are often left unannotated, and their putative molecular functions undisclosed. The ever-increasing growth in genomic data acquisition has magnified this predicament, leaving a massive amount of sequences of unknown function in its wake.

The sheer magnitude of these sequences has prevented their experimental characterization. This situation has left no alternative but to predict the function of these sequences through computational methods. However, the informational value conveyed by these annotations varies substantially. In part, this is due to the very nature of the ontologies that comprise the annotation terms, like the Gene Ontology (GO). Therefore we posit that organizing these uncharacterized sequences according to the extent to which they are annotated should be imperative. A considerable amount of work can be averted if we manage to identify which protein sequences were pliant to effective *in silico* annotation.

In this work we develop a new ensemble of metrics that allow to compute the information content of an annotated protein. The measures we present allow to numerically qualify a protein according to its annotations. We take into account three distinct Classification Systems: GO, Enzyme Commission (EC) numbers, and Conserved Domains (CD). We also show that a sequence dataset can be represented as a collection of distributions depicting spectra of protein annotation. Our results show that these distributions provide insight into the informational content of each individual protein according to the annotations it possesses. We also created an index that illustrates the extent to which a protein was annotated in relation to all three Classification Systems. To our knowledge, this metric is the first attempt at creating a unifying measure for protein annotation content and quality.

3.1 Introduction

Public sequence databases possess a tremendous amount of uncharacterized sequence data. Most gene products predicted from these sequences are usually classified as "uncharacterized", "putative", "predicted", "hypothetical", or "unknown" [1]. These sequences are often referred to as "genomic dark matter" or "functional dark matter". The rate at which these sequences are accumulating is distressing. In addition, constraints such as time and resources preclude their experimental characterization. The foregoing reasons leave no choice but to annotate these sequences via computational methods.

Among the core contributions to molecular biology, bioinformatics has introduced ontologies as a means to annotate gene products [2]. Ontologies are useful because they bypass the shortcomings of natural language descriptions [2]. By doing so, ontologies allow to automate not only the annotation process, but the reasoning over these annotations as well [3].

Perhaps the most widely known ontology within the biological sciences is the Gene Ontology (GO). The GO allows researchers to describe the function of a gene product through annotations (named GO terms) in the form of a controlled vocabulary [4]. Moreover, the use of ontologies like GO also allow to compare gene products based on their annotation [2]. These features allow for both detailed and large-scale sequence annotation to be conducted [4].

Nonetheless, there are numerous sources of molecular function annotations (i.e., Classification Systems) besides GO [4]. These include Enzyme Commission (EC) numbers, and Conserved Domains (CD), among many others. Additional Classification Systems provide enhanced molecular function descriptions when used in combination with GO [4].

Adding to this complexity, some annotations might be more specific than others. For instance, a protein whose sole annotation is "hydrolase" conveys less information regarding its molecular function than a second protein whose annotation is "amylase". In order to quantify the amount of information conveyed by a given term, one needs to calculate its Information Content (*IC*) [5, 6].

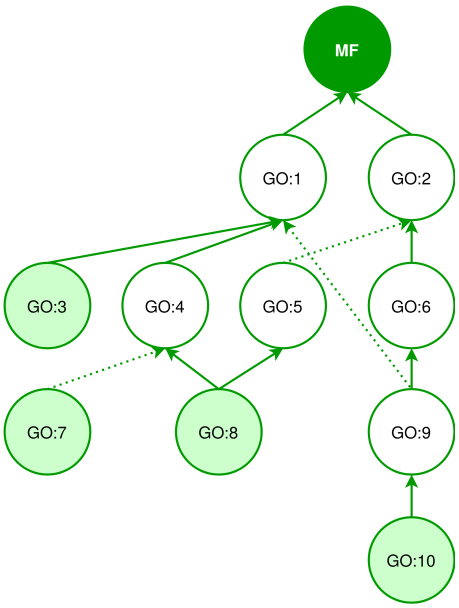
Shannon's definition states that the *IC* of a term is inversely proportional to the likelihood of its occurrence within a corpus [5, 6]. In this work, we define a corpus as a collection of sequences—i.e., a dataset. In a Classification System structured like a graph (e.g. GO, Figure 3.1.A), we have to consider the relations between terms to calculate the *IC* for each term [7]. An example of these relations are subsumption relations, like *is-a* or *part-of* [8].

We calculate the probability of a term occurring in a dataset by dividing its frequency by the frequency of the *root* term—i.e., the number of times any term occurs in the dataset [6]. Moreover, a term occurs if itself, or any of its descendants occur [6]. The *IC* of a term is given by the negative log likelihood of this probability [6], it is therefore a dataset-dependent measure [7]. By calculating the *IC* of a term we arrive at two observations. First, the greater the probability of a term occurring, the broader its representation, and the lower its *IC* value [6]. And conversely, the lower the probability of a term occurring, the greater its specificity, and the greater its *IC* value [6]. This probability increases as one moves towards the *root* of a graph, where the probability of the *root* term occurring is 1 [7].

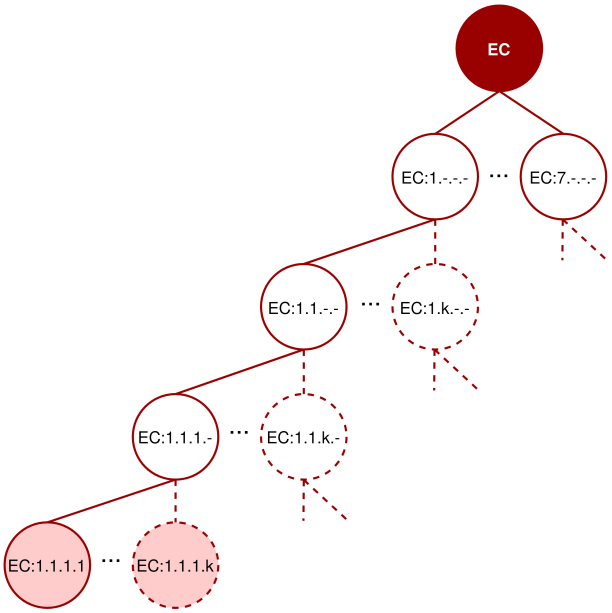
IC-based metrics have been extensively used for biological ontologies ever since the work of Lord et al. [7]. Yet, to our knowledge, these types of metrics have not been applied to EC numbers, nor CD identifiers. Despite being a numerical classification scheme for enzymes, we posit that EC numbers are also a *k*-ary tree. A *k*-ary tree is a rooted tree where each node has at most *k* children. Where *k* is the number of descendents of the parent node with the most offspring. All of its leaf nodes are at the same depth of the graph—i.e. the 4th-digit EC (Figure 3.1.B). This *k*-ary tree consists of four levels. The first level has seven nodes, where each node describes the enzyme class (e.g., EC:1.-.-.-, "Oxidoreductases"). The second and third levels include the enzyme subclass and sub-subclass. These levels represent the reaction a particular enzyme catalyzes in terms of several factors. These factors relate to the reactive species, the type of bond being acted upon, and group or product involved (e.g., EC:1.1.-.- , "Acting on the CH-OH group of donors"; and EC:1.1.1.-, "With NAD(+) or NADP(+) as acceptor"). The fourth level represents the overall reaction of that sub-subclass (e.g., EC:1.1.1.1, "Alcohol dehydrogenase"). Therefore, even though these subsumption relations are not strict *is-a* relations, the occurrence of a 4th-digit EC implies the occurrence of every digit it contains. Thus, we argue that these relations should be considered when calculating the *IC* for an EC number.

Classification systems might have different types of relations among its terms [8], or even the absence thereof. Such is the case of CD identifiers. These identifiers, like other Classification Systems, do not conform to a controlled vocabulary [4, 7]. We can calculate the *IC* for these terms nonetheless. By lacking relations they behave as sets of keywords, and may be considered "orphan" terms [7]. This way, every identifier has a single inheritance relation, directly and only, from a conceptual *root* term [7](Figure 3.1.C).

A



B



C

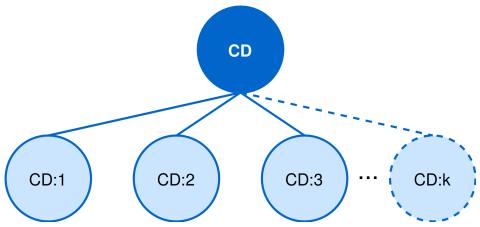


Figure 3.1: Graph structure example of each Classification System: GO (**A**), EC (**B**) and CD (**C**). The dark-color-fill nodes represent the *root* of each graph. The light-color-fill nodes represent leaves. **A**: The graph depicts a sub-DAG. The root term is the "molecular function" (MF) category of GO. Each solid line represents an *is-a* subsumption relation. The dotted lines represent a *part-of* subsumption relation. This sub-DAG structure example was adapted from [9]. **B**: EC is shown as a k-ary tree. The dots denote collapsed nodes. The dashed lines represent collapsed subsumption relationships. The dashed nodes represent the "k-most" nodes of that level, where k is the number of descendents of the parent node with the most offspring. These were collapsed in the graph for simplification purposes. **C**: The conjectural graph structure of CD PSSM identifiers. These are represented as a set of keywords. Every node is only related to a conceptual *root* term. The dots denote collapsed nodes, and the dashed node represents the "k-most" node of the only level that exists.

The calculation of *IC* for CD identifiers raises three implications. First, this reasoning implies that all CD identifiers share a single ancestor, the conceptual *root* term (Figure 3.1.C). Second, all identifiers are at the same level, which is the first and only one. Third, they have no descendants—i.e. they are all leaf nodes. Therefore, when calculating the probability of occurrence of an identifier, we cannot consider the number of times its descendants occurred, because there are none.

The *IC* of a term does not represent the full extent of information in an annotated protein. There are two main reasons for this. First, a protein might have multiple annotations. Second, these annotations might issue from multiple Classification Systems.

In this work we present a new family of *IC*-based measures, tailored for the computation of the informational value within an annotated protein. Our computations take into account a set of multiple terms from a given Classification System, as well as whether the Classification System in question conforms to a controlled vocabulary or not. This way, each metric can be calculated for the three Classification Systems used in the present work (i.e., EC, GO, and CD).

Ultimately we combine each measure calculated for a Classification System into an index. This index illustrates the extent to which a protein was annotated in relation to all three Classification Systems. To our knowledge, this metric is the first attempt at creating a unifying measure for protein annotation content and quality.

3.2 Methods

3.2.1 Datasets

The "Hypotheticals" dataset refers to 12,651,618 protein sequences. The "SwissProt" dataset refers to 235,543 protein sequences. Both datasets comprise sequences that were annotated with at least a GO term, an EC number, or a CD. These annotations issue from DeepGO [10], DeepEC [11], and RPS-BLAST [12], respectively. The sources from which these sequences were obtained, the selection criteria, all (pre-)processing steps that ensued, as well as the annotation processes, are described in section 2.2.

3.2.2 Rationale for the Development of Information-Theoretic Equations

We define a Classification System (C) as a Semantic-Base [8]:

$$C = \langle T, R \rangle \mid R = \{(t_i, t_j) : \forall t_i, t_j \in T\} \tag{3.1}$$

Where T is the set containing all terms t that belong to C, and R is the set of relations between these terms. A relation in R consists of a pair of terms (t_i, t_j) that belong to T. This definition of C ignores the type of relations between the terms [8]. R might contain subsumption relations (e.g. *is-a*), other types of relations, or even the absence thereof (i.e. $R = \varnothing$). Assuming that R contains subsumption relations, we define the ancestors (Anc) of a term t in T [8], as:

$$Anc(t) = \{a : (t, a) \in R^{+} \wedge\ t \in T\} \tag{3.2}$$

And the descendants ($Desc$) of a term t in T [8], as:

$$Desc(t) = \{d : (d, t) \in R^{+} \wedge\ t \in T\} \tag{3.3}$$

Where R^{+} is the transitive closure of R on the set T [8, 13]. Consider a corpus X to be a collection of protein sequences. Each protein sequence x in X can be annotated with a finite number of distinct terms t in T. Let ref be a predicate that evaluates to true when a protein x in a corpus X refers a term t in T [8]. This way, the frequency (f) of a term t in a corpus X is defined as:

$$f_X(t) = |\{x \in X : ref(x,t') \ \wedge \ t' \in \{t\} \cup Desc(t)\}| \tag{3.4}$$

Where t' comprises the term t and all of its descendants. Thus, the probability function of a term t occurring in a corpus X, can be written as:

$$p(t) = \frac{f_X(t)}{f_X(root)} \tag{3.5}$$

Where *root* is the root term of a C—i.e. the single common ancestor to all terms t in T. We calculate the information content (IC) of a term t in T according to Shannon [5]. Thus, the informational value of a term t is given by the negative log likelihood of its probability of occurrence within a given corpus [6]:

$$IC(t) = -\log_b p(t) \tag{3.6}$$

Where b is the base of the logarithm. In this work, all calculations involving the logarithm were made with $b = 10$. We can uniformize Equation 3.6 by dividing the value of IC by the maximum value in the scale—i.e. the frequency (f) of the *root* term in X [2]:

$$IC_u(t) = \frac{IC(t)}{\log_b f_X(root)} \tag{3.7}$$

Consider that a term t results from a computational prediction. Likewise, note that the prediction tools we used assign a score to every prediction. Therefore, assume that for every term t there is a corresponding prediction score s.

Let *annot* be a function that represents the input-output logic for a prediction tool. Given a protein x from a corpus X as input, annot outputs a set that contains pairs (t,s), as follows:

$$annot : x \mapsto \{(t_1,s_1),...,(t_n,s_n)\} \mid t \in T \wedge (s \in]0,1] \vee s \in \mathbb{R}_{\geqslant 0}) \tag{3.8}$$

Where t is a term that belongs to T in a C, and s is the prediction score for that t. In this work, the value of s might range from 0 to 1 (in the case of DeepEC or DeepGO), or belong to $\mathbb{R}_{\geqslant 0}$ (in case of bit scores from RPS-BLAST).

The semantic representation of a protein x is defined as the subset of T that characterizes a one-to-many mapping between a protein x in X, and terms t that belong to T in a C [8]:

$$T_x = \{t : (t,s) \in annot(x)\} \mid T_x \subset T \tag{3.9}$$

For each pair (t,s), T_x only contains the term t returned by *annot*, and not the prediction score s. Assume that T_x can represent the sub-graph associated with protein x. This might occur if T_x is the subset of a T in a C where R contains subsumption relations.

Let $depth$ be a function that, given a set of terms T, outputs the $depth$ value of each term t belonging to T in a C, provided that R contains subsumption relations:

$$depth : T \to \mathbb{N} \tag{3.10}$$

Thus, we determine M_x as the subset of T_x, containing the terms for which the value of $depth$ in the sub-graph is maximized:

$$M_x = \{t \in T_x : \operatorname{argmax} depth(t)\} \mid M_x \subseteq T_x \tag{3.11}$$

Notice that in the case of GO, specificity has been reported to be poorly related with depth in the graph [2]. The reason for this is that leaf nodes in the DAG differ substantially in their distance from the *root* [7]. However, we still chose the value of depth associated with a given term as a selection criterion. We do so in detriment of choosing the leaf nodes, for three reasons.

Firstly, the leaves in a sub-DAG might not correspond to leaf nodes in the complete GO DAG. Thus, they could lead to misinterpretation of downstream results. As an example, if a protein is annotated with the terms “binding” and “protein binding”, the only leaf node of the sub-DAG characterizing this protein is the term “protein binding”. Yet, this term is far from being a leaf node in the complete GO DAG.

Secondly, a leaf node in a sub-DAG might not be the most informative term characterizing a protein. We make this assumption minding that there might be other terms, at a greater depth, that even though not being leaves, might nonetheless possess greater informational value.

Thirdly, even if poorly related with specificity, the depth of a term is still a valid indicator of informational value. From an intrinsic graph perspective, the greater the distance from the

root of the graph, the more specific the terms [7]. In the present work we made use of the graph's structure in order to calculate the frequency of a term in a corpus (Equation 3.4), and consequently all equations branching from the latter. We regard our choice of depth over leaf nodes as reasonable.

Elaborating on the definition of ancestors (Equation 3.2), we establish the common ancestors (CA_x) of the terms m in M_x, as:

$$CA_x = \bigcap_{m \in M_x} Anc(m) \tag{3.12}$$

And the most informative common ancestors ($MICA_x$) as the subset of CA_x containing the common ancestors for which IC_u is maximized:

$$MICA_x = \{a \in CA_x : \operatorname{argmax} IC_u(a)\} \mid MICA_x \subseteq CA_x \tag{3.13}$$

According to the notation known as Iverson brackets [14], assume that a logical proposition L enclosed in square brackets is converted into 1 if satisfied, and 0 otherwise:

$$[L] = \begin{cases} 1 & \text{if } L \text{ is true;} \\ 0 & \text{otherwise.} \end{cases} \tag{3.14}$$

We define the Protein Information Content (PIC), of a set of terms T_x as the following equation:

$$PIC(T_x) = \sum_{m \in M_x} IC_u(m) - IC_u(MICA_x)[|M_x| > 1] \tag{3.15}$$

Where T_x is the semantic representation of a protein x in a corpus X; M_x is the subset of T_x that contains the terms at maximum depth of the sub-graph; $MICA_x$ is the set containing the common ancestor(s) with the maximum IC_u; and the square brackets denote Iverson brackets (Equation 3.14). Notice that, by definition, all ancestors in $MICA_x$ have the same IC_u value [8].

Considering the cardinality of M_x, and whether a $MICA_x$ exists, there can be 3 outcomes for Equation 3.15. First, if the $MICA_x$ is the root term, the PIC is equal to the sum of the IC_u values of the terms in M_x, given that $IC_u(root) = 0$. Second, if M_x contains a single term, the Iverson

brackets evaluate to 0, and the *PIC* is equal to the IC_u of that term alone, given that the sum over a set with a single element equals this element itself. Third, if M_x contains several terms, the *PIC* is equal to the sum of the IC_u values of these terms, minus the IC_u of the $MICA_x$.

Equation 3.15 can only be solved for a T_x from a C whose R contains subsumption relations—e.g., GO terms or EC numbers. The absence of semantic relations among CD PSSM identifiers preclude the use of Equation 3.15 to calculate *PIC*. This is due to the fact that, for a T_x containing CD PSSM identifiers, one cannot calculate M_x nor $MICA_x$. The reason for this is that in this C there is no concept of depth, nor that of ancestors. For these reasons, we alternatively define the *PIC* for a T_x, in a C, whose $R = \varnothing$, as:

$$PIC(T_x) = \max\{IC_u(t) : t \in T_x\} \mid R = \varnothing \tag{3.16}$$

Where the *PIC* for a semantic representation T_x, of a protein x in a corpus X, containing CD PSSM identifiers, is equal to the highest IC_u.

Elaborating on the output of Equation 3.8, we define a retrieval function S. This function returns the prediction score s calculated by *annot* (Equation 3.8), given a term t in T_x as input:

$$S : t \mapsto s \mid t \in T_x \wedge (s \in]0,1] \vee s \in \mathbb{R}_{\geqslant 0}) \tag{3.17}$$

Where T_x is the semantic representation of a protein x in a corpus X. Using the retrieval function S, we can include the prediction scores s provided by *annot* into the calculation of *PIC*. This way, we are able to associate the IC_u of a given term t in T_x, with the degree of confidence for that prediction. Thus, given a semantic representation T_x, of a protein x in a corpus X, belonging to a C whose R contains subsumption relations, we can define a prediction-score-weighted version of *PIC*, named PIC_S, as:

$$PIC_S(T_x) = \sum_{m \in M_x} (IC_u \cdot S)(m) - (IC_u \cdot S)(MICA_x)[|M_x| > 1] \tag{3.18}$$

Where S is the retrieval function defined in Equation 3.17. The product of IC_u and S follows the algebraic function multiplication, such that given the product of two arbitrary functions f and g, we write $f(x) \cdot g(x) = (f \cdot g)(x)$. Akin to Equation 3.15, the square brackets denote Iverson brackets (Equation 3.14). The same 3 outcomes apply when considering the cardinality of M_x and the existence of a $MICA_x$.

The semantic representation T_x, of a protein x in a corpus X, might issue from a C whose R contains subsumption relations. Such is the case with GO terms or EC numbers. In this work, EC numbers were predicted by DeepEC. DeepEC uses 3 convolutional neural networks (CNNs) to predict the enzymatic function of a protein sequence. The first CNN predicts whether the protein is an enzyme or not. The second CNN predicts the 3rd-digit EC number. And the third CNN predicts the 4th-digit EC number.

To mitigate false-positives, DeepEC only provides results for proteins that were predicted to be enzymes by the first CNN, and whose predicted 3rd-digit EC number is a prefix of the predicted 4th-digit one. DeepEC also performs homology analysis if one of the CNNs fails to predict an EC number. It also outputs to the user the log files corresponding to each CNN prediction, and its prediction scores. The predictions from homology analysis lack prediction scores. For that reason they were excluded from all analyses.

Conversely to DeepGO, all results from DeepEC contain EC numbers that are at the maximum depth of the graph—i.e., the 4th-digit. In accordance with DeepGO, DeepEC can assign a protein sequence with more than one term at the maximum depth. To calculate the CA_x and the $MICA_x$ for the protein sequences with multiple EC assignments, we decomposed each 3rd-digit EC number prediction and retrieved its 2nd and 1st-digit.

However, DeepEC does not provide prediction scores for the 2nd and 1st-digit EC. This is due to the fact that there are no CNNs liable to do so. To overcome this, we assumed that each 2nd and 1st-digit EC had the same prediction score s as that of the 3rd-digit EC prediction.

Next, we sought to develop the prediction-score-weighted version of Equation 3.16. To do so, we had to consider several factors. First, a semantic representation T_x, of a protein x in a corpus X, issuing from this C, represents a set of CD PSSM identifiers. Second, the prediction tool that assigns these terms is RPS-BLAST, and the prediction scores s returned by this tool are bit scores. Third, the bit score distributions for the datasets used in this work are heavily influenced by outliers. Fourth, the bit score scale does not range from 0 to 1.

To solve these issues we had to normalize the bit scores, and then rescale them. By doing this we ensured that the resulting bit score distributions were not as influenced by outliers, and that their values ranged from 0 to 1.

To this end, we first performed a quantile transformation on the bit score cumulative distribution function (generalized as F_Y for the random variable Y). We achieved this by using

the quantile function of the standard normal distribution (Φ^{-1}). Afterwards we rescaled the resulting distribution with min-max feature scaling. We named the equation that includes both of these steps Q, and define it as:

$$Q(y) = \frac{y' - y'_{min}}{y'_{max} - y'_{min}} \mid y' \in (\Phi^{-1} \circ F_Y)(y) \tag{3.19}$$

Where the circle denotes the function composition operation, such that given two arbitrary functions f and g, we write $g(f(x)) = (g \circ f)(x)$. After these transformations we only considered the transformed bit scores whose original value was $\geqslant 80$ for downstream processing.

As a result of this normalization procedure, we are able to define the prediction-score-weighted version of Equation 3.16, as:

$$PIC_S(T_x) = \max\{IC_u(t) \cdot (Q \circ S)(t) : t \in T_x\} \mid R = \varnothing \tag{3.20}$$

Where, S is the retrieval function defined in Equation 3.17; Q represents the normalization procedure defined in Equation 3.19; and the circle denotes the function composition operation.

We decided that further equations built upon the concept of Protein Information Content, should use the prediction-score-weighted versions (Equation 3.18 and 3.20), and not the unweighted ones (Equation 3.15 and 3.16). We made this choice for the following reason. If a protein sequence has a predicted term t whose IC_u value is high, but whose prediction score s is low, then the resulting Protein Information Content should reflect that balance.

Ultimately, we intended to devise a global equation. This equation should include the three PIC_S values, one for each C used in this work—i.e., GO terms, EC numbers, and CDs. To do so, we first define the total number of distinct terms annotated to proteins x in a corpus X, for a given C, as the cardinality resulting from the union of all semantic representations T_x, provided all terms t in the union set are distinct:

$$N_X(C) = \left| \left\{ \forall t_i, t_j \in \bigcup_{x \in X} T_x : t_i \neq t_j \right\} \right| \tag{3.21}$$

Secondly, we define a triple named Z that contains each C used in this work, where C represents any of the three Classification Systems:

$$Z = \langle EC, GO, CD \rangle \mid C \in Z \tag{3.22}$$

Thirdly, we define a weight measure, so that we are able to highlight the contributions of each C in comparison to the remainder. We formulate the weight of a C (w_C), as the total number of distinct terms, for that C, annotated to proteins x in a corpus X, divided by the sum of the total number of distinct terms annotated to proteins x in a corpus X for each C in Z, as follows:

$$w_C = \frac{N_X(C)}{\sum_{C \in Z} N_X(C)} \tag{3.23}$$

Fourthly, assume that for a protein x in a corpus X, there is a superset ($T_{x,Z}$), that comprises a semantic representation T_x, for each C in Z, such that:

$$T_{x,Z} = \{T_{x,C} : C \in Z\} \tag{3.24}$$

We then applied the normalization procedure described in Equation 3.19, to each PIC_S distribution (one for each C in Z). These distributions required normalization before being used in the global equation for two reasons. First, the PIC_S calculation differs when taking into account the presence (Equation 3.18), or absence (Equation 3.20) of subsumption relations among the terms in that C. Second, the PIC_S distributions presented different scales. By normalizing the PIC_S distributions, we ensure that they were fit to a standard normal distribution, and that they scale from 0 to 1.

Thus, for a protein x in a corpus X, with an associated superset $T_{x,Z}$ comprising a semantic representation T_x for each C in Z, and assuming a weight w_C that highlights the contributions of each C in Z; we define the global equation, named Protein Annotation Index (PAI), as a weighted arithmetic mean:

$$PAI(T_{x,Z}) = \sum_{C \in Z} (Q \circ PIC_S)(T_{x,C}) \cdot w_C \tag{3.25}$$

Where, Q represents the normalization procedure defined in Equation 3.19, and the circle denotes the function composition operation. In Figure 3.2 we show a diagram that illustrates the logic behind this rationale, as well as the workflow undertaken for each dataset.

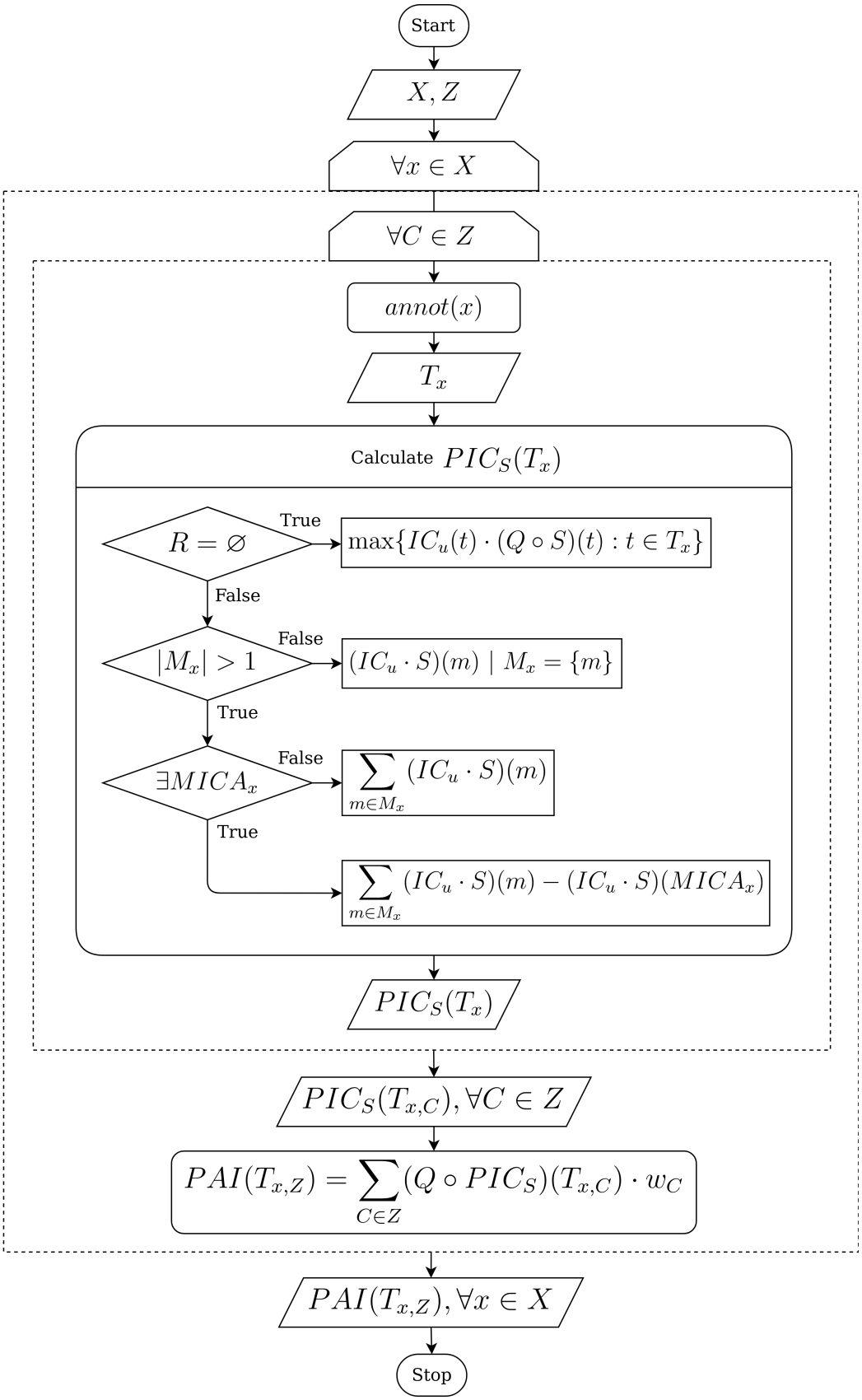


Figure 3.2: Flowchart representation of the rationale behind the information-theoretic equations, and the workflow undertaken for each dataset. The oval shapes represent the beginning and the end of the workflow for each dataset. The parallelograms represent inputs and outputs. The rectangles without top corners represent a *for* loop, and the dashed rectangles enclose the processes that take place inside that loop. The diamonds represent conditional choices. The rectangles with sharp corners represent the outcomes of each choice. The rectangles with rounded corners represent functions.

3.2.3 Minimum Method Thresholding Algorithm

We calculated the value of *PAI* (Equation 3.25) for each protein x in a corpus X—i.e., each protein in each dataset. The collection of *PAI* values for a dataset constitutes a *PAI* distribution. A *PAI* distribution will be simply denoted as *PAI* henceforth.

Histogram-based thresholding algorithms are nonparametric, unsupervised procedures that automatically select a critical value (i.e., threshold) from an intensity histogram [15]. These algorithms allow to partition a grayscale image into two classes. The pixels whose value exceeds the threshold are assigned to one class, and the remainder to the other. This process is known as image thresholding [15].

Grayscale images are usually represented as an intensity histogram. Each histogram bin depicts a gray-level. The intensity of each bin corresponds to the number of pixels in that level. The number of gray-levels, and consequently that of histogram bins, for 8-bit grayscale images is 256 [16]. Programming toolboxes that offer collections of thresholding algorithms—e.g., Python's Scikit-Image [17]—default to 256 histogram bins.

Given that we are using *PAI* as input, we needed to calculate the number of histogram bins a priori. We did this by using NumPy's [18] "`histogram_bin_edges`" function, with the "`bins`" parameter set to "`auto`". The "`auto`" setting defines the optimal number of histogram bins as the maximum between the output of Sturges formula [19], and that of the number of bins calculated by the Freedman-Diaconis (FD) rule [20]. Following the notation set in the previous section, we rewrite Sturges formula as:

$$k_{Sturges} = \lceil \log_2 |X| \rceil + 1 \tag{3.26}$$

Where X represents a corpus, consisting of a collection of protein sequences x, and $k_{Sturges}$ is the number of histogram bins. We also rewrite the FD rule as:

$$h_{FD} = 2\frac{\mathrm{IQR}(PAI)}{\sqrt[3]{|X|}} \tag{3.27}$$

Where $IQR(PAI)$ is the interquartile range of the data, and h_{FD} is the bin-width. To estimate the number of bins for *PAI* using the FD rule, we must calculate the difference between the maximum and the minimum value of *PAI* and divide it by h_{FD}. This is due to the fact that

the FD rule is not used to calculate the number of bins, but rather their width. The number of histogram bins k_{FD}, using the bin-width h_{FD}, for PAI, is given by:

$$k_{FD} = \left\lceil \frac{PAI_{max} - PAI_{min}}{h_{FD}} \right\rceil \tag{3.28}$$

Thus, the optimal number of histogram bins k^*, is determined by:

$$k^* = \max\left\{k_{Sturges}, k_{FD}\right\} \tag{3.29}$$

Adapting the notation used by Glasbey [15], we define a histogram B as a collection of bins b:

$$B = \langle b_1, ..., b_{k^*} \rangle \tag{3.30}$$

Where b_i is the number of data-values from PAI that belong to index i, and k^* is the optimal number of bins (Equation 3.29). The Minimum Method algorithm [21] proposes that, the threshold for segmenting a histogram, is the value i at which the bin counts b_i are minimized, in a valley between maxima [15]. This definition assumes that the histogram B is bimodal. Bimodality can be achieved by smoothing each bin b in B:

$$b'_i = \frac{b_{i-1} + b_i + b_{i+1}}{3} \mid i \in [1, k^*] \ \wedge \ b_0 = b_{k^*+1} = 0 \tag{3.31}$$

Where b' is the smoothed bin. This process is repeated for all b in B until only two local maxima are found. The final smoothed histogram B' can be written as:

$$B' = \langle b'_1, ..., b'_{k^*} \rangle \mid \sum_{i=1}^{k^*} [b'_{i-1} < b'_i > b'_{i+1}] = 2 \tag{3.32}$$

Where the square brackets are Iverson brackets (Equation 3.14). Assuming a smoothed bimodal histogram B' containing bins b', of which only two are local maxima, the threshold (τ) is equal to the value of i for which b'_i is minimized, such that:

$$\tau = \underset{i}{\operatorname{argmin}}\{ \ b'_i \ : \ \nexists j \in [1, k^*] \ \ (b'_j < b'_i) \ \} \tag{3.33}$$

In Figure 3.3 we show a diagram that illustrates the logic behind this method, and the workflow undertaken for each dataset, after being subjected to the steps described in subsection 3.2.2 (see Figure 3.2).

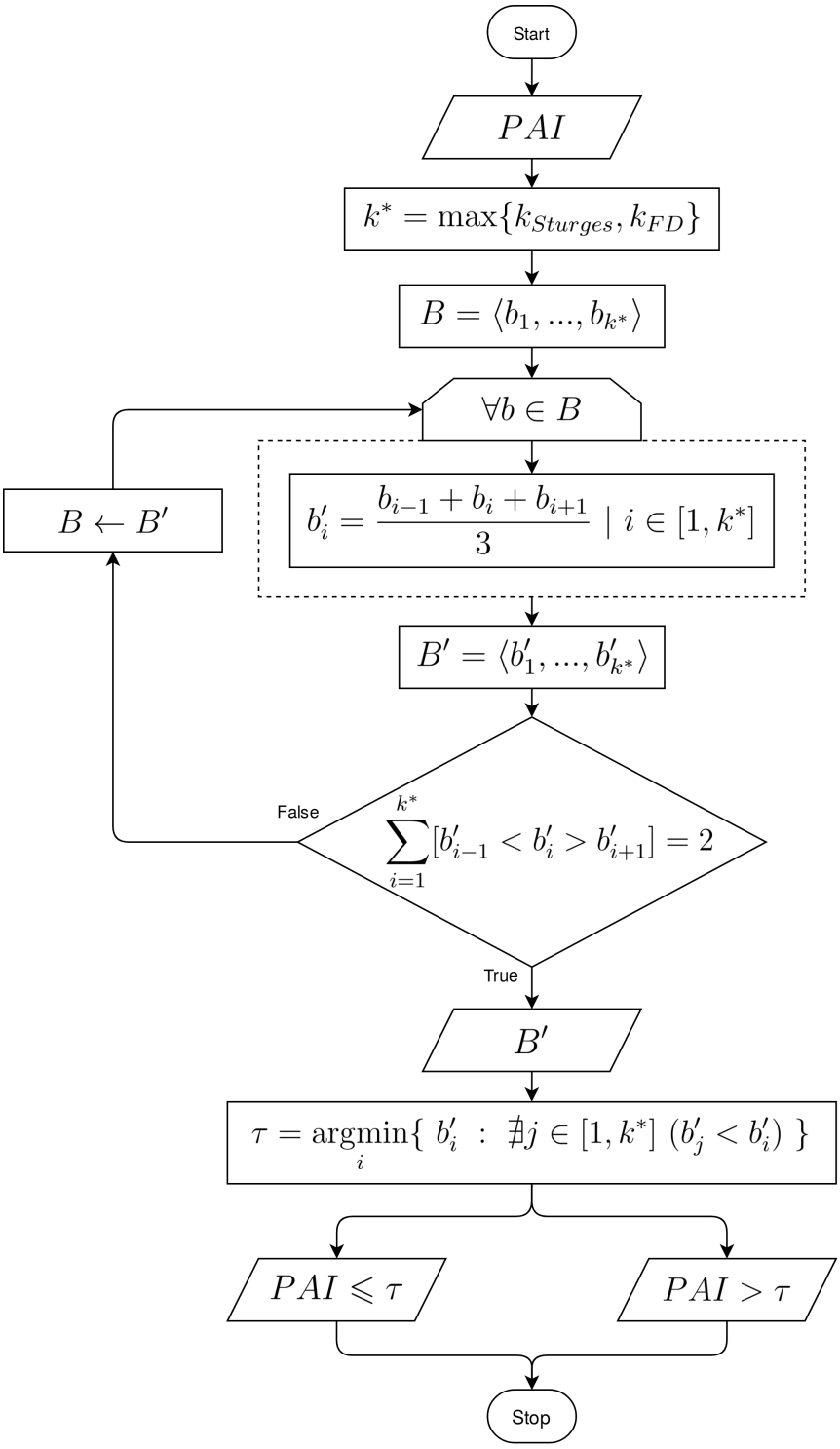


Figure 3.3: Flowchart representation of the Minimum Method Thresholding Algorithm, and the workflow undertaken for each dataset. The oval shapes represent the beginning and the end of the workflow for each dataset. The parallelograms represent inputs and outputs. The rectangles without top corners represent a *for* loop, and the dashed rectangles enclose the processes that take place inside that loop. The diamonds represent conditional choices. The rectangles with sharp corners represent variable declarations. $B \leftarrow B'$ denotes variable reassignment. *PAI* refers to the distribution of values gathered by calculating Equation 3.25 for each protein within a dataset.

3.3 Results and Discussion

3.3.1 The IC_u distributions

Continuing the work of chapter 2, our next objective was to validate the protein sequence annotations of the Hypotheticals dataset. We also wanted to know if these annotations allowed for the molecular function representation of the proteins enclosed therein. To do so, we had to quantify, and qualify, the information conveyed by the annotation terms in each Classification System—i.e., GO, EC, and CD.

We calculated the *IC* for every distinct term assigned to the proteins in each dataset. We did this for each Classification System. We also uniformized the *IC* values (IC_u) so that these would range from 0 to 1 (Equation 3.7). This step yielded six distributions—one for each Classification System, per dataset (Figure 3.4).

The Hypotheticals dataset has a lower diversity of EC numbers (1,964), than SwissProt (2,044)(Figure 3.4.A). However, relative to SwissProt, the Hypotheticals dataset shows an increased representation of ECs with an IC_u value between $\sim$0.62 and $\sim$0.86, and from $\sim$0.96 to 1. These distributions differ significantly (two-sample K-S p-value $= 0.0$, Mann–Whitney U p-value $= 0.0001$). They are also heteroscedastic (Levene p-value $= 0.0001$). The EC IC_u distribution of the Hypotheticals dataset has a heavy tail (kurtosis $= 0.7978$), whereas the EC IC_u distribution of SwissProt has a thinner one (kurtosis $= -0.0829$). In both datasets the EC IC_u distribution is leaning towards higher values. This skewness is more prominent in the Hypotheticals dataset (skewness $= -0.828$), than in SwissProt (skewness $= -0.648$).

In the case of the GO DAG, specificity is known to be poorly related with depth [2]. The reason for this is that, in the GO DAG, leaf nodes differ substantially in their distance from the *root* [7]. We posit that this is not the case for EC numbers for three reasons.

First, every protein with an EC, has this number obligatorily at a 4th-digit resolution—i.e. the maximum depth of the graph. Second, in the EC Classification System all leaves are at the same level of depth in the graph. Third, this level is that of maximum depth, and in this case, specificity (Figure 3.1.B).

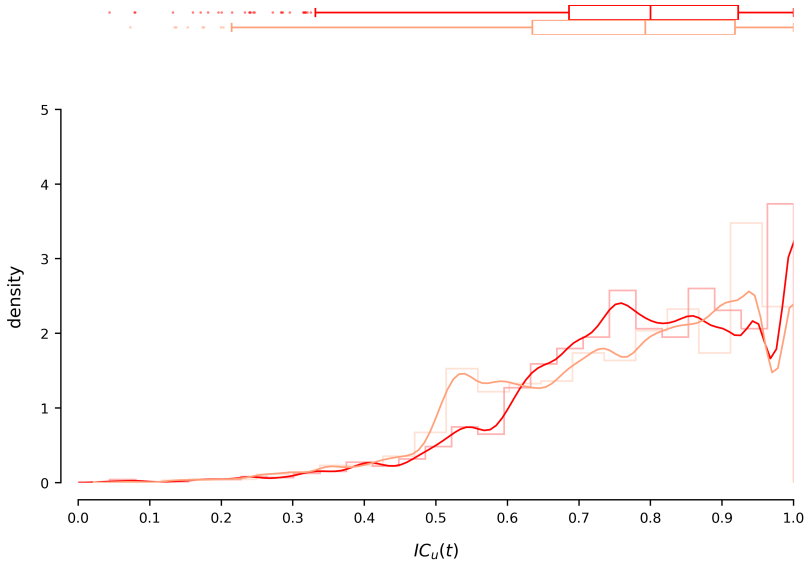
Conversely, DeepGO annotates proteins with GO terms at different levels of depth (see Figure 3.1.A). This implies a greater representation of shallower—and arguably less specific—terms in the GO IC_u distribution (Figure 3.4.B), in comparison to that of EC (Figure 3.4.A). The GO

IC_u distribution of the Hypotheticals dataset shows steady density values between the IC_u of 0.4 and 0.9. This implies similar representation of both unspecific and specific GO terms. It is also a recurring observation. We saw that only a small fraction of proteins possess a GO term annotation deeper than the 3rd level of the DAG (see Figure 2.8). We also reported that the two most populous over-represented terms were located at the depth levels of 1, and 2 of the GO DAG (see Figure 2.14.A).

This might be explained by the fact that, in the GO IC_u distribution of the Hypotheticals dataset, even if a protein is annotated with a specific GO, its ancestors also occur via *is-a* relationship. Thus increasing the probability of occurrence of unspecific terms. This differs from the context of the EC IC_u distribution. In the GO IC_u distribution the ancestors that occur by subsumption increase the representation of the already existing, unspecific sole annotations. Whereas the ancestors in the EC IC_u distribution occur exclusively by subsumption.

The GO IC_u distribution of SwissProt leans towards higher IC_u values (skewness $= -0.7686$) than that of the Hypotheticals dataset (skewness $= -0.2112$). It also shows greater density between the IC_u of $\sim$0.63 and 0.9, and lower density between $\sim$0.27 and 0.55. This contrasts with the GO IC_u distribution of the Hypotheticals dataset. It also suggests that there is a greater representation of specific terms annotated to proteins from SwissProt, in comparison to the Hypotheticals dataset.

A

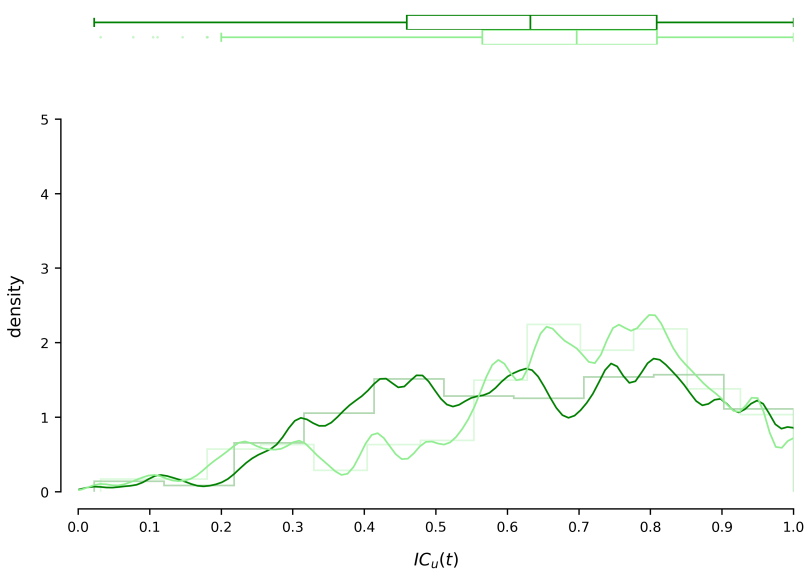


	Hypotheticals	SwissProt
count	1,964	2,044
mean	0.7835	0.7634
std	0.1659	0.1759
min	0.0436	0.0727
25%	0.6859	0.6346
50%	0.7997	0.7925
75%	0.9223	0.9178
max	1.0	1.0
S-W	0.942	0.9481
S-W p-value	0.0	0.0
skewness	-0.828	-0.648
kurtosis	0.7978	-0.0829

	Statistic	p-value
two-sample K-S	0.1091	0.0
Levene	15.1532	0.0001
Mann-Whitney U	1,867,488.5	0.0001

Figure 3.4. (Continued on next page.)

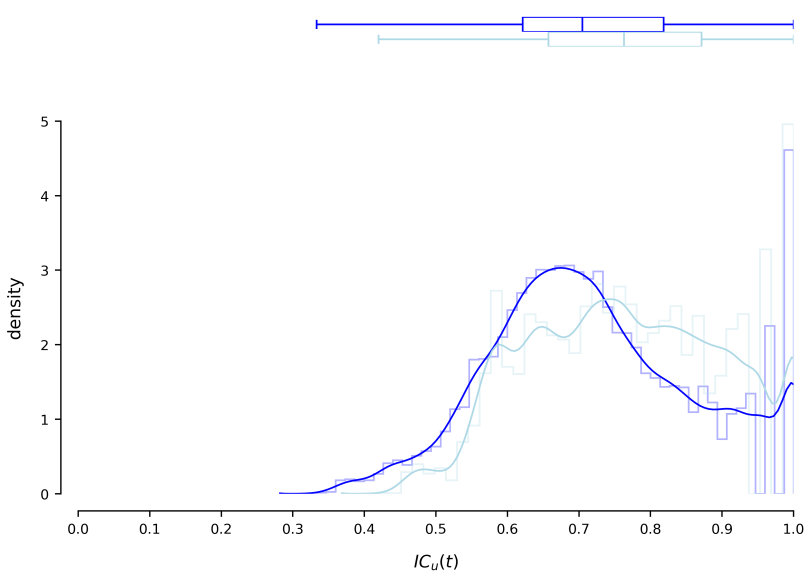
B



	Hypotheticals	SwissProt
count	358	233
mean	0.6279	0.6614
std	0.2205	0.216
min	0.0222	0.0312
25%	0.459	0.5646
50%	0.632	0.6968
75%	0.8086	0.809
max	1.0	1.0
S-W	0.9753	0.9445
S-W p-value	0.0	0.0
skewness	-0.2112	-0.7686
kurtosis	-0.785	0.0076

	Statistic	p-value
two-sample K-S	0.1535	0.0022
Levene	3.0531	0.0811
Mann-Whitney U	37,540.0	0.02

C



	Hypotheticals	SwissProt
count	33,967	19,706
mean	0.7209	0.7676
std	0.1436	0.1355
min	0.333	0.4197
25%	0.6211	0.6572
50%	0.7048	0.7629
75%	0.8182	0.8713
max	1.0	1.0
S-W	0.9785	0.974
S-W p-value	0.0	0.0
skewness	0.2081	-0.0008
kurtosis	-0.507	-0.9069

	Statistic	p-value
two-sample K-S	0.1649	0.0
Levene	3.3042	0.0691
Mann-Whitney U	271,597,536.5	0.0

Figure 3.4: IC_u distribution of each Classification System: EC (**A**), GO (**B**), and CD (**C**); and for each dataset: Hypotheticals (darker color), and SwissProt (lighter color). Each plot is colored according to its annotation type: EC (red), GO (green), CD (blue). Each plot contains a KDE, an histogram outline and a top boxplot, for each dataset. The boxplots share the same x-axis as the main plot. A table of summary statistics and tests is shown to the right of each plot. The cardinality refers to the number of distinct terms annotated to proteins from each dataset. t: a given term; S-W: Shapiro-Wilk test of normality; two-sample K-S: two-sample Kolmogorov-Smirnov test; Levene: Levene's test of homoscedasticity; Mann-Whitney U: Mann-Whitney U test.

The two GO IC_u distributions are homoscedastic (Levene p-value $= 0.0811$), and dissimilar (two-sample K-S p-value $= 0.0022$, Mann–Whitney U p-value $= 0.02$). Additionally, the GO IC_u distribution of the Hypotheticals dataset has a thinner tail (kurtosis $= -0.785$), than that

of SwissProt (kurtosis $= 0.0076$). This result also emphasizes the increased representation of unspecific annotations for the Hypotheticals dataset.

The CD IC_u distribution of the Hypotheticals dataset partially resembles a Normal distribution (Figure 3.4.C). Albeit for an increase in density from the IC_u of 0.9 upward. This is shown by both datasets. This density at $IC_u \geqslant 0.9$ suggests an increased representation of highly specific terms. The almost-Normal appearance might be due to two factors. First, this Classification System has the greatest number of distinct terms annotated to the Hypotheticals dataset (33,967). Second, this Classification System is non-hierarchical. Therefore, in the absence of inheritance relations, the IC of a term only relies on the probability of its occurrence. Which might explain why this probability density function seems to approach a normal distribution.

On the other hand, SwissProt is leaning towards higher IC_u values (skewness $= -0.0008$), than the Hypotheticals dataset (skewness $= 0.2081$). It also has a thinner tail (kurtosis $= -0.9069$), than the Hypotheticals dataset (kurtosis $= -0.507$), and almost half the number of distinct terms (19,706). Both distributions differ significantly (two-sample K-S p-value $= 0.0$, Mann–Whitney U p-value $= 0.0$), and are homoscedastic (Levene p-value $= 0.0691$).

3.3.2 Protein Information Content (PIC)

The IC_u cannot represent the entire informational value of an annotated protein. It also lacks the possibility to include multiple terms into its calculation. Therefore, the cardinality of IC_u is oriented towards the term, and not the protein. An equation that calculates the overall information content from multiple terms could be useful for proteins annotated with more than one term. It could also be of use if one wants to consider multiple terms to compute the information content; yet does not want to dilute the specificity of the result. This might happen if one were to apply the arithmetic mean. For these reasons, we conceived two metrics that could be useful towards the solution of these hindrances. We will explain these metrics in the following paragraphs.

We define the semantic representation [8] of a protein sequence, as a set of terms that describe its molecular function. In this work, the semantic representation of a protein might be one of three types of sets. It might contain: (i) terms that are at the deepest level of the graph (i.e., 4th-digit ECs); (ii) terms at different depths of the graph (i.e., GO terms); or (iii) terms that behave like keywords (i.e., CD PSSM identifiers). We can assort these types of semantic

representation into two main groups: those whose terms belong to a Classification System with subsumption relations (i.e., EC and GO); and that which does not (i.e., CDs).

We could not come up with a single equation for both groups, due to their different nature. Therefore, we developed two equations instead, one for each group. Note that these two equations are different means to achieve the same goal. This goal is to calculate the information content of a protein via its semantic representation. Each equation is specific to the nature of the group it was developed for. Yet they are the same metric. Therefore we named both as Protein Information Content (*PIC*, Equation 3.15 and 3.16).

For the semantic representations that belong to a Classification System with subsumption relations, the *PIC* is equal to the sum of the IC_u of the deepest terms in the subgraph, minus the IC_u of the most informative common ancestor (*MICA*, Equation 3.15). Whereas for the semantic representations that belong to a Classification System without subsumption relations, the *PIC* is equal to the maximum IC_u (Equation 3.16).

Each computational method we used assigns a score for every term it predicts. These scores range from 0 to 1, in the case of DeepEC and DeepGO; or are bit scores in the case of RPS-BLAST. If we incorporate these prediction scores into the calculation of *PIC*, we can establish a prediction-score-weighted version of these equations. We named these equations PIC_S (Equation 3.18 and 3.20). Just as *PIC*, PIC_S also comprises two equations. Each attending to the presence or absence of subsumption relations in a Classification System. Contrasting with *PIC* however, the PIC_S associates the IC_u of a predicted term with the degree of confidence of that prediction. We achieve this by multiplying the IC_u of a term with its respective prediction score.

For the semantic representations that belong to a Classification System with subsumption relations, the PIC_S is equal to the sum of the paired products between the IC_u of the deepest terms in the subgraph, and their prediction scores, minus the product of the IC_u of the *MICA* with its prediction score (Equation 3.18). Whereas for the semantic representations that belong to a Classification System without subsumption relations, the PIC_S is equal to the maximum among the products of the IC_u of a term and the prediction score for that term (Equation 3.20).

For the last case of PIC_S, the prediction scores come in the form of bit scores. Given that these do not fit in a scale from 0 to 1, we needed to normalize them first (Equation 3.19). By normalizing them, we ensure that they share the same scale of the prediction scores provided by DeepEC and DeepGO.

3.3.3 The PIC and PIC$_S$ distributions

We calculated the *PIC* and the *PIC*$_S$ for all proteins from both datasets. We did this for each Classification System. This generated twelve distributions (Figure 3.5, 3.6, and 3.7). In this section we will address the *PIC* and *PIC*$_S$ distributions while comparing them to one another. We will do as such for each Classification System.

The semantic representations for the EC Classification System (Figure 3.5), show *PIC* (Figure 3.5.A) and *PIC*$_S$ (Figure 3.5.B) maxima that exceed 1. Indeed, if a protein has multiple 4th-digit EC assignments, *PIC* (Equation 3.15) may be greater than 1. This outcome depends on three factors: (i) the number of 4th-digit ECs; (ii) their IC_u; and (iii) the IC_u of *MICA*. This scenario can also be envisioned for *PIC*$_S$ (Equation 3.18).

In the case of *PIC*$_S$ the same outcome will depend on two additional factors: the prediction scores of the 4th-digit ECs, and the prediction score of *MICA*. Consequently, we posit that these values relate to proteins with more than one 4th-digit EC, possibly representing proteins with multiple molecular functions. This might explain why these proteins were left uncharacterized in the first place, given the added difficulty in classifying a protein with multiple molecular functions.

We theorize that for the proteins with multiple 4th-digit EC assignments there might be four distinct scenarios taking place. These scenarios depend on whether a *MICA* exists, its IC_u, and whether *PIC* or *PIC*$_S$ is calculated. These scenarios also apply to proteins with multiple GOs at the maximum depth of the sub-DAG. Albeit with some differences, as disclosed later on.

As a first scenario, suppose that a protein has multiple EC assignments. Assume that these assignments share a *MICA* with a high IC_u (e.g., a common 3rd-digit EC). The outcome of *PIC* being greater than 1 will depend on the number, and IC_u, of the ECs—provided that their sum exceeds 1 after subtracting the IC_u of *MICA* (see Equation 3.15).

In a second scenario, consider the same setting as the first. Suppose that besides having a high IC_u the *MICA* also has a high prediction score. Yet, assume that we calculate *PIC*$_S$ instead. If the prediction scores of the ECs are low, the resulting *PIC*$_S$ may be less than 1 after subtracting the product of *MICA*'s IC_u with its prediction score (see Equation 3.18).

Consider a third scenario, where a protein has multiple assignments, as in the previous examples. However, in the present case the *MICA* is unspecific (e.g., shares the 1st-digit EC), or is absent altogether (e.g., the 4th-digit EC numbers have no ancestry in common). In this

context, even two ECs with an average IC_u (e.g., IC_u = 0.51), will equate to a *PIC* that exceeds the value of 1. The magnitude of this *PIC* will be mainly constrained by the number of ECs. This scenario differs from the first in the sense that, in the presence of an unspecific *MICA*—or absence thereof—the outcome of *PIC* is dictated by the number of EC assignments, even if not specific.

As the fourth and last scenario, assume the same setting as in the third, but we calculate PIC_S instead. In this case, we are able to coerce the sum of the IC_u for these terms, by striking a balance with their prediction scores. Thus, by calculating PIC_S instead of *PIC* we ensure that if this value exceeds 1, then it must be due to the presence of high prediction scores. This translates to high confidence in these predictions, even if bearing median informational value.

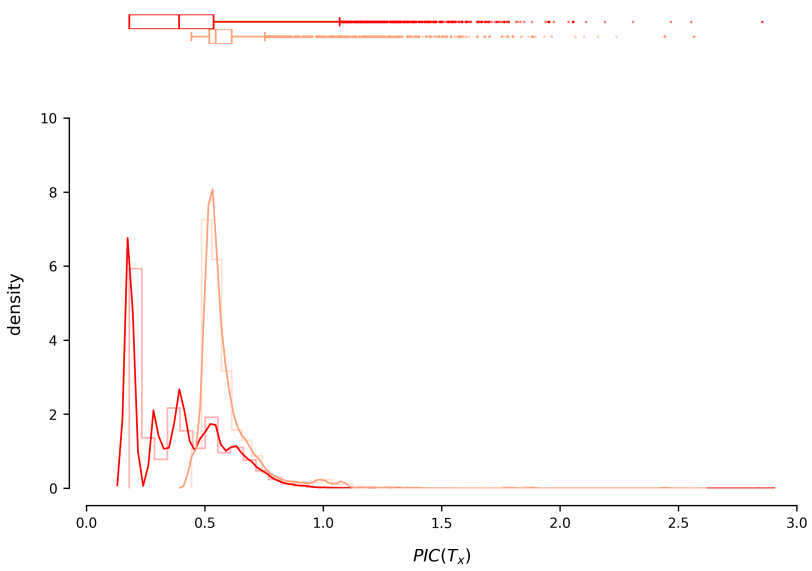
The PIC_S distribution (Figure 3.5.B) for the Hypotheticals dataset shows a lower minimum (0.0907) than that of *PIC* (0.1813, Figure 3.5.A), but an identical 25% percentile (0.1813). This suggests that upon considering the prediction scores, there were some values that shifted towards the left. Arguably this is due to the fact that the prediction scores for these terms were low.

The *PIC* distribution has a skewness = 0.8441, while that of PIC_S is 1.0702. Both distributions show positive kurtosis. PIC_S has a heavier tail (kurtosis = 2.0542), than *PIC* (kurtosis = 1.3814). For both *PIC* and a PIC_S, these values of skewness and kurtosis might be due to the outliers whose values exceed 1. Thus creating right-skewed distributions, with heavy tails to the right.

Most proteins from the SwissProt dataset have *PIC* (Figure 3.5.A) and PIC_S (Figure 3.5.B) values centered between 0.5 and 0.6. Each distribution shows a density maximum around 0.51. They also have higher minima than those of the Hypotheticals dataset (*PIC* min = 0.4438, PIC_S min = 0.2242). Both SwissProt distributions are skewed towards the right (*PIC* skewness = 4.4315, PIC_S skewness = 4.4373), and present the heaviest tails out of both datasets (*PIC* kurtosis = 34.2537, PIC_S kurtosis = 35.138).

The *PIC* and PIC_S distributions for SwissProt appear nearly identical. Both graphically and statistically. This might be due to the fact that DeepEC included proteins from SwissProt in its training set [11]. Meaning that the prediction scores for the SwissProt proteins will be equal to, or nearing 1. Thus having no significant weight upon the "raw" informational content provided by *PIC*.

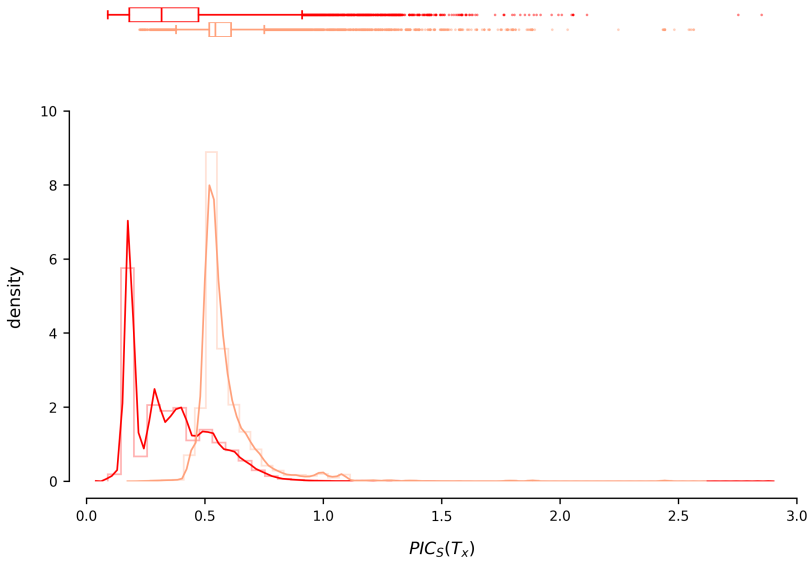
A



	Hypotheticals	SwissProt
count	336,584	151,886
mean	0.3957	0.5933
std	0.1981	0.1473
min	0.1813	0.4438
25%	0.1813	0.5193
50%	0.3917	0.5455
75%	0.5365	0.6131
max	2.8538	2.5647
S-W	0.897	0.6315
S-W p-value	0.0	0.0
skewness	0.8441	4.4315
kurtosis	1.3814	34.2537

	Statistic	p-value
two-sample K-S	0.6326	0.0
Levene	50,876.4817	0.0
Mann-Whitney U	10,588,633,737.5	0.0

B



	Hypotheticals	SwissProt
count	336,584	151,886
mean	0.3533	0.5902
std	0.1751	0.1459
min	0.0907	0.2242
25%	0.1813	0.5182
50%	0.3178	0.5452
75%	0.473	0.6114
max	2.8507	2.564
S-W	0.8912	0.6413
S-W p-value	0.0	0.0
skewness	1.0702	4.4373
kurtosis	2.0542	35.138

	Statistic	p-value
two-sample K-S	0.7243	0.0
Levene	30,688.5175	0.0
Mann-Whitney U	7,046,116,217.5	0.0

Figure 3.5: *PIC* (**A**) and PIC_S (**B**) distributions for the EC Classification System; and for each dataset: Hypotheticals (darker color), and SwissProt (lighter color). Each plot contains a KDE, an histogram outline and a top marginal boxplot, for each dataset. The boxplots share the same x-axis as the main plot. A table of summary statistics and tests is shown to the right of each plot. The cardinality refers to the number of proteins annotated with at least one 4th-digit EC and belonging to each dataset. T_x: Semantic representation of a given protein x; S-W: Shapiro-Wilk test of normality; two-sample K-S: two-sample Kolmogorov-Smirnov test; Levene: Levene's test of homoscedasticity; Mann-Whitney U: Mann-Whitney U test.

The *PIC* distributions for GO (Figure 3.6.A) also show maxima that exceed 1. This suggests the existence of proteins with multiple molecular functions, as posited for EC before. Mind that DeepGO might annotate a protein with multiple GOs at different levels of depth. Even if only considering the terms at maximum depth of a sub-DAG, these might not be at the maximum

depth of the entire GO DAG. Therefore, the IC_u for these terms might not add up to a *PIC* greater than 1. Furthermore, the PIC_S distribution values for GO (Figure 3.6.B) do not exceed nor equate 1, showing lower maxima instead (Hypotheticals dataset max $= 0.7952$, SwissProt max $= 0.8844$). This implies that the prediction scores for these annotations are lower than those for the ECs. Thus driving the maxima leftwards below the value of 1, by weighing down specific terms whose confidence is low.

The majority of proteins from the Hypotheticals dataset have a *PIC* below that of 0.2 (Figure 3.6.A). Whereas for SwissProt this value is below that of 0.3 (Figure 3.6.A). This may relate to the fact that DeepGO can provide a protein with a term at a deeper or shallower level as its only annotation. This translates to a greater representation of unspecific terms in the *PIC* distribution. These values are lower in the PIC_S distribution (Figure 3.6.B).

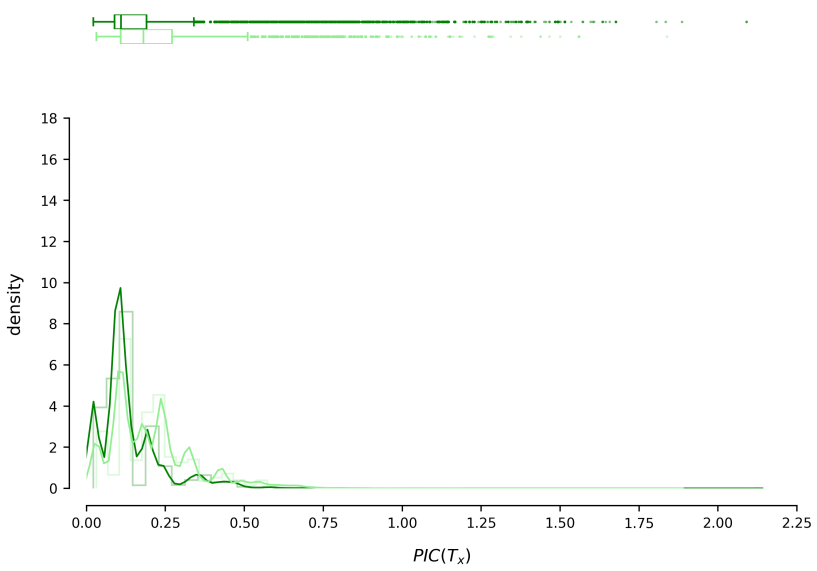
Most proteins from the Hypotheticals dataset have PIC_S values lower than 0.07. The majority of proteins from SwissProt have PIC_S values lower than 0.11. This outlines the fact that the prediction scores for these annotations are low. Thus, they drive the already low *PIC* values further down.

Overall, the *PIC* and PIC_S distributions from GO have lower values than those of EC. However, in the Hypotheticals dataset, the number of proteins with GO annotation (12,644,075) surpasses that of proteins with EC annotation (336,584). Few of these proteins possess a GO annotation deeper than the 3rd level of the DAG (see Figure 2.8). Consequently these low values of *PIC* and PIC_S for GO are expected.

The *PIC* and PIC_S distributions for the CD Classification System from the Hypotheticals dataset (Figure 3.7) resemble Normal distributions. However, the Shapiro-Wilk test for both *PIC* and PIC_S indicates otherwise (p-value $= 0.0$). This resemblance might share the explanation we proposed for the IC_u distribution of CD (Figure 3.4.C). That is, with no concept of depth, nor that of ancestors, the *PIC* and PIC_S equations used for CD (see Equation 3.16 and 3.20) produce distributions that approach Normality.

Akin to the IC_u distribution of CD (Figure 3.4.C), in Figure 3.7.A each *PIC* value corresponds to the maximum IC_u among those calculated for the semantic representation of a protein (Equation 3.16). And in Figure 3.7.B, each PIC_S value equals the product of this maximum IC_u with its transformed bit score (Equation 3.20). In all of these scenarios the values shown rely exclusively on the probability of occurrence of a CD PSSM identifier.

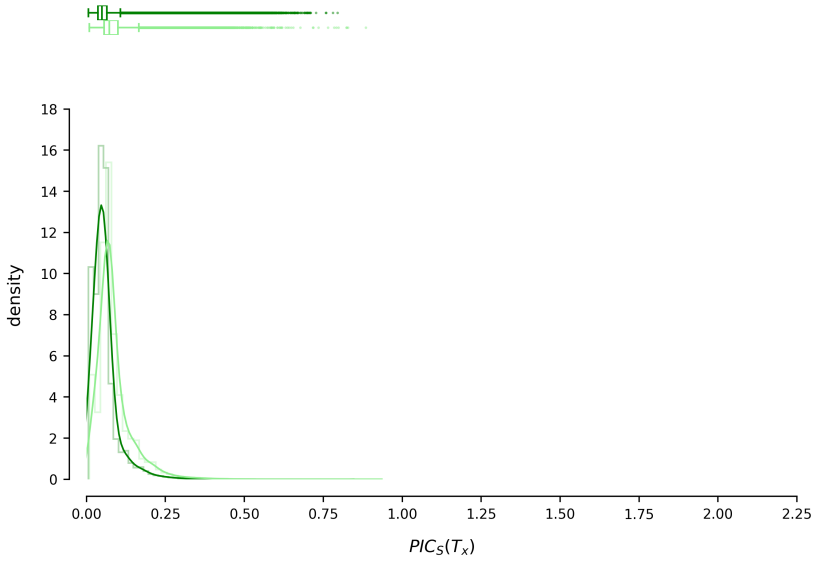
A



	Hypotheticals	SwissProt
count	12,644,075	235,464
mean	0.1344	0.2061
std	0.1024	0.1381
min	0.0222	0.0312
25%	0.0893	0.1081
50%	0.1095	0.1809
75%	0.1906	0.2707
max	2.0898	1.8379
S-W	0.8431	0.8854
S-W p-value	0.0	0.0
skewness	1.9993	1.4156
kurtosis	6.1897	3.3005

	Statistic	p-value
two-sample K-S	0.3517	0.0
Levene	48,792.4007	0.0
Mann-Whitney U	1.02704463922e+12	0.0

B



	Hypotheticals	SwissProt
count	12,644,075	235,464
mean	0.0572	0.0856
std	0.0414	0.0556
min	0.0067	0.0094
25%	0.0361	0.0561
50%	0.0493	0.0728
75%	0.0648	0.1002
max	0.7952	0.8844
S-W	0.7808	0.8051
S-W p-value	0.0	0.0
skewness	2.7556	2.3307
kurtosis	12.4909	9.2109

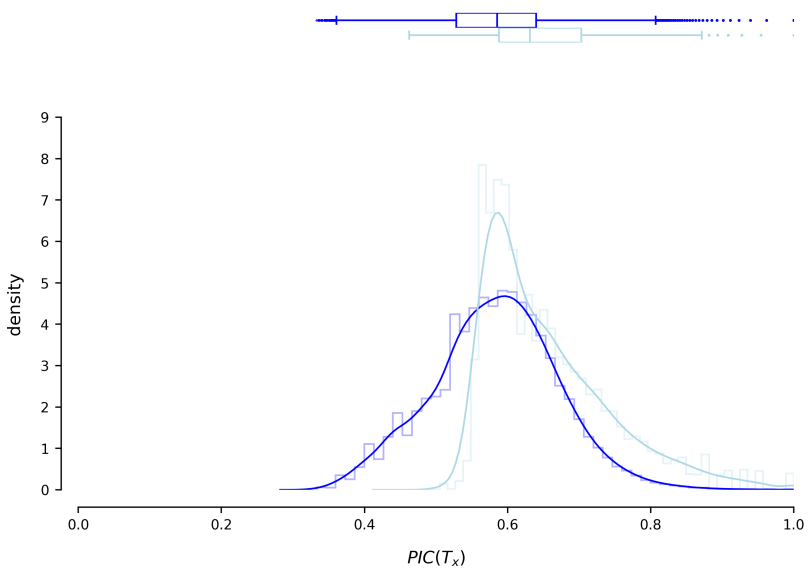
	Statistic	p-value
two-sample K-S	0.3742	0.0
Levene	17,460.6158	0.0
Mann-Whitney U	8.45344623196e+11	0.0

Figure 3.6: *PIC* (**A**) and PIC_S (**B**) distributions for the GO Classification System; and for each dataset: Hypotheticals (darker color), and SwissProt (lighter color). Each plot contains a KDE, an histogram outline and a top marginal boxplot, for each dataset. The boxplots share the same x-axis as the main plot. A table of summary statistics and tests is shown to the right of each plot. The cardinality refers to the number of proteins annotated with at least one GO term and belonging to each dataset. T_x: Semantic representation of a given protein x; S-W: Shapiro-Wilk test of normality; two-sample K-S: two-sample Kolmogorov-Smirnov test; Levene: Levene's test of homoscedasticity; Mann-Whitney U: Mann-Whitney U test.

In Figure 3.7.A the *PIC* distributions of both datasets follow a different trend than those shown for EC (Figure 3.5.A), and GO (Figure 3.6.A). This might be because *PIC* is being calculated differently for CD. By considering only the maximum IC_u, we attain the highest minimum values for both *PIC* and PIC_S, among all the Classification Systems, and for both datasets

(Hypotheticals: PIC min $= 0.333$, PIC_S min $= 0.2053$; SwissProt: PIC min $= 0.4625$, PIC_S min $= 0.265$). This inflation of both PIC and PIC_S distributions might also be due to the fact that we only considered CD PSSM identifiers whose bit score was $\geqslant 80$ prior to transformation (see subsection 3.2.2). Thus selecting the most significant domain annotation(s) per protein.

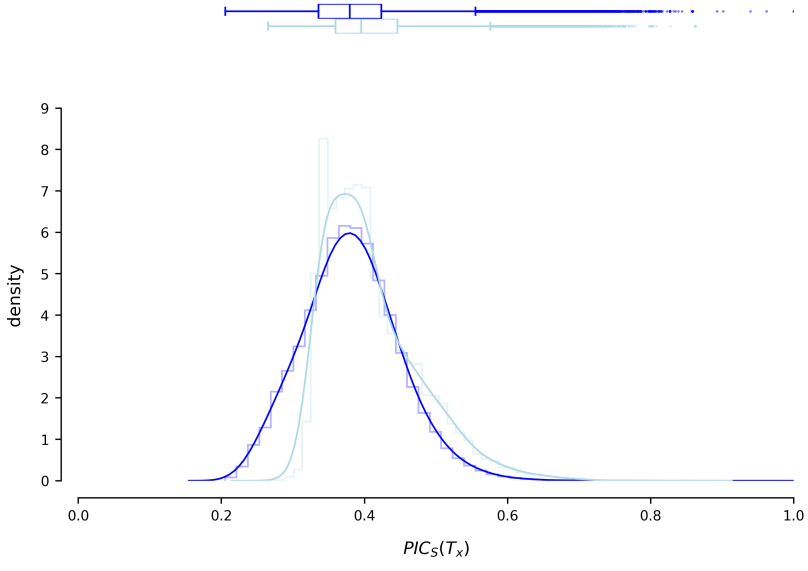
A



	Hypotheticals	SwissProt
count	6,142,383	231,627
mean	0.5832	0.6556
std	0.0878	0.0887
min	0.333	0.4625
25%	0.5281	0.588
50%	0.5854	0.6311
75%	0.6398	0.703
max	1.0	1.0
S-W	0.9917	0.8953
S-W p-value	0.0	0.0
skewness	0.085	1.2119
kurtosis	0.2957	1.2588

	Statistic	p-value
two-sample K-S	0.3465	0.0
Levene	131.0386	0.0
Mann-Whitney U	4.16941621004e+11	0.0

B



	Hypotheticals	SwissProt
count	6,142,383	231,627
mean	0.3814	0.4101
std	0.0673	0.0671
min	0.2053	0.265
25%	0.3358	0.3594
50%	0.3794	0.3953
75%	0.4236	0.4461
max	1.0	0.8624
S-W	0.9876	0.9186
S-W p-value	0.0	0.0
skewness	0.3322	1.1426
kurtosis	0.3729	1.406

	Statistic	p-value
two-sample K-S	0.1799	0.0
Levene	593.0364	0.0
Mann-Whitney U	5.63817587656e+11	0.0

Figure 3.7: PIC (**A**) and PIC_S (**B**) distributions for the CD Classification System; and for each dataset: Hypotheticals (darker color), and SwissProt (lighter color). Each plot contains a KDE, an histogram outline and a top marginal boxplot, for each dataset. The boxplots share the same x-axis as the main plot. A table of summary statistics and tests is shown to the right of each plot. The cardinality refers to the number of proteins annotated with at least one CD and belonging to each dataset. T_x: Semantic representation of a given protein x; S-W: Shapiro-Wilk test of normality; two-sample K-S: two-sample Kolmogorov-Smirnov test; Levene: Levene's test of homoscedasticity; Mann-Whitney U: Mann-Whitney U test.

Similarly to EC and GO, upon introducing the transformed bitscores (Equation 3.20), the PIC_S distributions for both datasets shift to the left, in comparison to PIC. The two PIC_S distributions have similar interquartile ranges, and near identical standard deviations (Hypotheticals std $= 0.0673$, SwissProt std $= 0.0671$), in spite of differing significantly (two-sample K-S p-value $= 0.0$, Mann–Whitney U p-value $= 0.0$), and being heteroscedastic (Levene p-value $= 0.0$).

The maximum value of PIC for either dataset is 1; whereas the maximum value of PIC_S for the Hypotheticals dataset is also 1, but that of SwissProt is 0.8624. This indicates that, in the Hypotheticals dataset, there are proteins whose product between the maximum IC_u and the transformed bit score is greater than that of the protein with the highest scoring attributes from SwissProt.

3.3.4 Protein Annotation Index (PAI)

We sought to create a global equation that enclosed the PIC_S from each Classification System. We also decided that this equation should enclose PIC_S instead of PIC. We chose to do so because the prediction scores ascribe each annotation with its degree of confidence. If a protein has a term whose IC_u value is high, but whose prediction score is low, then the end product should reflect that balance.

To this end, we first created a weight measure that highlighted the contribution of each Classification System to the global equation. We reckon that this measure is necessary, given that one Classification System might provide more information than the remainder. We define the weight of a Classification System (w_C), as the number of distinct terms from that Classification System, that were annotated to the proteins of a dataset (Equation 3.21); divided by the total number of distinct terms in all Classification Systems that were annotated to the proteins of a dataset (Equation 3.23). This is therefore a dataset-dependent metric. Using the Hypotheticals dataset as an example, the weights of the Classification Systems EC, GO, and CD are, respectively:

$$w_{EC} = \frac{1,964}{1,964+358+33,967} \approx 0.0541 \tag{3.34}$$

$$w_{GO} = \frac{358}{1,964+358+33,967} \approx 0.0098 \tag{3.35}$$

$$w_{CD} = \frac{33,967}{1,964 + 358 + 33,967} \approx 0.9360 \tag{3.36}$$

The lowest w_C for the Hypotheticals dataset is that of GO (w_{GO}). It accounts for 0.98% of total term diversity. This is due to the fact that there are only 358 distinct terms represented in the Hypotheticals dataset (see Figure 3.4.B). On the other hand, w_{EC} accounts for 5.41% of total term diversity, amounting to 1,964 distinct terms (see Figure 3.4.A). Although most proteins from the Hypotheticals dataset are presumably non-enzymes (see Figure 2.7), those that have EC annotation do so at the deepest level of the graph—i.e., the 4th-digit EC. This allows for w_{EC} to account for five times more term diversity than that of w_{GO}, given that the deeper the ontology level, the greater the number of distinct annotations, and thus the greater the probability of distinct term representation. The heftiest Classification System is CD, with w_{CD} accounting for 93.60% of the total term diversity. This was expected minding that the number of distinct CDs in the Hypotheticals dataset is 33,967 (see Figure 3.4.C). This is also a favorable result, because the existence of CD annotation might act as a "limiting factor"—i.e., it might dictate whether an amino-acid sequence is a protein or not. We make this assumption based on the premise that the sequences without CD annotation might not be proteins at all.

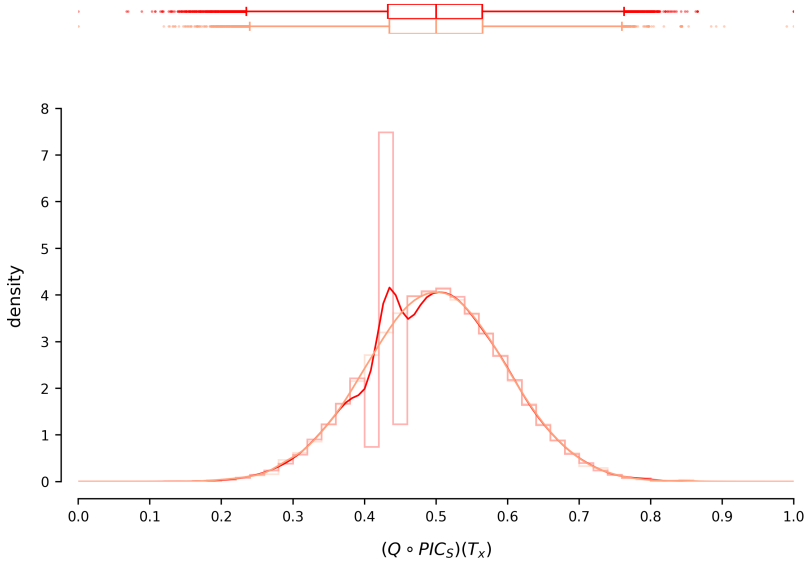
Afterwards, we normalized each PIC_S distribution using the Q function (Equation 3.19). We did this for both datasets. The PIC_S distributions had to be normalized before we could proceed for two reasons. First, the calculation of PIC_S differs when taking the presence (Equation 3.18), or absence (Equation 3.20) of subsumption relations in a Classification System into account. Second, the PIC_S distributions show different scales (see Figure 3.5.B, 3.6.B, and 3.7.B). By normalizing them, we ensure that they have been fit to a standard normal distribution, and that their scale ranges between 0 and 1. Thus being apt to undergo further processing. We named these normalized distributions $Q \circ PIC_S$ (Figure 3.8).

Both $Q \circ PIC_S$ distributions for EC (Figure 3.8.A) show a median value $\sim$0.5 (median $= 0.4999$ for the Hypotheticals dataset, and median $= 0.5001$ for SwissProt). The Hypotheticals distribution shows a local maxima near $Q \circ PIC_S = 0.44$. This might have lead to higher skewness (0.0108), in comparison to SwissProt (0.0091). This local maxima might be related to the one near the value of $PIC_S = 0.18$ (Figure 3.5.B) after normalization. The Hypotheticals and SwissProt distributions are dissimilar according to the two-sample K-S p-value $= 0$. However, the Mann-Whitney U p-value $= 0.4751$ indicates otherwise.

The $Q \circ PIC_S$ distributions for GO (Figure 3.8.B) have values of skewness and kurtosis close to 0 (Hypotheticals dataset skewness $= -0.0098$, kurtosis $= -0.007$; SwissProt dataset skewness $= -0.0075$, kurtosis $= -0.0093$). The SwissProt distribution shows a S-W p-value $= 0.6339$. This does not allow us to reject the null hypothesis for any significance level, thus suggesting Normality for this distribution. The Mann-Whitney U test p-value $= 0.151$ does not allow us to reject the null hypothesis, for any significance level. This indicates that both distributions are similar. Yet, the two-sample K-S p-value $= 0$ implies otherwise. Both distributions share the same standard deviation (0.0961). This is also shown by the Levene p-value $= 0.9253$, which does not allow us to reject the null hypothesis, for any significance level. Thus suggesting that these distributions are homoscedastic.

Both $Q \circ PIC_S$ distributions for CD (Figure 3.8.C) show mean and median values close to 0.5, and similar quantiles. The values of skewness and kurtosis are close to 0 for both distributions (Hypotheticals dataset skewness $= -0.01$, kurtosis $= -0.0532$; SwissProt dataset skewness $= -0.0076$, kurtosis $= -0.016$). These distributions are also dissimilar (two-sample K-S p-value $= 0.0247$, Mann-Whitney U test p-value $= 0.0113$). They show similar standard deviations (Hypotheticals std $= 0.0962$, SwissProt std $= 0.0963$). This is also supported by Levene p-value $= 0.5916$, which does not allow us to reject the null hypothesis, for any significance level. Hence hinting at homoscedasticity.

A

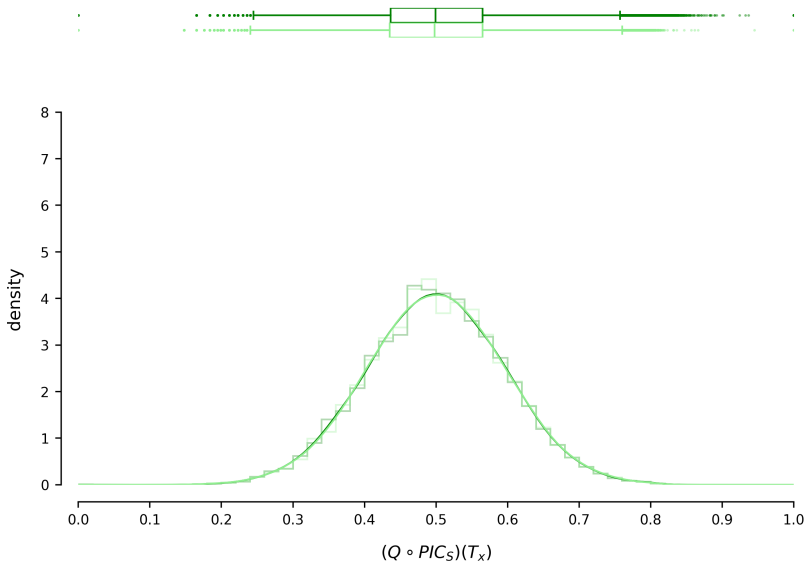


	Hypotheticals	SwissProt
count	336,584	151,886
mean	0.5001	0.5
std	0.0956	0.0964
min	0.0	0.0
25%	0.4328	0.4348
50%	0.4999	0.5001
75%	0.5648	0.5651
max	1.0	1.0
S-W	0.9977	0.9999
S-W p-value	0.0	0.0168
skewness	0.0108	0.0091
kurtosis	-0.0169	0.0177

	Statistic	p-value
two-sample K-S	0.0708	0.0
Levene	5.694	0.017
Mann-Whitney U	25,558,352,119.5	0.4751

Figure 3.8. (Continued on next page.)

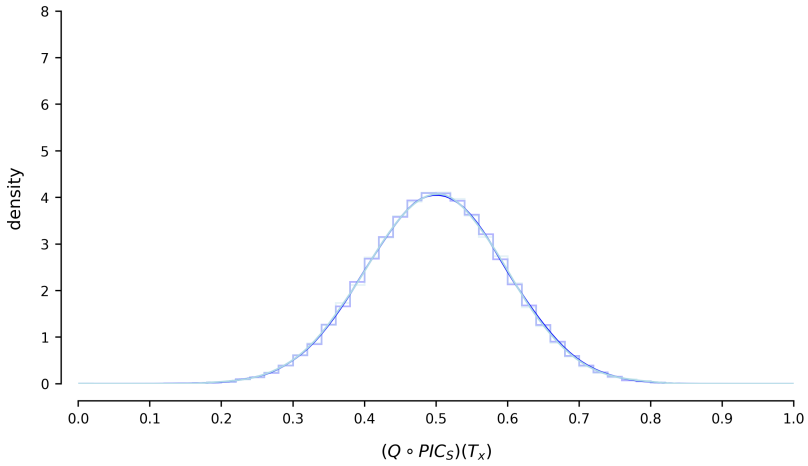
B



	Hypotheticals	SwissProt
count	12,644,075	235,464
mean	0.4998	0.5001
std	0.0961	0.0961
min	0.0	0.0
25%	0.4367	0.4354
50%	0.4994	0.4981
75%	0.5649	0.5652
max	1.0	1.0
S-W	0.9216	1.0
S-W p-value	0.0	0.6339
skewness	-0.0098	-0.0075
kurtosis	-0.007	-0.0093

	Statistic	p-value
two-sample K-S	0.0159	0.0
Levene	0.0088	0.9253
Mann-Whitney U	1.48676762331e+12	0.151

C



	Hypotheticals	SwissProt
count	6,142,383	231,627
mean	0.5001	0.4996
std	0.0962	0.0963
min	0.0	0.0
25%	0.435	0.4345
50%	0.5002	0.4999
75%	0.5654	0.5647
max	1.0	1.0
S-W	0.9871	0.9999
S-W p-value	0.0	0.0
skewness	-0.01	-0.0076
kurtosis	-0.0532	-0.016

	Statistic	p-value
two-sample K-S	0.0031	0.0247
Levene	0.2879	0.5916
Mann-Whitney U	7.09388721188e+11	0.0113

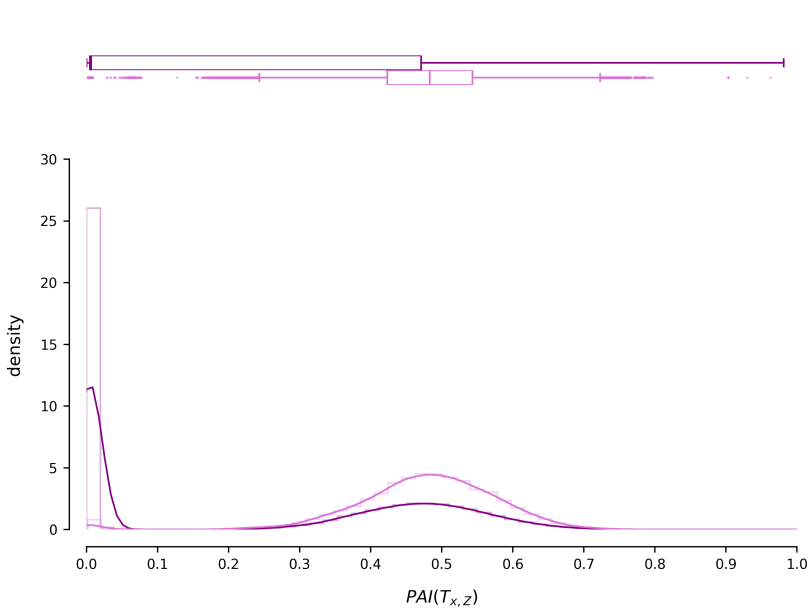
Figure 3.8: $Q \circ PIC_S$ distribution for each Classification System: EC (**A**), GO (**B**), and CD (**C**); and for each dataset: Hypotheticals (darker color), and SwissProt (lighter color). Each plot is colored according to its annotation type: EC (red), GO (green), CD (blue). Each plot contains a KDE, an histogram outline and a top marginal boxplot, for each dataset. The boxplots share the same x-axis as the main plot. A table of summary statistics and tests is shown to the right of each plot. The cardinality refers to the number of proteins with at least one annotated term for a given Classification System and belonging to each dataset. T_x: Semantic representation of a given protein x; S-W: Shapiro-Wilk test of normality; two-sample K-S: two-sample Kolmogorov-Smirnov test; Levene: Levene's test of homoscedasticity; Mann-Whitney U: Mann-Whitney U test.

We define the global equation as the weighted arithmetic mean of the products between the $Q \circ PIC_S$ for each Classification System, and its w_C (Equation 3.25). We named this equation Protein Annotation Index (*PAI*). *PAI* should allow to arrange the proteins from the Hypotheti-

cals dataset according to their global informational value. This way, the *PAI* distribution should portray a spectrum for different degrees of protein annotation.

We calculated *PAI* for all proteins from both datasets. This generated two distributions, one for each dataset (Figure 3.9). The *PAI* distribution for the Hypotheticals dataset shows a local maxima near 0. The boxplot at the top also suggests that at least 50% of the proteins from this dataset have a *PAI* value between 0 and 0.0066—i.e., those spanning from the left whisker ($Q_1 - 1.5 \times IQR$) to the median (Q_2). This corresponds to ∼6 million proteins with extremely low values of *PAI*. The distribution also shows a Normal density curve between 0.2 and 0.8. These observations indicate that the distribution is skewed to the right (skewness $= 0.2514$). The value of kurtosis is very low (kurtosis $= -1.7024$), which suggests a distribution with very light tails—i.e., without outliers. Indeed, there appear to be no outliers for this distribution, which can be confirmed by the boxplot at the top.

The *PAI* distribution for SwissProt shows minor density near 0. This density has a value lower than 1, whereas the one shown by the Hypotheticals dataset is greater than 10. The minimum value of *PAI* for SwissProt is 0.0018. This entails that, in contrast to the Hypotheticals dataset, the lowest value of global informational content in SwissProt does not equate the absence of information.



	Hypotheticals	SwissProt
count	12,651,618	235,543
mean	0.2329	0.4757
std	0.2428	0.1072
min	0.0	0.0018
25%	0.0048	0.4233
50%	0.0066	0.4832
75%	0.4709	0.5432
max	0.9813	0.9625
S-W	0.7517	0.919
S-W p-value	0.0	0.0
skewness	0.2514	-1.3139
kurtosis	-1.7024	4.4605

	Statistic	p-value
two-sample K-S	0.4989	0.0
Levene	91,930.6086	0.0
Mann-Whitney U	6.98535471687e+11	0.0

Figure 3.9: *PAI* distribution for each dataset: Hypotheticals (darker color), and SwissProt (lighter color). Each plot contains a KDE, an histogram outline and a top marginal boxplot, for each dataset. The boxplots share the same x-axis as the main plot. A table of summary statistics and tests is shown to the right of each plot. The cardinality refers to the number of proteins with at least one annotated term, for at least one Classification System, belonging to each dataset. $T_{x,Z}$: Superset containing the semantic representations of a given protein x, for each Classification System; S-W: Shapiro-Wilk test of normality; two-sample K-S: two-sample Kolmogorov-Smirnov test; Levene: Levene's test of homoscedasticity; Mann-Whitney U: Mann-Whitney U test.

Moreover, 50% of the proteins from the SwissProt dataset have a value of *PAI* between 0.0018 and 0.4832—i.e., from the minimum to Q_2—and 25% of these range from 0.4233 to 0.4832—i.e., from Q_1 to Q_2. The remaining 50% range from 0.4832 to 0.9625—i.e., from Q_2 to the maximum.

Curiously, the *PAI* maximum for SwissProt is lower than that of the Hypotheticals dataset (0.9625 versus 0.9813, respectively). This implies that the global informational content of the highest scoring protein(s) in the Hypotheticals dataset surpasses that of SwissProt. The SwissProt distribution also depicts a Normal density curve between the value of 0.2 and 0.8. In contrast to the *PAI* distribution for the Hypotheticals dataset, that of SwissProt is skewed to the left (skewness $= -1.3139$), and has heavy tails (kurtosis $= 4.4605$). These two distributions differ from one another (two-sample K-S p-value $= 0.0$, Mann–Whitney U p-value $= 0.0$), and are heteroscedastic (Levene p-value $= 0.0$).

3.3.5 Subsection and Analysis of the PAI distributions

The *PAI* distribution for both datasets appeared to be bi-modal. We speculated that the local maxima near 0 correspond to proteins with unspecific annotations. We also wondered to what extent did the Classification System weights influence the differentiation of values in each distribution. This directed us to further investigate these distributions.

To do so, we first needed to calculate a partition threshold. This threshold would allow us to separate each distribution into two, so that these could be studied individually. To calculate this threshold we used a histogram-based thresholding algorithm known as the Minimum method [15, 21] (see subsection 3.2.3). The resulting threshold value (τ) was 0.1529 for the *PAI* distribution of the Hypotheticals dataset (Figure 3.10.A), and 0.1271 for that of SwissProt (Figure 3.10.B). Both values of τ are located in a valley between local maxima (Figure 3.10.A and 3.10.B). By using the Minimum method on the *PAI* distributions we created two individual partitions, per dataset. One partition contains the values of *PAI* that are lesser or equal to τ. The other partition contains the values of *PAI* that are greater than τ. These partitions will be henceforth referred to as the "left" and "right" partitions, respectively. In the following sections we will analyze each partition individually; compare it between datasets; and then compare the "left" and the "right" partition against one another, for the same dataset. We will also briefly overview some annotations from these proteins.

3.3.5.1 The "left" PAI partition

The left partition of the Hypotheticals dataset (Figure 3.10.A) encloses 51.46% (6,509,908) of its total number of proteins (12,651,618). On the other hand, the left partition of SwissProt (Figure 3.10.B) encloses only ∼1.67% (3,928) of its total number of proteins (235,543). Thus, the proportion of proteins from SwissProt that failed to be effectively annotated is lower than that of the Hypotheticals dataset. The $Q \circ PIC_S(T_{x,CD})$ distribution on the left panel of Figure 3.10.A ranges from ∼0.09 to ∼0.19. It also shows an histogram bin whose $Q \circ PIC_S(T_{x,CD}) = 0$. We also observe that only 673 proteins are annotated with CD, accounting for 0.01% of the proteins present on the left partition. There are only 3 most-informative CD annotations among these 673 proteins: "Signal transduction histidine kinase" (CDD:223715): 671 proteins; "ATPase components of ABC transporters with duplicated ATPase domains" (CDD:223562): 1 protein; and "ABC-type glutathione transport system ATPase component, contains duplicated ATPase domain" (CDD:224048): 1 protein.

A

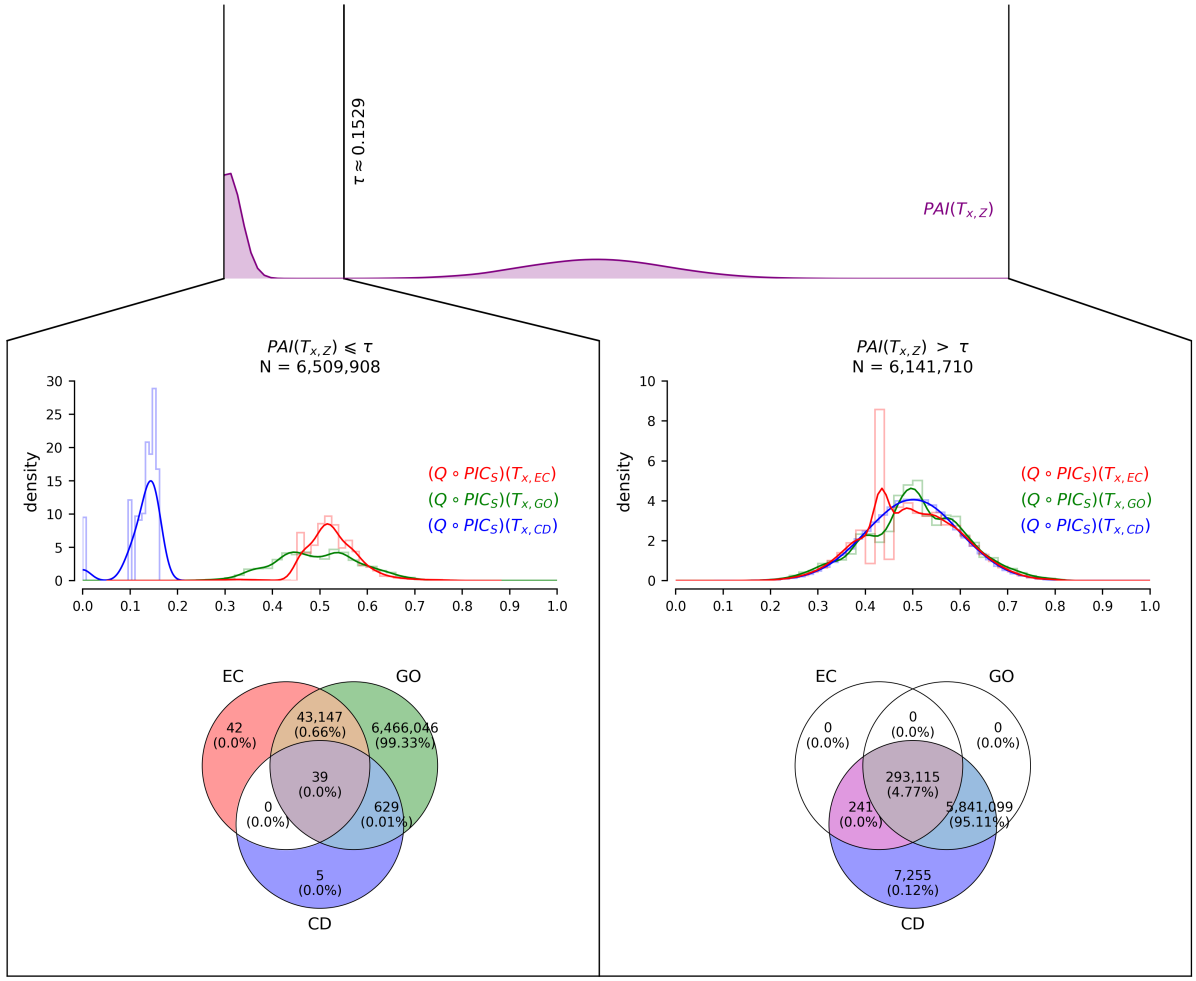


Figure 3.10. (Continued on next page.)

B

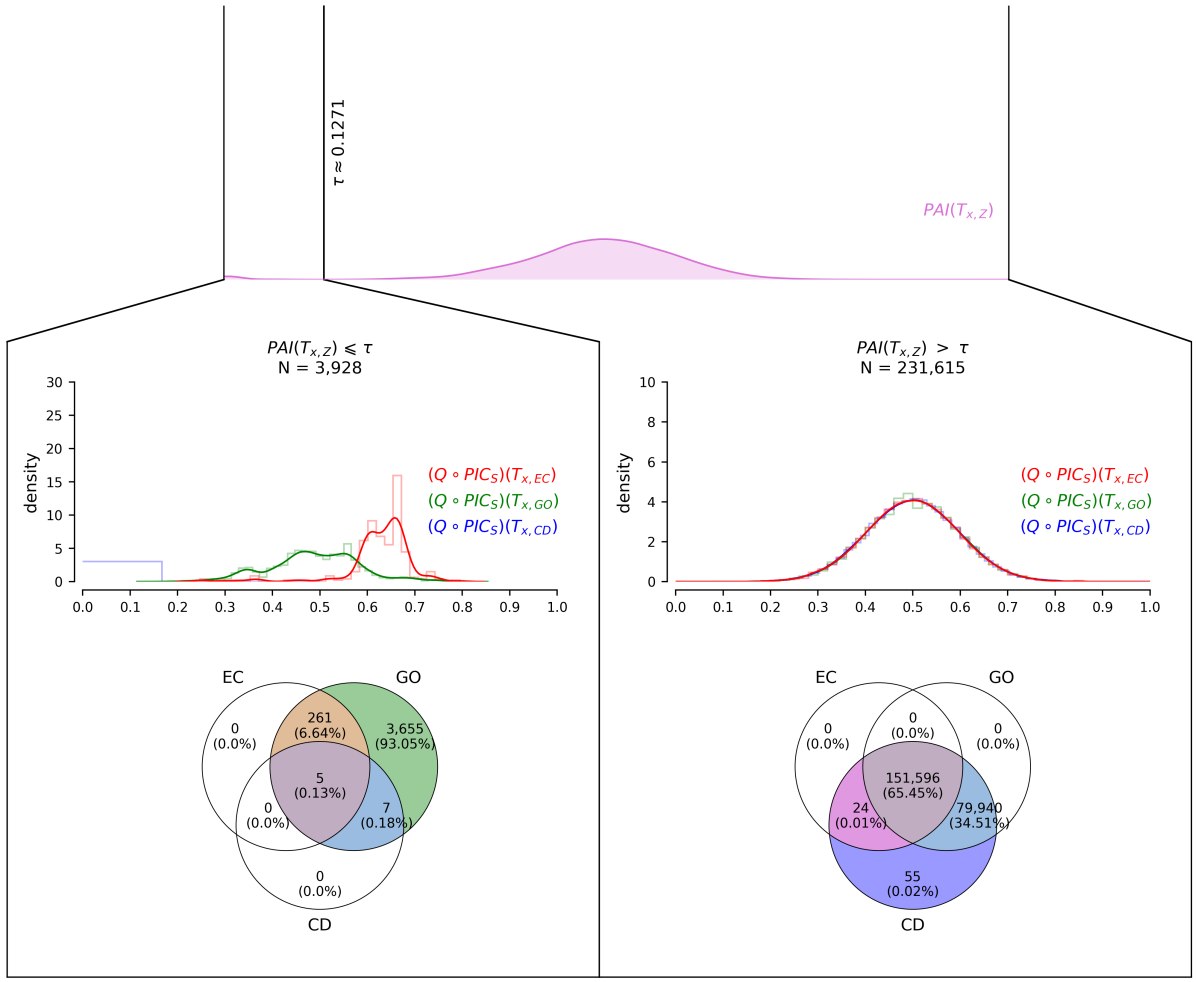


Figure 3.10: Subsection analysis portraits of the *PAI* distributions. Top: *PAI* distribution of the Hypotheticals dataset (**A**); and *PAI* distribution of SwissProt (**B**). The vertical line subsetting each *PAI* distribution depicts the Minimum Method threshold (τ). Left panel: data relating to proteins whose *PAI* is lesser than or equal to τ. Right panel: data relating to proteins whose *PAI* is greater than τ. Middle: $Q \circ PIC_S(T_{x,C})$ distributions for EC (red), GO (green), and CD (blue) respective to each partition. Bottom: Venn diagram depicting the protein counts for each Classification System, and intersections thereof, respective to each partition.

Attending to their $Q \circ PIC_S(T_{x,CD})$ values, we posit that these CDs are ubiquitous among the Hypotheticals dataset. Therefore they provide little to no information about the proteins they are annotated to. Moreover, CD is the Classification System that contributes the most to the final value of *PAI* (see Equation 3.36). For this reason a low value of $Q \circ PIC_S(T_{x,CD})$ will dictate a low *PAI*. Thus it is expected for the proteins with the lowest values of $Q \circ PIC_S(T_{x,CD})$ to be allocated to the left partition.

The SwissProt $Q \circ PIC_S(T_{x,CD})$ distribution on the left (Figure 3.10.B) shows a similar scenario. This distribution contains a single histogram bin, ranging from 0 to ∼0.18. This bin contains only 12 proteins, none of which have CD as their sole annotation.

The $Q \circ PIC_S(T_{x,GO})$ values for the left partition of the Hypotheticals dataset are spread throughout most of the axis. The presence of high values of $Q \circ PIC_S(T_{x,GO})$ on the left is due to the fact that this Classification System contributes less than 1% to the final value of *PAI* (see Equation 3.35). The Venn diagram shows that 99.33% (6,466,046) of these proteins are only annotated with GO. These proteins share 453 deepest GO assignments. Each assignment may include more than one GO at the maximum depth. From 453 assignments, the top 5 most represented are: "protein binding" (GO:0005515): 2,300,145 proteins; "DNA binding" (GO:0003677): 1,071,879 proteins; "binding" (GO:0005488): 992,332 proteins; both "catalytic activity" (GO:0003824) and "binding" (GO:0005488): 797,943 proteins; and "hydrolase activity" (GO:0016787): 317,523 proteins. These top 5 assignments represent 84,74% out of 99.33%, yet they carry very little information.

Nonetheless, the presence of proteins only annotated with GO, by itself, justifies the existence of high scoring $Q \circ PIC_S(T_{x,GO})$ values on the left. This is for the reason that: even if a protein was annotated with the most informative GO term in the Hypotheticals dataset, with a perfect prediction score (i.e., $Q \circ PIC_S(T_{x,GO}) = 1$); if it lacked any annotation from the two remainder Classification Systems, its *PAI* value would not be greater than:

$$(Q \circ PIC_S)(T_{x,GO}) \cdot w_{GO} = 1 \times 0.0098 = 0.0098 \tag{3.37}$$

Indeed, these proteins solely annotated with GO might explain the local maximum near 0 on the PAI distribution. We consider this to be reasonable. Mainly due to the fact that DeepGO appears to be the most permissive out of the 3 Classification Systems used. Thus its contribution for the final value of *PAI* should be minor. This also allows to group the proteins possessing only GO as their source of annotation on the left side of the *PAI* distribution. Enabling us to discriminate the proteins that were not pliant to annotation.

The left partition of SwissProt shows a similar assortment of $Q \circ PIC_S(T_{x,GO})$ values and respective density. The respective Venn diagram depicts the same scenario as that of the Hypotheticals dataset, where 93.05% of proteins have GO as their sole source of annotation.

The remaining proteins on the left partition of the Hypotheticals dataset amount to 0.67%. The majority of which (0.66%) have both GO and EC annotations, and only 629 (0.01%) have both CD and GO. Note that in this partition there are more proteins with EC annotation, than with CD.

There are 39 (∼0.0%) proteins annotated by all three Classification Systems. These 39 proteins share the same 4th-digit EC—"Histidine kinase" (EC:2.7.13.3). They also share the same most informative CD—"Signal transduction histidine kinase" (CDD:223715). Both assignments appear semantically consensual. However, these 39 proteins share 8 different deepest GO assignments: "binding" (GO:0005488): 18 proteins; "protein binding" (GO:0005515): 9 proteins; both "catalytic activity" (GO:0003824) and "binding" (GO:0005488): 4 proteins; "protein kinase activity" (GO:0004672): 3 proteins; "phosphotransferase activity, nitrogenous group as acceptor" (GO:0016775): 2 proteins; "ATPase activity" (GO:0016887): 1 protein; "organic cyclic compound binding" (GO:0097159): 1 protein; and "transmembrane receptor protein tyrosine kinase activity" (GO:0004714): 1 protein. From these assignments only "protein kinase activity" (GO:0004672) appears to be in agreement with the foregoing EC and CD annotations.

On the left partition from SwissProt (Figure 3.10.B), the percentage of proteins annotated with other Classification Systems besides GO is tenfold (6.95%) that of the Hypotheticals dataset. The majority of these proteins are also annotated with both GO and EC (6.64%).

Most $Q \circ PIC_S(T_{x,EC})$ values for the left partition of the Hypotheticals dataset (Figure 3.10.A) are located between ∼0.45 and ∼0.61. SwissProt (Figure 3.10.B) has a greater density of $Q \circ PIC_S(T_{x,EC})$ values between ∼0.58 and ∼0.68. The presence of these average $Q \circ PIC_S(T_{x,EC})$ values might relate to a similar scenario to that which was observed for GO above. Akin to GO, the Classification System contribution of EC (i.e., w_{EC}) only accounts for ∼5% of the value of *PAI* (see Equation 3.34). The majority of proteins on the left that are annotated with EC also have GO annotation but lack CD. This is the case for both datasets (Figure 3.10.A and 3.10.B). Thus, using the Hypotheticals dataset as an example: even if they were annotated with the most informative terms from their Classification System, paired with perfect prediction scores, the value of *PAI* would not be greater than:

$$\begin{aligned}(Q \circ PIC_S)(T_{x,GO}) \cdot w_{GO} + (Q \circ PIC_S)(T_{x,EC}) \cdot w_{EC} \\ = (1 \times 0.0098) + (1 \times 0.0541) = 0.0639\end{aligned} \tag{3.38}$$

Which is still lower than τ, and thus belonging to the left. Conversely to what was observed for the proteins annotated with GO, those solely annotated with EC are just 42 (∼0.0%). There are no proteins with both EC and CD annotation—with the exception of the 39 (∼0.0%) that

were annotated by all three Classification Systems (as discussed above). On the other hand, the left distribution of the SwissProt dataset (Figure 3.10.B) does not enclose proteins solely annotated with EC, CD, nor the intersection thereof—except for 5 (0.13%) that were annotated by all 3 Classification Systems.

3.3.5.2 The "right" PAI partition

The right *PAI* partition depicts a different scenario from the left. In both datasets (Figure 3.10.A and 3.10.B), all three $Q \circ PIC_S(T_{x,C})$ distributions are akin to those in Figure 3.8. In the Hypotheticals dataset (Figure 3.10.A), the right partition has less than half ($\sim$48,54%) of the total number of proteins. In contrast, that of SwissProt (Figure 3.10.B) has 98% of the total number of proteins. Analogously to what was reported for the left partition, this result implies that the proportion of effectively annotated proteins belonging to SwissProt is far greater than its counterpart from the Hypotheticals dataset.

The $Q \circ PIC_S(T_{x,CD})$ distributions are distinct from those on the left. They also show greater values. This is the case for both datasets (Figure 3.10.A and 3.10.B). The presence of a protein on the right seems to require the presence of a CD annotation with a greater informational value—i.e., annotation specificity. This is due to the fact that CD contributes the most to the value of *PAI* (see Equation 3.36).

The $Q \circ PIC_S(T_{x,GO})$ distributions for both datasets appear to span the same range, and the same densities, as their counterparts on the left. This is to be expected given that the contribution of GO to the value of *PAI* is negligible ($\sim$1%). It also suggests that the proteins on the right are selected based on their values of $Q \circ PIC_S(T_{x,CD})$ and $Q \circ PIC_S(T_{x,EC})$.

The $Q \circ PIC_S(T_{x,EC})$ distributions of both datasets show dissimilar spread and density between the left and right partitions. On the left partition of the Hypotheticals dataset (Figure 3.10.A) the bulk of the $Q \circ PIC_S(T_{x,EC})$ distribution density is limited to the values between 0.45 and 0.61. The same is true for SwissProt (Figure 3.10.B), but for the range of 0.58 to 0.68 instead. On the right partition there is a greater representation of lower $Q \circ PIC_S(T_{x,EC})$ values in both datasets. This might be a consequence of the fact that the proteins with EC annotation were also annotated by CD—as shown by the Venn diagrams of both datasets (Figure 3.10.A and 3.10.B).

Therefore, these proteins would be allocated to this partition nonetheless, on the account of

their $Q \circ PIC_S(T_{x,CD})$ value alone. Meaning that, if a protein has a $Q \circ PIC_S(T_{x,CD})$ value whose product with w_{CD} is greater than τ, it will be allocated to the right partition. Using the w_{CD} calculated for the Hypotheticals dataset as an example, this condition may be written as:

$$(Q \circ PIC_S)(T_{x,CD}) \cdot w_{CD} > \tau \Leftrightarrow (Q \circ PIC_S)(T_{x,CD}) \cdot 0.9360 > 0.1529 \tag{3.39}$$

A protein whose product of $Q \circ PIC_S(T_{x,CD})$ with w_{CD} is less than τ might also be allocated to the right, provided it has an annotation from one of the remainder Classification Systems. This last condition is required so that either the value of $Q \circ PIC_S(T_{x,EC})$ or $Q \circ PIC_S(T_{x,GO})$ bridges the gap between the value of $Q \circ PIC_S(T_{x,CD})$ and that of τ.

Both datasets have no proteins on the right partition solely annotated with EC, GO, or the intersection thereof. This result is due to what was explained above for Equation 3.38. Therefore the lack of proteins with only EC and/or GO annotations on this partition was to be expected. However, there are 7,255 (0.12%) proteins on the right partition of the Hypotheticals dataset solely annotated with CD, whereas its counterpart count on the left was just 5 (∼0.0%). These 7,255 proteins share 1,255 unique most-informative CD assignments. We posit that these assignments are rare CDs. For instance, the most represented CD in this group is "Uncharacterized membrane protein" (CDD:227665), which is shared by only 202 proteins.

SwissProt on the other hand has no proteins solely annotated with CD on the left, while on the right it has 55 (0.02%). Being exclusively annotated by CD does not seem to be the rule, but rather the exception. On the right partition of the Hypotheticals dataset, nearly all proteins have both CD and GO annotation, totalling 5,841,099 (95.11%). These proteins share 215,582 distinct GO-CD assignment combinations. In the equivalent subset of the left partition, this number of proteins was just 629 (0.01%).

Although it suffices to satisfy Equation 3.39 to be on the right, there are 0.12% of proteins where this might be the case. Suggesting that past τ the proteins are more likely to be annotated by at least two Classification Systems. This assumption is substantiated by 99.88% of proteins on the right, versus their left counterpart of 0.67%. A similar result is also observed on the right partition of SwissProt, where the proteins annotated by at least two Classification Systems total 99.97%, whereas on the left they are 0.31%.

On the Hypotheticals dataset, the majority of proteins annotated by all three Classification Systems are on the right partition, with the exception of 39 (∼0.0%) that are on the left. The

number of proteins on the right that match these criteria is 293,115 (4.77%). The fact that these proteins only comprise 4.77% of the total can be explained by supposing that most proteins from this dataset might not be enzymes at all, as postulated before. In contrast, the majority of proteins annotated by all Classification Systems on the right partition of SwissProt amount to 65.45%. Those only annotated by both GO and CD comprise 34.51%. This result implies that most proteins from the SwissProt dataset appear to be enzymes.

On the right partition of the Hypotheticals dataset, there are 241 (0.0%) proteins annotated just by EC and CD. There is no counterpart of this subset on the left. The same pattern is observed for SwissProt, but with a count of 24 (0.01%) proteins on the right instead. These 241 proteins from the Hypotheticals dataset share 126 combinations of 4th-digit ECs with most informative CDs. From these, the combination of "Histidine kinase" (EC:2.7.13.3) with "PAS fold" (CDD:315525) is the most represented, being shared by 23 proteins.

3.4 Conclusions

In this work we developed a family of metrics that allow to numerically qualify the annotations within a protein sequence dataset. These metrics classify the semantic representation of a given sequence according to its informational value and quality. To test and validate these metrics, we used both a gold-standard dataset as well as $\sim$12 million proteins that had been unyielding to annotation by their databases of origin.

By calculating *PIC* and/or PIC_S for all proteins annotated by each Classification System, we have shown that it is possible to represent an entire dataset as a collection of distributions. By calculating *PIC* we tap into the "raw" informational value enclosed by a semantic representation, whereas by calculating PIC_S, we assess the balance between the overall specificity of a semantic representation, and the extent to which we ought to trust it. These distributions depict spectra of protein annotation, and their values provide insight into the informational content of these annotations. Moreover, *PIC* and PIC_S also allow to disclose whether these annotations relate to multiple molecular functions.

Combining the PIC_S measures calculated for each Classification System enabled the creation of an index, which we named Protein Annotation Index (*PAI*). The *PAI* allows to systematize protein annotations according to the full extent of their informational value. Consequently, the *PAI* enables the creation of a global distribution that characterizes different degrees of protein annotation within a given dataset. Additionally, the *PAI* is both scalable and modular. It can be extended to include other Classification Systems, and modified according to specific needs.

We have also shown that by calculating the value of *PAI* for all proteins in a dataset, one can unravel distinct sub-distributions relating to proteins with differing levels of annotation. This development allowed us to distinguish which proteins had been more thoroughly annotated in each of the datasets we used.

After applying a thresholding algorithm, we were further able to select a value of *PAI* that allowed for the subsection of each dataset among two major groups of annotated proteins. We posit that these groups tentatively portray proteins that were effectively annotated, versus those that were not. By creating this dichotomy we hope to expedite the targeting of protein sequences that were the most pliant to *in silico* annotation. In addition, we aspire to facilitate the selection of the most promising protein sequences for further analyses, and ensuing experimental validation.

References

[1] Erdin S, Lisewski AM, Lichtarge O. Protein function prediction: towards integration of similarity metrics. Curr Opin Struct Biol. 2011;21(2):180–188.

[2] Pesquita C, Faria D, Bastos H, Ferreira AEN, Falcão AO, Couto FM. Metrics for GO based protein semantic similarity: a systematic evaluation. BMC Bioinformatics. 2008;9 Suppl 5:S4.

[3] Azuaje F, Al-Shahrour F, Dopazo J. Ontology-Driven Approaches to Analyzing Data in Functional Genomics. In: Bioinformatics and Drug Discovery. Humana Press; 2006. p. 67–86. Available from: https://doi.org/10.1385%2F1-59259-964-8%3A67.

[4] Furnham N. Complementary Sources of Protein Functional Information: The Far Side of GO. Methods Mol Biol. 2017;1446:263–274.

[5] Shannon CE. A Mathematical Theory of Communication. Bell System Technical Journal. 1948;27(3):379–423.

[6] Resnik P. Semantic Similarity in a Taxonomy: An Information-Based Measure and its Application to Problems of Ambiguity in Natural Language; 1999.

[7] Lord PW, Stevens RD, Brass A, Goble CA. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics. 2003;19(10):1275–1283.

[8] Couto FM, Lamurias A. Semantic Similarity Definition. In: Encyclopedia of Bioinformatics and Computational Biology. Elsevier; 2019. p. 870–876.

[9] Peng J, Uygun S, Kim T, Wang Y, Rhee SY, Chen J. Measuring semantic similarities by combining gene ontology annotations and gene co-function networks. BMC Bioinformatics. 2015;16:44.

[10] Kulmanov M, Khan MA, Hoehndorf R, Wren J. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. Bioinformatics. 2018;34(4):660–668.

[11] Ryu JY, Kim HU, Lee SY. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. Proc Natl Acad Sci U S A. 2019;116(28):13996–14001.

[12] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–3402.

[13] Lidl R, Pilz G. Applied Abstract Algebra. Springer Science & Business Media; 1997.

[14] Graham RL, Graham RE, Knuth DL, Knuth DE, Patashnik O. Concrete Mathematics: A Foundation for Computer Science. Addison Wesley Publishing Company; 1989.

[15] Glasbey CA. An analysis of histogram-based thresholding algorithms. CVGIP Graph Models Image Process. 1993;55(6):532–537.

[16] Sankur B. Survey over image thresholding techniques and quantitative performance evaluation. J Electron Imaging. 2004;13(1):146.

[17] van der Walt S, Schönberger JL, Nunez-Iglesias J, Boulogne F, Warner JD, Yager N, et al. scikit-image: image processing in Python. PeerJ. 2014;2:e453.

[18] Harris CR, Millman KJ, van der Walt SJ, Gommers R, Virtanen P, Cournapeau D, et al. Array programming with NumPy. Nature. 2020;585(7825):357–362.

[19] Sturges HA. The choice of a class interval. J Am Stat Assoc. 1926;21(153):65–66.

[20] Freedman D, Diaconis P. On the histogram as a density estimator:L 2 theory. Z Wahrscheinlichkeitstheorie verw Gebiete. 1981;57(4):453–476.

[21] Prewitt JM, Mendelsohn ML. The analysis of cell images. Ann N Y Acad Sci. 1966;128(3):1035–1053.

Chapter 4

The Biotechnological Potential within Functional Dark Matter of Prokaryotic Origin

Abstract

Prokaryotic gene products of unknown molecular function represent an astounding amount of uncharted sequence space. These elusive elements might comprise as much as 50% of total gene content in newly sequenced genomes. Efforts to annotate these sequences are often limited by their dependence on sequence similarity searches against reference genomes. This has led to a continuous agglomeration of uncharacterized genomic elements among public databases.

Increasing evidence indicates that these genes of unknown function might be a treasure trove of biotechnological potential. Previous reports have already unearthed several enzymes of industrial interest from within this functional "dark matter". Our previous work has suggested that a substantial amount of these sequences do code for putative enzymes.

In the present work we identified which sequences from a $\sim$12 million *de novo*-annotated protein sequence dataset belonged to enzyme classes with known biotechnological or industrial potential. We report 13,734 protein sequence representatives issuing from 64 enzyme sub-classes; 34 of which belonged to the Hydrolase class (EC:3.-.-.-).

We focus our efforts in characterizing a subset of 48 putative Cutinases (EC:3.1.1.74). We compare these cutinases with a dataset of polyethylene terephthalate (PET)-hydrolases. We report that the characteristic motifs are highly conserved. The foregoing 48 cutinases share the greatest amino-acid identity with two fungal PET hydrolases: the *Fusarium solani pisi* cutinase (FsC), and the *Humicola insolens* cutinase (HiC). To conclude, we posit that these 48 novel cutinases might withstand harsh industrial conditions as estimated from their imputed physico-chemical characteristics. We put forward that these 48 cutinases might be promising candidates for ensuing PET-degradation assays.

4.1 Introduction

Despite the advances in both sequencing technologies and in our understanding of microbial genomes, the biological function of ∼35% of microbial gene products remain a mystery [1]. In some cases the fraction of genes of unknown function can amount to as much as 50%, such as in newly sequenced genomes [2] and uncultured microbial candidate taxa [3–7]. In more extreme cases (e.g., metagenomic studies), these percentages can span from 85% [8] to 99% of total gene content [9, 10]. This has led to the coinage of the term microbial dark matter (MDM, [4]); alluding to the unexploited fraction of microbial diversity, along with its uncharacterized functional and metabolic potential. We will refer to this uncharacterized functional potential as "functional dark matter" (FDM).

These genomic elements of unknown function are a concealed treasure trove of biotechnological potential [7, 11–13]. Previous studies have already unearthed enzymes of biotechnological relevance from the MDM. These include, but are not limited to, cellulases [1], lipases [14], alcohol dehydrogenases [15, 16], enzymes that catalyze organophosphorus compounds [17], and others displaying enhanced stability under industrial conditions, as recently reviewed [12]. Therefore we posit that the sequences encased in the FDM may lead to escalating innovative biotechnological solutions in a world of ever-increasing societal demands.

Science has profited tremendously from enzymes of microbial origin, as substantiated by the work of numerous Nobel Prize laureates [7]. Examples of this are restriction enzymes [18], DNA polymerases [19] coupled to the advent of the polymerase chain reaction [20], and more recently the CRISPR-cas9 system [21]. Enzymes are also the backbone of numerous industries [14, 22–29]. Reactions catalyzed by enzymes are thought to follow the rules of green chemistry—they are safer, faster, and generate less waste than chemical catalysts [12]. The unmatched eco-friendly potential of enzymes is of vital use in the industry to mitigate the rampant overconsumption of our planet's resources [12]. In addition, the discovery of novel enzymes of biotechnological or industrial interest is critical for the growth of the industrial enzymes market [7]. This market amounted to 9.9 billion USD in 2019, and it is projected to reach 14.9 billion USD by 2027 [12]. Hence, microbial gene discovery can greatly push progress and development of new mechanisms and compounds of societal relevance.

Our previous work has shown that a substantial amount of uncharacterized protein sequences of prokaryotic origin code for putative enzymes (see chapter 2). In addition, we have

also indicated that the majority of these putative enzymes are pliant to *in silico* annotation, and that their potential functions can already be subject to further scrutiny (see chapter 3). For all the aforementioned reasons, in this work we sought to identify which proteins from the Hypotheticals dataset (see section 2.2 and 3.2.1) belonged to enzyme classes with known biotechnological or industrial potential. Besides showcasing the potential of putative enzymes from our dataset, we additionally direct our efforts towards the characterization of a specific group of enzymes: that of Cutinases.

Cutinases (EC:3.1.1.74) are serine esterases [30]. Their active site is characterized by a catalytic triad composed of serine (S), aspartate (D), and histidine (H) residues [30, 31]. They share this catalytic triad with several lipases and serine proteases [30]. Cutinases were first reported in phytopathogenic fungi that grow on cutin—a biopolyester that acts as a structural component of the cuticular layer in leaves [32, 33]. Besides biopolyesters, they are able to hydrolyse lipids, waxes and synthetic esters, among other substrates [30]. These enzymes are highly promiscuous, being able to hydrolyse roughly 78 substrates [34]. This versatility allows for their application in numerous industries, namely: agricultural, bioremediation, cosmetics, detergents, fine chemicals, foodstuffs, textile, and polymer chemistry [32]. The most pressing feature of cutinases is tied to the hydrolysis of high molecular weight polyesters, like polyethylene terephthalate (PET) [31].

PET is reported to be the most abundant polyester plastic [35]. It is mainly used in the textile and packaging industries [34, 35]. Nearly 70 million tons of PET are manufactured worldwide per year [35]. PET is commonly recycled through thermomechanical means, resulting in its loss of mechanical properties [36]. As such, PET is preferentially synthesized *de novo*, and its waste continues to accumulate in ecosystems throughout the globe [35, 37]. Therefore, PET hydrolases are in the spotlight of biotechnological applications. These applications include biorecycling, biocatalysis, waste treatment, and sustainable polymer chemistry [31].

In this chapter we report 48 novel putative cutinases issuing from the Hypotheticals dataset of previously unnanotated protein sequences. We compare these 48 cutinases with 30 bona fide PET hydrolases. We perform *in silico* analyses such as hierarchical clustering, Multiple Sequence Alignment (MSA), and the imputation of physico-chemical characteristics. We conclude that these putative cutinases might be promising candidates for ensuing biochemical characterization. Overall, the case study presented herein provides a proof-of-concept for the entire workflow developed throughout this thesis.

4.2 Methods

4.2.1 Datasets

4.2.1.1 Hypotheticals dataset and SwissProt

The Hypotheticals and SwissProt datasets refer to the same datasets as those described in section 2.2 and subsection 3.2.1. Both datasets comprise sequences that were annotated with at least a GO term, an EC number, or a CD. These annotations issue from DeepGO [38], DeepEC [39], and RPS-BLAST [40], respectively. The sources from which these sequences were obtained, the selection criteria, all (pre-)processing steps that ensued, as well as the annotation processes, are described in section 2.2.

In order to gather potential enzymes of biotechnological interest from each dataset, and ensure stringency, we selected the protein sequences that met the following criteria: (i) with EC, GO, and CD annotations; (ii) whose value of *PAI* was greater than the partitioning threshold (i.e., $PAI > \tau$, see chapter 3); (iii) annotated by a single 4th-digit EC assignment; and (iv) whose 4th-digit EC number was represented on Table 1.1. We selected proteins with a single EC assignment in order to exclude potential promiscuous enzymes, fusions and moonlighting proteins, which could convolute the interpretation of the results. We will refer to the group of proteins from each dataset that met the criteria above as that dataset's "biotechnological potential subset" (BPS). The Hypotheticals dataset BPS refers to 13,734 protein sequences. The SwissProt dataset BPS refers to 4,955 protein sequences.

4.2.1.2 PET hydrolase dataset

We gathered the 21 accession numbers from Table 1 in Carr et al. [31]. For each accession number, and depending on its datasource, we retrieved the protein sequence in FASTA format by performing API requests to either the NCBI protein database via the Efetch utility (`eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=protein&id=<ID>&rettype=fasta&retmode=text`); the Protein Data Bank (PDB) (`rcsb.org/fasta/entry/<ID>/display`); UniProt (`uniprot.org/uniprot/<ID>.fasta`); or UniParc (`uniprot.org/uniparc/<ID>.fasta`).

Most PET hydrolases described in Table 1 from Kawai et al. [33] were already present in Carr et al. [31]. Thus, we only retrieved the accession numbers for 9 proteins that were not:

Tfu_0882; Tfu_0883; Est1; Est119; TfAXE; Cut1; Cut2; Cbotu_EstA; and Cbotu_EstB. Each associated accession number either issued from the NCBI nucleotide database, or the NCBI protein database. In case an accession number issued from the nucleotide database, we first performed an API request to NCBI via the Elink utility (`eutils.ncbi.nlm.nih.gov/entrez/eutils/elink.fcgi?dbfrom=nuccore&db=protein&id=<ID>&idtype=acc`), in order to retrieve the accession number for the corresponding protein. Once all NCBI nucleotide accession numbers had been mapped to the NCBI protein database we retrieved the protein sequences in FASTA format by performing API requests to the NCBI protein database via the Efetch utility (`eutils.ncbi.nlm.nih.gov/entrez/eutils/efetch.fcgi?db=protein&id=<ID>&rettype=fasta&retmode=text`).

4.2.2 Multiple Sequence Alignment Analysis

We used Clustal Omega [41] (version `1.2.1`), to generate a pairwise distance matrix of amino-acid identity among 48 putative cutinases from the Hypotheticals dataset, along with the 30 reported PET hydrolases described in subsubsection 4.2.1.2.

We performed the Multiple Sequence Alignment between the two chosen PET hydrolases (HiC and FsC), and the 48 putative cutinases from the Hypotheticals dataset, using T-COFFEE [42] (version `13.45.47.aba98c5`), in "`regular`" mode. The Multiple Sequence Alignment visualization was created with TEXshade [43].

4.2.3 Physico-Chemical Metadata

We used BioPython's [44] `ProtParam` module to calculate the physico-chemical properties for each of the two chosen PET hydrolases (HiC and FsC), and the 48 putative cutinases from the Hypotheticals dataset, using a FASTA file containing their protein sequences as input. The physico-chemical properties calculated were: sequence length; molecular weight; grand average of hydropathicity [45]; aromaticity [46]; isoelectric point, and instability index [47].

4.3 Results and Discussion

4.3.1 Enzymes of biotechnological relevance

The biotechnological potential subset (BPS) of each dataset is shown in Figure 4.1, comprising a total of 64 4th-digit ECs. We arranged the proteins from the BPS into two figures, A and B. Figure 4.1.A shows the proteins whose 4th-digit EC belongs to the hydrolase class (i.e., EC:3.-.-.-), amounting to 34 ECs. Figure 4.1.B shows all other enzyme classes, comprising 30 ECs. We chose to show the hydrolases in a separate figure for two reasons. First, given that most enzymes of industrial relevance are of hydrolytic nature [14]. Second, because this class had the greatest number of 4th-digit ECs within each BPS.

The *PAI* values from the Hypotheticals BPS show both greater variance and greater maxima than those from SwissProt, for 50 out of 64 4th-digit ECs (Figure 4.1.A and B). The SwissProt BPS on the other hand shows greater median values of *PAI* for 36 EC classes (Figure 4.1.A and B). The Hypotheticals BPS has eight-fold more hydrolases for the selected EC numbers than the SwissProt BPS (Figure 4.1.A). However, when addressing each hydrolase class individually, this is only true for 27 out of 34 ECs (Figure 4.1.A). This was expected given the difference in size between datasets. Curiously, the SwissProt BPS has a greater representation of proteins from the remainder enzyme classes than the Hypotheticals BPS (Figure 4.1.B). However, this increased representation by SwissProt is only shown by 15 out of 30 ECs (Figure 4.1.B).

There are 6 EC numbers with proteins from the Hypotheticals BPS, that lack SwissProt counterparts (Figure 4.1.A and B). These ECs are: Cutinase (EC:3.1.1.74); Mannanase (EC:3.2-.1.25); Microbial collagenase (EC:3.4.24.3); Lactic acid dehydrogenase (EC:1.1.2.4); Protein disulfide-isomerase (EC:5.3.4.1); and Tyrosinase (EC:1.14.18.1).

Intrigued by this, we queried our data and found that there were 5 protein representatives for Tyrosinase (EC:1.14.18.1); 9 for Cutinase (EC:3.1.1.74); and 4 for Microbial collagenase (EC:3.4.24.3) present in the SwissProt dataset (data not shown). Despite having both EC and GO annotations, these proteins lacked CD annotation. Therefore they were unable to meet the first and the second filtering criteria enunciated above, and were thus left out from the BPS as a consequence. The remainder 3 EC numbers—i.e., Mannanase (EC:3.2.1.25); Lactic acid dehydrogenase (EC:1.1.2.4); and Protein disulfide-isomerase (EC:5.3.4.1)—do not possess any SwissProt counterparts whatsoever.

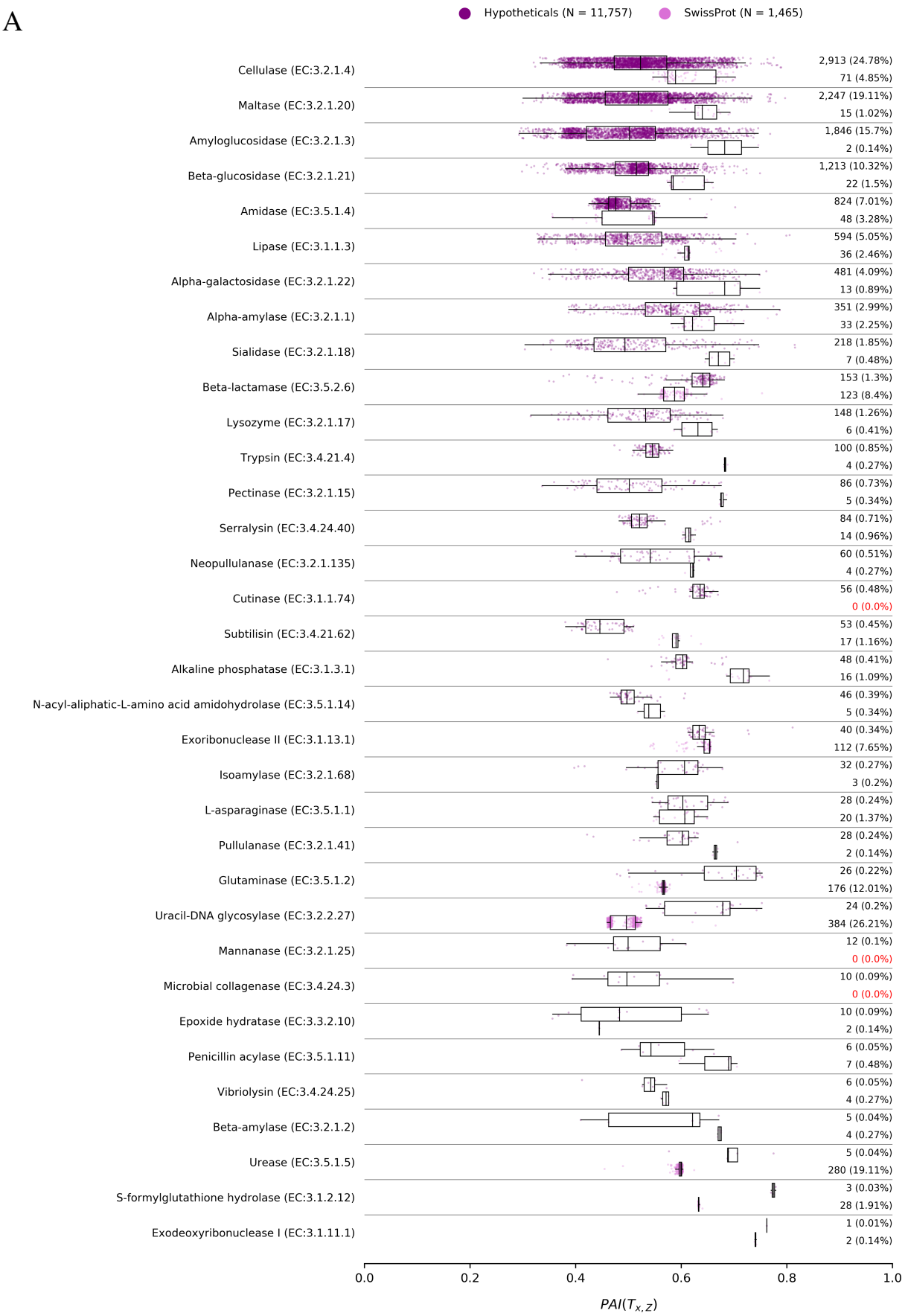


Figure 4.1. (Continued on next page.)

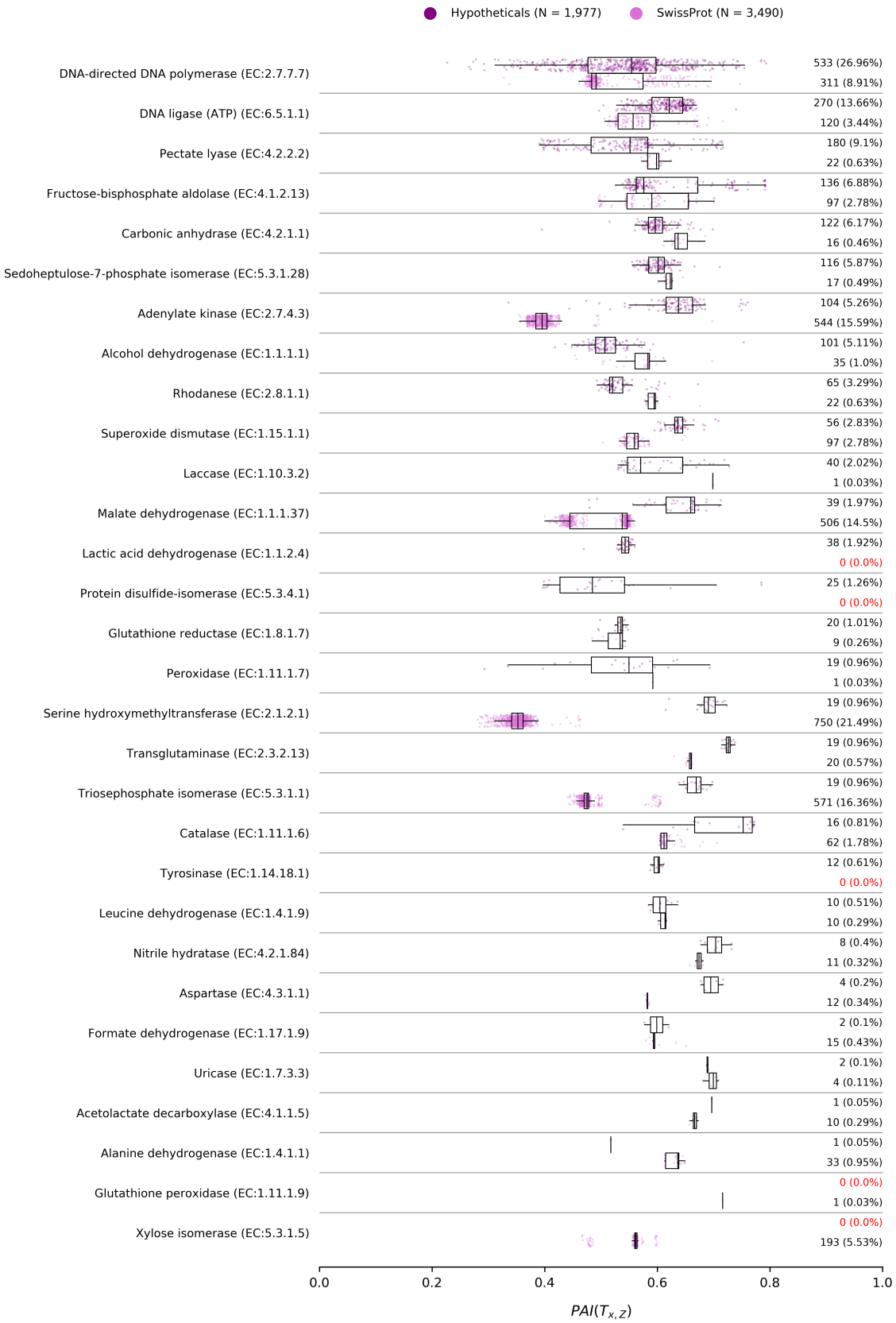


Figure 4.1: Subset of enzymes with known biotechnological potential; from each dataset; with EC, GO, and CD annotation; whose value of $PAI > \tau$; and whose 4th-digit EC is represented in Table 1.1. Only the proteins with a single EC assignment were considered. These proteins are grouped by their 4th-digit EC number. Each row refers to a distinct EC number. The number of proteins from each dataset is shown to the right of each boxplot. The number "0" is shown in red if a dataset lacks protein representatives for an EC number. N: total number of proteins from each dataset shown in each figure that met the selection criteria. The percentages were calculated in relation to each N. (**A**) Hydrolases; (**B**) other enzyme classes.

The latter is due to the fact that SwissProt lacks protein representatives for these EC numbers issuing from the Archaea or the Bacteria domains [1,2,3].

Conversely, there are 2 EC numbers with proteins from the SwissProt BPS, but with no representation from the Hypotheticals BPS (Figure 4.1.B). These ECs are Glutathione peroxidase (EC:1.11.1.9), and Xylose isomerase (EC:5.3.1.5). We further queried the full extent of proteins with EC annotation from the Hypotheticals dataset and found no representatives for either EC.

The antioxidant activity of some glutathione peroxidases depends on the presence of a selenocysteine (SeCys) residue at the catalytic site [48, 49]. Upon retrieving protein sequences from our data sources, we selected for those absent of ambiguous or rare amino-acid codes, including SeCys (see section 2.2). This might explain the lack of glutathione peroxidases in the Hypotheticals dataset. Yet, it does not explain the presence of a glutathione peroxidase from SwissProt—which was also subject to our filtering criteria (see section 2.2). We posited that this glutathione peroxidase must be an isozyme without a SeCys residue. Upon inspecting its sequence we observed that it does not contain SeCys [4].

Xylose isomerase (also known as Glucose isomerase [50]), is ubiquitous among prokaryotes [51]. Additionally, this enzyme is extensively used in the industry to produce high fructose corn syrup [24]. We speculate that this enzyme is absent from the Hypotheticals dataset because it has been comprehensively studied. Therefore, orthologs of this enzyme are unlikely to elude annotation by public databases.

4.3.2 Cutinases: a case study

We have set to study the 56 putative cutinases issuing from the BPS of the Hypotheticals dataset (Figure 4.1.A). We wanted to know what was the most-informative CD for each of these proteins—i.e., that upon which the $Q \circ PIC_S(T_{x,CD})$ value was calculated.

We queried this BPS and grouped the 56 putative cutinases by their most-informative CD (Table 4.1). For 48 out of 56 putative cutinases the most-informative CD was indeed "Cutinase". This suggests a consensus between the predicted 4th-digit EC number (i.e., EC:3.1.1.74) and the most-informative CD. The remaining proteins had different most-informative CDs. Even

[1] https://www.uniprot.org/uniprot/?query=EC%3A3.2.1.25+AND+reviewed%3Ayes&sort=score
[2] https://www.uniprot.org/uniprot/?query=EC%3A1.1.2.4+AND+reviewed%3Ayes&sort=score
[3] https://www.uniprot.org/uniprot/?query=EC%3A5.3.4.1+AND+reviewed%3Ayes&sort=score
[4] https://www.uniprot.org/uniprot/O08368.fasta

so, 5 out of the remaining 8 assignments are related to cutinase-like activity.

Some authors characterize cutinases as intermediates between lipases and esterases, given their activity towards soluble esters and lipids [30]. Therefore, we posit that the 3 assignments for "Triacylglycerol esterase/lipase EstA, alpha/beta hydrolase fold" are reasonable. Additionally, cutinases belong to the family of α/β hydrolases [30]; explaining the assignment for the protein whose most-informative CD is "Alpha/beta hydrolase family".

Moreover, thioesterase domains are homologous to α/β hydrolases [52], and they also share the same catalytic triad consisting of S-H-D residues [52]. Thus justifying the presence of the "Surfactin synthase thioesterase subunit" CD assignment within this subset of putative cutinases.

As a next step, we wanted to understand whether these putative cutinases were related to bona fide PET hydrolases. To do so, we gathered 30 reported PET hydrolase protein sequences by their access numbers in Carr et al. [31], and Kawai et al. [33]. We then selected the 48 putative cutinases that had the "Cutinase" (CDD:307296) most-informative CD assignment (Table 4.1). We only considered these 48 sequences to ensure stringency in the downstream results. Afterwards we generated a pairwise distance matrix of amino-acid identity (AAI) for all 78 proteins, in order to see how they related to one another (Figure 4.2).

By subsetting the dendrograms in Figure 4.2 at the third hierarchical level—in a top to bottom direction—we can distinguish 3 clusters and 1 singleton. We reckon that subsetting the dendrograms at this level is reasonable. Our choice appears to be supported by the clustering patterns in the heatmap. From left to right, we can notice that the first cluster is composed by two *Clostridium botulinum* esterases Cbotu_EstA and Cbotu_EstB. These two esterases share $\geqslant$ 40% AAI. The singleton consists of *Bacillus subtilis* esterase BsEstB. This esterase shows $<$ 20% AAI to every other protein sequence in the heatmap.

Table 4.1: Most-informative Conserved Domains for 56 putative Cutinases from the Hypotheticals dataset BPS.

CD	description	counts
CDD:307296	Cutinase	48
CDD:224001	Triacylglycerol esterase/lipase EstA, alpha/beta hydrolase fold	3
CDD:215128	methyl indole-3-acetate methyltransferase	2
CDD:225749	Surfactin synthase thioesterase subunit	1
CDD:227020	Microcystin-dependent protein (function unknown)	1
CDD:315383	Alpha/beta hydrolase family	1

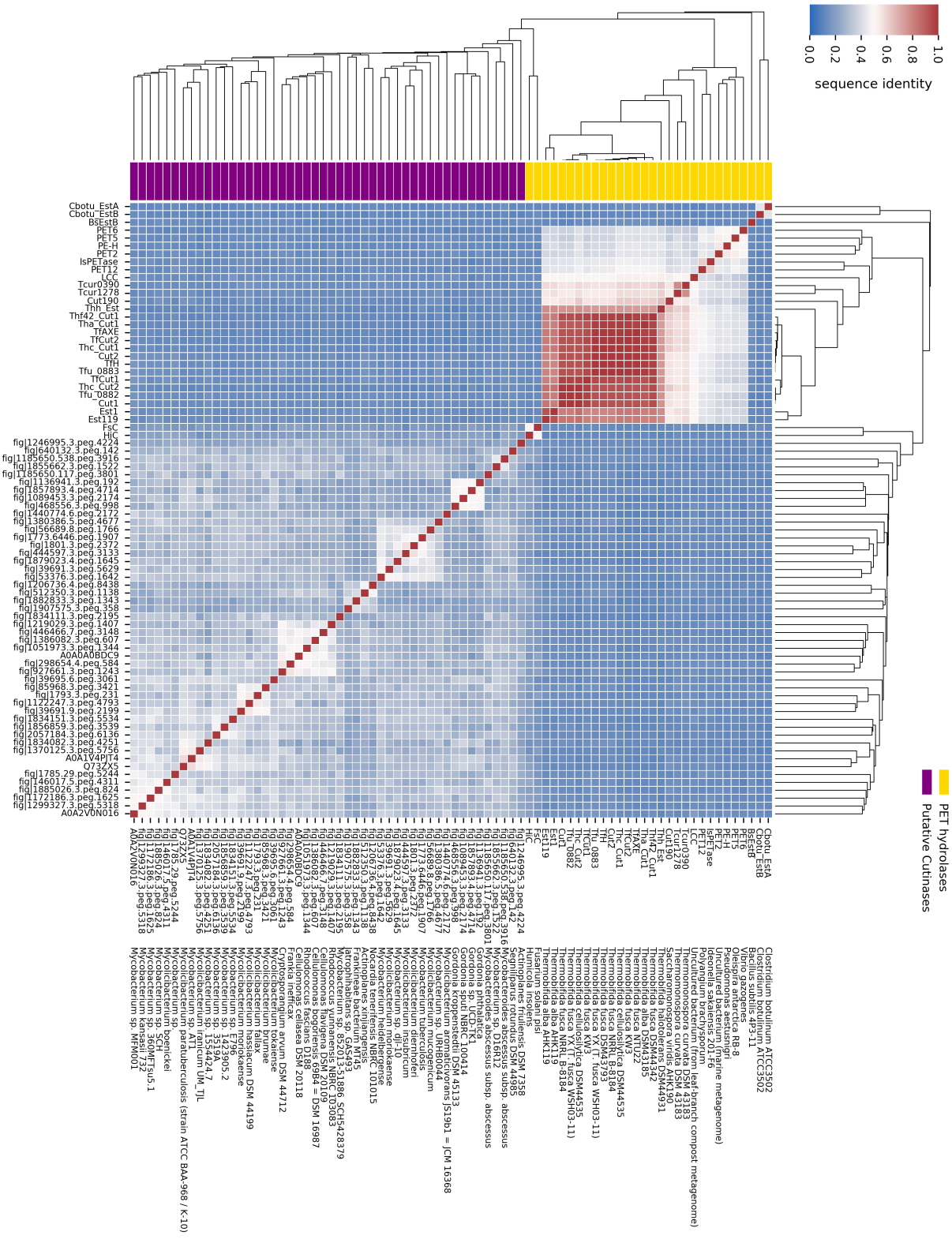


Figure 4.2: Amino-acid sequence identity heatmap of 78 cutinases. There are 48 putative cutinases from the Hypotheticals dataset (purple label), and 30 reported PET hydrolases gathered from the bibliography [31, 33] (yellow label). The species/strain of origin is shown to the right of the y-axis for each protein. A dendrogram depicting the hierarchical linkage clustering for these sequences is shown on top and to the left of the heatmap. The dendrograms were generated using SciPy's "single" method, also known as "nearest point algorithm".

The second cluster contains the majority of reported PET hydrolases. These PET hydrolases all share $\geqslant$ 40% AAI among themselves. Of particular interest is the subcluster of PET hydrolases issuing from the *Thermobifida* genus, within which all share $\geqslant$ 60% AAI. A similar result has been reported by Carniel et al. [53].

The third and last cluster encloses the 48 putative cutinases. This major cluster consists of remotely-related sequences ($\leqslant$ 50% AAI). Within this cluster there appear to be 8 sub-clustering patterns. Some of these patterns seem to correlate with the taxa from which these proteins originate—e.g., the *Gordonia* genus sub-cluster, and the sub-clusters containing both *Mycobacterium* and *Mycolicibacterium*, minding their recent demarcation as different genera (see [54]). Moreover, all 48 putative cutinases issue from the phylum Actinobacteria. This result is congruent with previous reports. These reports state that to date most bacterial isolates with PET hydrolytic potential are found within this phylum [31, 34, 55, 56].

To our surprise, both the *Fusarium solani pisi* cutinase (FsC), and the *Humicola insolens* cutinase (HiC) appear to be clustered together with the 48 putative cutinases. Albeit sharing low amino-acid identity (20% $\leqslant$ AAI $<$ 40%) with the rest of this cohort.

To date, FsC and HiC are the only known PET hydrolases issuing from the domain Eukarya [53]. Furthermore, FsC and HiC seem to share greater AAI with the 48 putative cutinases, than with the remaining members of the reported PET hydrolases. This is an interesting result because Carniel et al. previously reported that both FsC and HiC also stood out in their analyses [53]. According to the authors, this was due to the fact that these cutinases were more similar to one another than to the remainder PET hydrolases of bacterial origin [53].

The PET hydrolytic activities of both FsC and HiC have been formerly assessed by Ronkvist et al. [57]. HiC showed the ability to completely degrade low crystallinity PET, with near 100% weight loss within 96h [57]. Whereas FsC achieved 5% weight loss on the same substrate [57]. In addition, HiC was also reported as being capable of hydrolysing crystalline PET ([57], reviewed in [31, 33]).

Intrigued by the fact that these functionally-verified PET hydrolases clustered together with the 48 putative cutinases, we chose to select both HiC and FsC as references for further inquiries. We were specially interested in knowing whether the 48 putative cutinases shared the same motifs as these two cutinases. To this end we performed a Multiple Sequence Alignment (MSA) among HiC, FsC, and the 48 putative cutinases (Figure 4.3).

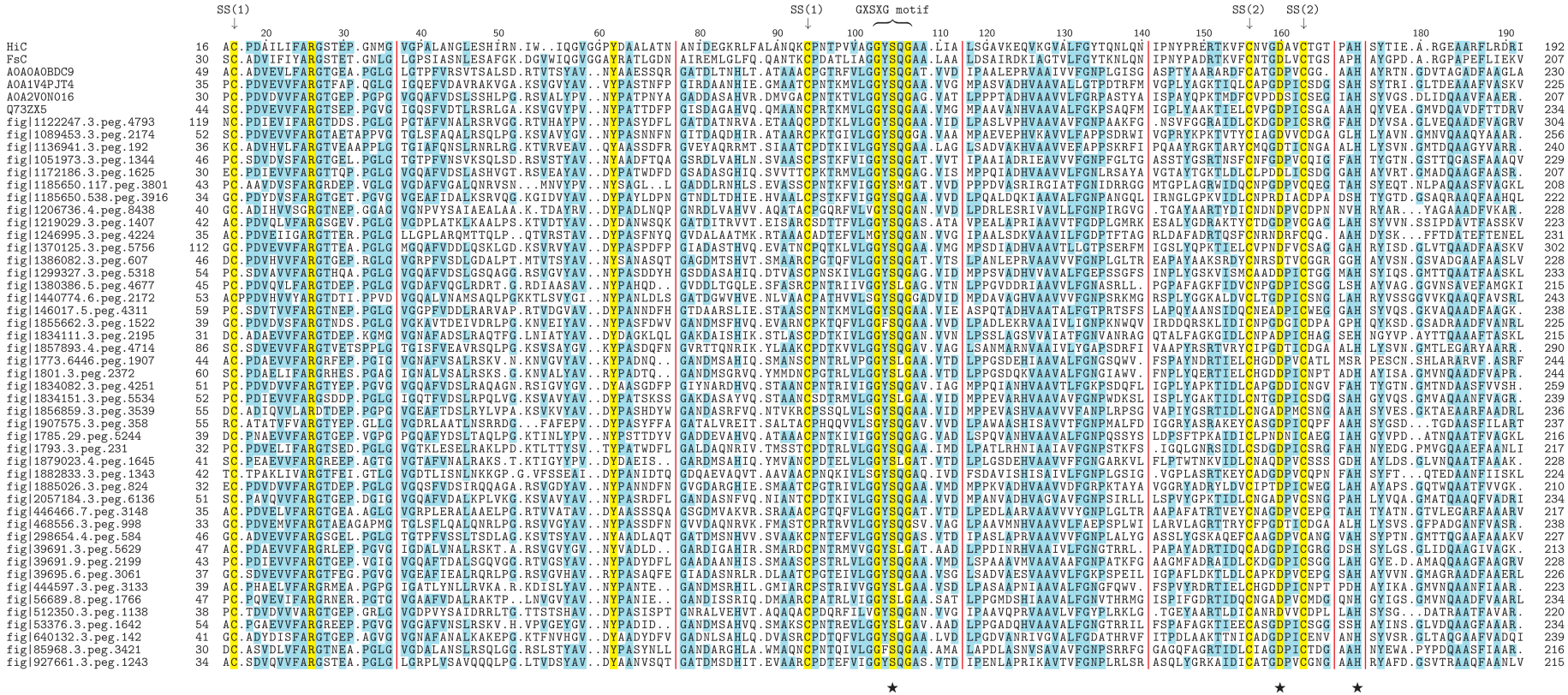


Figure 4.3: Multiple Sequence Alignment (MSA) among 48 putative cutinases and two PET hydrolases: HiC and FsC. The range shown refers to the amino-acid coordinates for the cutinase domain of HiC. The ruler on top denotes the residue number for the HiC sequence. The residue numbers where the alignment started and ended are shown to the right and to the left of each sequence, respectively. Alignment sites with low occupancy were omitted, and the red vertical lines show where these truncations took place. Residues that are conserved for all sequences are highlighted in yellow, and those that are conserved for $\geqslant 50\%$ of sequences are highlighted in blue. Amino-acids comprising the catalytic triad are annotated with a star. SS: disulfide bridge.

The S-D-H catalytic triad is conserved in all 48 putative cutinases (Figure 4.3). There is a single sequence (`fig|1773.6446.peg.1907`) whose catalytic Histidine residue did not align with the remaining sequences in the MSA. The GXSXG pentapeptide motif enclosing the nucleophilic serine is fully conserved in all sequences. This motif contains the same five residues as both HiC and FsC (GYSQG) for 36 (75%) out of 48 putative cutinases. There are also four cysteine residues that are fully conserved in all sequences. These cysteine residues are thought to form disulfide bonds in cutinases [58], and disulfide bonds are known to enhance protein thermostability [59]. Carr et al. outline that functionally-verified PET hydrolases of bacterial origin have a C-terminal disulfide bond, reported to confer thermal and kinetic stability to these enzymes [31]. Moreover, FsC is reported to achieve optimal performance at 50°C [57]. On the upside, HiC maintains maximum initial activity between 70°C and 80°C [57]. Given the conservation of these cysteine residues likely involved in disulfide bond formation, across all 48 putative cutinases, we argue that it is reasonable to hypothesize that the latter might show some degree of thermostability akin to that of FsC or HiC. Likewise, minding the conservation of key residues such as those of the catalytic triad and the pentapeptide motif, we also posit that the tertiary structure of these 48 putative cutinases might resemble that of FsC and/or HiC.

Next we surveyed the physico-chemical properties of the 48 putative cutinases, together with HiC and FsC. For each of these protein sequences, we computed the (i) aromaticity coefficient; (ii) grand average of hydropathicity (GRAVY); (iii) instability index; (iv) isoelectric point (pI); (v) molecular weight (MW); and (vi) sequence length. The values for each of these properties, and for each protein sequence are shown in Figure 4.4.

Most ($\sim$75%) putative cutinases have a sequence length of < 250 a.a, although the maximum is near 350 a.a. This contrasts with the sequence length of HiC (194 a.a.)—which is the minimum value—but not as much with the length of FsC (214 a.a.); given that there are putative cutinases whose sequence length is $\leqslant 214$ a.a. Analogously, most putative cutinases have a MW of < 26 kDa; with the maximum reaching near 35 kDa. The minimum value for MW is that of HiC (20 kDa); and there are putative cutinases whose MW is equal or less than that of FsC (22 kDa).

Increased positive values of GRAVY suggest increased hydrophobicity [45]. In this regard, both FsC and HiC are on the hydrophilic end of the distribution, with a GRAVY score of -0.1196 and -0.0335, respectively. On the other hand, 47 putative cutinases are more hydrophobic than FsC, and 46 are more hydrophobic than HiC. The minimum GRAVY score is

shown by a putative cutinase with a value of -0.2182, being more hydrophilic than both FsC and HiC. It should be emphasized that GRAVY estimates the global hydrophobicity of a protein from its entire amino-acid sequence [60]. It does not take into account the tertiary structure of the protein, nor the amount of residues in its hydrophobic core [60].

Therefore, the increased hydrophobicity portrayed by most putative cutinases, in comparison to FsC and HiC, might be implicated with the existence of a larger hydrophobic core, which skewes the GRAVY score towards more positive values. Alternatively, it might be due to the fact that most putative cutinases have a sequence length that is greater than that of HiC or FsC, coupled with the observation that whereas the cutinase domain of HiC starts at residue number 16 and spans up to residue number 192 (see Figure 4.3); some putative cutinases have the same cutinase domain starting at residue number 30 or even 119. This entails a greater number of amino-acids that do not belong to the cutinase domain to be used as input for the GRAVY score calculation for these proteins, and thus possibly generating less accurate predictions.

The pI corresponds to the pH at which a given protein has a net charge of zero [61]. While FsC shows a pI $\approx$ 6.79, and HiC a pI = 8.43, most putative cutinases show either a pI $\leqslant$ 6, or pI $\geqslant$ 8.5; with the maximum and minimum values being pI $\approx$ 4 and pI $\approx$ 10, respectively. Even though protein stability at a given pH depends not only on its net charge, but also on the difference in pKs (i.e., the negative logarithm of the dissociation constant K) of the ionizable groups among the folded and the unfolded protein [62]; this result hints at the fact that these putative cutinases might have either a more acidic, or a more basic optimal pH than those of the two chosen references. In theory, these putative cutinases could be promising candidates for biotechnological applications, minding that they might withstand harsher industrial conditions (see [12]), than those allowed for FsC or HiC.

As a rule of thumb, proteins with an instability index below 40 are probably stable *in vitro* [63]. Most (79%) putative cutinases have an instability index $<$ 40, together with FsC and HiC. Thus suggesting that theoretically they would be stable *in vitro*. Moreover, while FsC and HiC have an aromaticity ratio of $\sim$6%, the putative cutinases have wider-spread values, ranging from 4.6% to 9.2%. This greater spread of aromaticity ratios, as portrayed by the putative cutinases, might also be related with the hypothetical situation proposed above, where lengthier sequences might show different ratios of aromatic residues that might not obligatorily fall within the range of the cutinase domain.

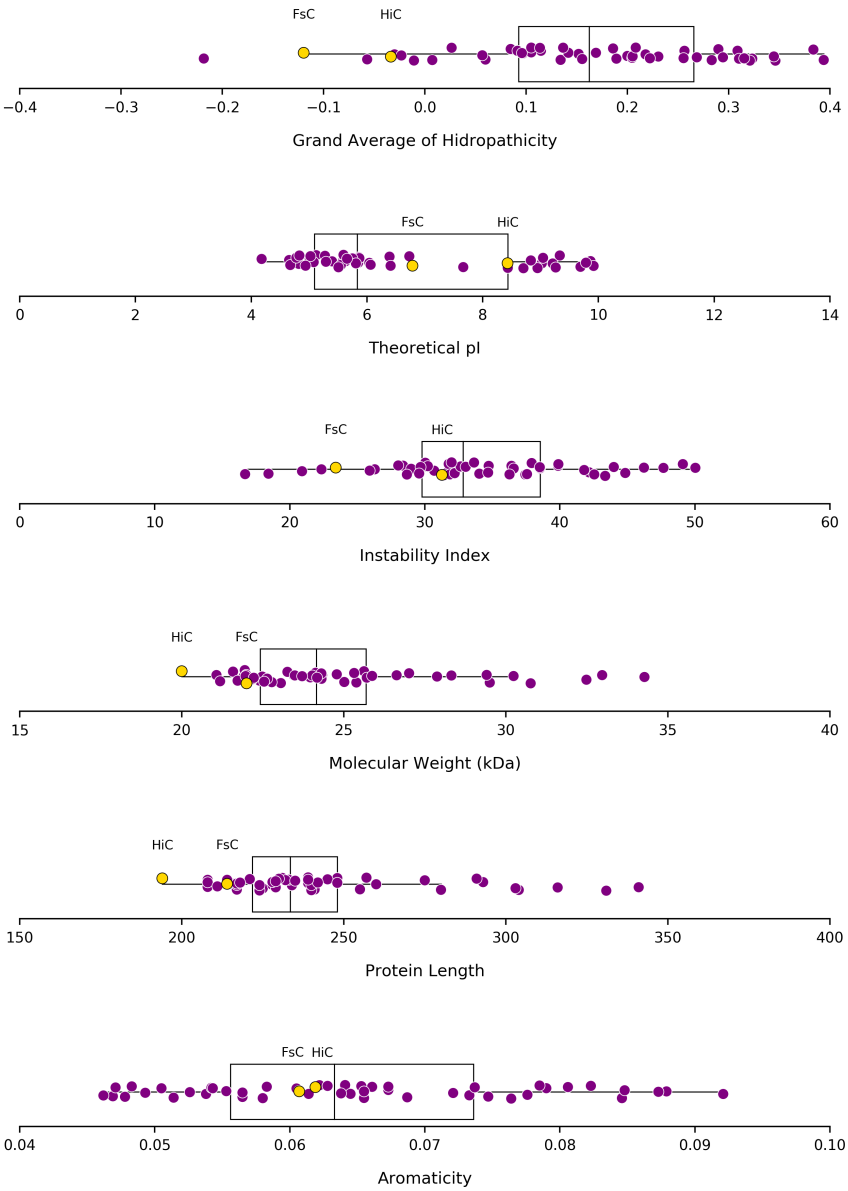


Figure 4.4: *In silico*-generated physico-chemical properties for the putative cutinases. There are 48 putative cutinases from the Hypotheticals dataset BPS (purple), and two PET hydrolases (yellow). The data-points referring to HiC and FsC are explicitly labeled for each property shown.

Reminiscing on the work developed throughout chapter 2, one might recall that these putative cutinases are in fact cluster representative sequences at a 60% global identity resolution. Thus, for these 48 cluster representatives, there are a total of 254 cluster members (data not shown) that share $\geqslant$ 60% sequence identity with their respective cluster representative. We tentatively suggest that it is reasonable for these cluster members to inherit the same annotations—and consequently the same predicted molecular function—as their representatives. Thus, in this work we report 254 candidate proteins that might be at a prime for further data analysis and scrutiny, and possibly future *in vitro* assays in the search for novel PET hydrolases.

4.4 Conclusions

Starting with a dataset comprising *de novo*-annotated (see chapter 2) protein sequences of prokaryotic origin, we first created a subset enclosing the most extensively annotated ones. We ensured that their Protein Annotation Index (*PAI*) value was greater than the partitioning threshold (see chapter 3), and that they had been annotated by all three Classification Systems—i.e., EC, GO and CD.

We then selected the sequences whose predicted enzyme subclass (i.e., 4th-digit EC) was of known biotechnological and/or industrial interest (see Table 1.1). This enabled us to disclose 64 enzyme subclasses, amounting to a total of 13,734 protein cluster representatives. The *PAI* values for the sequences issuing from 50 out of 64 enzyme subclasses exhibited greater variance and greater maxima than their counterparts issuing from SwissProt.

From the foregoing 64 enzyme subclasses, 34 belonged to the Hydrolase class (i.e., EC:3.-.-.-), totaling 11,757 protein cluster representatives. We report that for 27 out of these 34 Hydrolase subclasses, our dataset contains eight times more potential Hydrolases than the gold-standard cohort. Additionally, we report 3 Hydrolase subclasses with prokaryotic protein sequence representatives—Mannanase (EC:3.2.1.25); Lactic acid dehydrogenase (EC:1.1.2.4); and Protein disulfide-isomerase (EC:5.3.4.1)—that to our knowledge, do not have representatives in SwissProt issuing from either Archaea or Bacteria domains. These 13,734 putative enzymes represent prime candidates of recognized industrial interest. These putative enzymes might be of interest to prospective researchers aiming to characterize them to a greater extent.

Among 34 Hydrolase subclasses, we chose that of Cutinases (EC:3.1.1.74) as a case study. We chose to investigate the Cutinase subclass based on the catalytic versatility of its members [34], industrial pluripotentiality [32], and reported ability to hydrolize high molecular weight polyesters, like polyethylene terephthalate (PET) [31]. We report 48 novel putative cutinases that possessed the characteristic cutinase CD. These held the most promise as their annotation depicted a consensus between the EC and CD classification systems. We also account that all 48 putative cutinases issue from the phylum Actinobacteria. This result is in agreement with previous reports.

Afterwards, we compared the foregoing 48 putative cutinases with 30 bona fide PET hydrolases. Surprisingly, we found that the PET hydrolases that shared the greatest amino-acid identity to these 48 putative cutinases were of fungal origin. These were the *Fusarium solani*

pisi cutinase (FsC), and the *Humicola insolens* cutinase (HiC). Upon performing a Multiple Sequence Alignment (MSA) among HiC, FsC, and the 48 putative cutinases, we were able to confirm that the S-D-H active site is present throughout all sequences. The characteristic pentapeptide motif was highly conserved, and 36 (75%) out of 48 putative cutinases contained the same five residues in this motif as both HiC and FsC. Four cysteine residues involved in disulfide bond formation were also highly conserved. The conservation of these motifs suggested three things. First, that these 48 putative cutinases are possibly functional. Second, that they might be thermostable to some extent. Third, that the tertiary structure of these 48 putative cutinases might resemble that of FsC or HiC.

Lastly, upon calculating several physico-chemical properties for these 48 putative cutinases, we suggest that these might withstand harsher industrial conditions than those allowed for FsC or HiC. We conclude that these putative cutinases might be promising candidates for biotechnological applications, and further disclose that these 48 cluster representative sequences relate to a total of 254 cluster members. These 254 protein sequences might be at a prime for further analysis in the search for novel PET hydrolases. Moreover, the workflow we presented in this chapter might be used to further characterize any of the putative enzyme subclasses disclosed herein.

References

[1] Piao H, Froula J, Du C, Kim TW, Hawley ER, Bauer S, et al. Identification of novel biomass-degrading enzymes from genomic dark matter: Populating genomic sequence space with functional annotation. Biotechnol Bioeng. 2014;111(8):1550–1565.

[2] Al-Shahib A, Breitling R, Gilbert DR. Predicting protein function by machine learning on amino acid sequences – a critical evaluation. BMC Genomics. 2007;8(1):78.

[3] McLean JS, Lombardo MJ, Badger JH, Edlund A, Novotny M, Yee-Greenbaum J, et al. Candidate phylum TM6 genome recovered from a hospital sink biofilm provides genomic insights into this uncultivated phylum. Proc Natl Acad Sci U S A. 2013;110(26):E2390–9.

[4] Marcy Y, Ouverney C, Bik EM, Lösekann T, Ivanova N, Martin HG, et al. Dissecting biological "dark matter" with single-cell genetic analysis of rare and uncultivated TM7 microbes from the human mouth. Proc Natl Acad Sci U S A. 2007;104(29):11889–11894.

[5] Garza DR, Dutilh BE. From cultured to uncultured genome sequences: metagenomics and modeling microbial ecosystems. Cell Mol Life Sci. 2015;72(22):4287–4308.

[6] Becraft ED, Dodsworth JA, Murugapiran SK, Ohlsson JI, Briggs BR, Kanbar J, et al. Single-Cell-Genomics-Facilitated Read Binning of Candidate Phylum EM19 Genomes from Geothermal Spring Metagenomes. Appl Environ Microbiol. 2015;82(4):992–1003.

[7] Bernard G, Pathmanathan JS, Lannes R, Lopez P, Bapteste E. Microbial dark matter investigations: how microbial studies transform biological knowledge and empirically sketch a logic of scientific discovery. Genome Biol Evol. 2018;.

[8] Lobb B, Kurtz DA, Moreno-Hagelsieb G, Doxey AC. Remote homology and the functions of metagenomic dark matter. Front Genet. 2015;6:234.

[9] Mokili JL, Rohwer F, Dutilh BE. Metagenomics and future perspectives in virus discovery. Curr Opin Virol. 2012;2(1):63–77.

[10] Dutilh BE. Metagenomic ventures into outer sequence space. Bacteriophage. 2014;4(4):e979664.

[11] Rashid M, Stingl U. Contemporary molecular tools in microbial ecology and their application to advancing biotechnology. Biotechnol Adv. 2015;33(8):1755–1773.

[12] Sysoev M, Grötzinger SW, Renn D, Eppinger J, Rueping M, Karan R. Bioprospecting of Novel Extremozymes From Prokaryotes-The Advent of Culture-Independent Methods. Front Microbiol. 2021;12:630013.

[13] Chen R, Wong HL, Burns BP. New Approaches to Detect Biosynthetic Gene Clusters in the Environment. Medicines (Basel). 2019;6(1).

[14] Verma S, Meghwanshi GK, Kumar R. Current perspectives for microbial lipases from extremophiles and metagenomics. Biochimie. 2021;182:23–36.

[15] Akal AL, Karan R, Hohl A, Alam I, Vogler M, Grötzinger SW, et al. A polyextremophilic alcohol dehydrogenase from the Atlantis II Deep Red Sea brine pool. FEBS Open Bio. 2019;9(2):194–205.

[16] Grötzinger SW, Karan R, Strillinger E, Bader S, Frank A, Al Rowaihi IS, et al. Identification and Experimental Characterization of an Extremophilic Brine Pool Alcohol Dehydrogenase from Single Amplified Genomes. ACS Chem Biol. 2018;13(1):161–170.

[17] Singh BK. Organophosphorus-degrading bacteria: ecology and industrial applications. Nat Rev Microbiol. 2009;7(2):156–164.

[18] Smith HO, Wilcox KW. A restriction enzyme from Hemophilus influenzae. I. Purification and general properties. J Mol Biol. 1970;51(2):379–391.

[19] Brock TD, Freeze H. Thermus aquaticus gen. n. and sp. n., a nonsporulating extreme thermophile. J Bacteriol. 1969;98(1):289–297.

[20] Saiki RK, Gelfand DH, Stoffel S, Scharf SJ, Higuchi R, Horn GT, et al. Primer-directed enzymatic amplification of DNA with a thermostable DNA polymerase. Science. 1988;239(4839):487–491.

[21] Jinek M, Chylinski K, Fonfara I, Hauer M, Doudna JA, Charpentier E. A programmable dual-RNA-guided DNA endonuclease in adaptive bacterial immunity. Science. 2012;337(6096):816–821.

[22] Li S, Yang X, Yang S, Zhu M, Wang X. Technology prospecting on enzymes: application, marketing and engineering. Comput Struct Biotechnol J. 2012;2:e201209017.

[23] Bruno S, Coppola D, di Prisco G, Giordano D, Verde C. Enzymes from Marine Polar Regions and Their Biotechnological Applications. Mar Drugs. 2019;17(10).

[24] Singh R, Kumar M, Mittal A, Mehta PK. Microbial enzymes: industrial progress in 21st century. 3 Biotech. 2016;6(2):174.

[25] Gurung N, Ray S, Bose S, Rai V. A broader view: microbial enzymes and their relevance in industries, medicine, and beyond. Biomed Res Int. 2013;2013:329121.

[26] Robinson PK. Enzymes: principles and biotechnological applications. Essays Biochem. 2015;59:1–41.

[27] Cabrera MÁ, Blamey JM. Biotechnological applications of archaeal enzymes from extreme environments. Biol Res. 2018;51(1):37.

[28] Ramesh A, Harani Devi P, Chattopadhyay S, Kavitha M. Commercial Applications of Microbial Enzymes. In: Microorganisms for Sustainability. Singapore: Springer Singapore; 2020. p. 137–184.

[29] Meghwanshi GK, Kaur N, Verma S, Dabi NK, Vashishtha A, Charan PD, et al. Enzymes for pharmaceutical and therapeutic applications. Biotechnol Appl Biochem. 2020;67(4):586–601.

[30] Nyyssölä A. Which properties of cutinases are important for applications? Appl Microbiol Biotechnol. 2015;99(12):4931–4942.

[31] Carr CM, Clarke DJ, Dobson ADW. Microbial Polyethylene Terephthalate Hydrolases: Current and Future Perspectives. Front Microbiol. 2020;11:571265.

[32] Pio TF, Macedo GA. Cutinases: properties and industrial applications. Adv Appl Microbiol. 2009;66:77–95.

[33] Kawai F, Kawabata T, Oda M. Current knowledge on enzymatic PET degradation and its possible application to waste stream management and other fields. Appl Microbiol Biotechnol. 2019;103(11):4253–4268.

[34] Danso D, Chow J, Streit WR. Plastics: Environmental and Biotechnological Perspectives on Microbial Degradation. Appl Environ Microbiol. 2019;85(19).

[35] Tournier V, Topham CM, Gilles A, David B, Folgoas C, Moya-Leclair E, et al. An engineered PET depolymerase to break down and recycle plastic bottles. Nature. 2020;580(7802):216–219.

[36] Ragaert K, Delva L, Van Geem K. Mechanical and chemical recycling of solid plastic waste. Waste Manag. 2017;69:24–58.

[37] Yoshida S, Hiraga K, Takehana T, Taniguchi I, Yamaji H, Maeda Y, et al. A bacterium that degrades and assimilates poly(ethylene terephthalate). Science. 2016;351(6278):1196–1199.

[38] Kulmanov M, Khan MA, Hoehndorf R, Wren J. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. Bioinformatics. 2018;34(4):660–668.

[39] Ryu JY, Kim HU, Lee SY. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. Proc Natl Acad Sci U S A. 2019;116(28):13996–14001.

[40] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–3402.

[41] Sievers F, Wilm A, Dineen D, Gibson TJ, Karplus K, Li W, et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. Mol Syst Biol. 2011;7:539.

[42] Notredame C, Higgins DG, Heringa J. T-Coffee: A novel method for fast and accurate multiple sequence alignment. J Mol Biol. 2000;302(1):205–217.

[43] Beitz E. TEXshade: shading and labeling of multiple sequence alignments using LATEX2 epsilon. Bioinformatics. 2000;16(2):135–139.

[44] Cock PJA, Antao T, Chang JT, Chapman BA, Cox CJ, Dalke A, et al. Biopython: freely available Python tools for computational molecular biology and bioinformatics. Bioinformatics. 2009;25(11):1422–1423.

[45] Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol. 1982;157(1):105–132.

[46] Lobry JR, Gautier C. Hydrophobicity, expressivity and aromaticity are the major trends of amino-acid usage in 999 Escherichia coli chromosome-encoded genes. Nucleic Acids Res. 1994;22(15):3174–3180.

[47] Guruprasad K, Reddy BV, Pandit MW. Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. Protein Eng. 1990;4(2):155–161.

[48] Margis R, Dunand C, Teixeira FK, Margis-Pinheiro M. Glutathione peroxidase family - an evolutionary overview. FEBS J. 2008;275(15):3959–3970.

[49] Herbette S, Roeckel-Drevet P, Drevet JR. Seleno-independent glutathione peroxidases. More than simple antioxidant scavengers. FEBS J. 2007;274(9):2163–2180.

[50] Saha BC, Jordan DB, Bothast RJ. Enzymes, Industrial (overview). In: Encyclopedia of Microbiology. Elsevier; 2009. p. 281–294.

[51] Bhosale SH, Rao MB, Deshpande VV. Molecular and industrial aspects of glucose isomerase. Microbiol Rev. 1996;60(2):280–300.

[52] Montavon TJ, Bruner SD. Nonribosomal Peptide Synthetases. In: Comprehensive Natural Products II. Elsevier; 2010. p. 619–655.

[53] Carniel A, Waldow VdA, Castro AMd. A comprehensive and critical review on key elements to implement enzymatic PET depolymerization for recycling purposes. Biotechnol Adv. 2021;52:107811.

[54] Gupta RS, Lo B, Son J. Phylogenomics and Comparative Genomic Studies Robustly Support Division of the Genus into an Emended Genus and Four Novel Genera. Front Microbiol. 2018;9:67.

[55] Danso D, Schmeisser C, Chow J, Zimmermann W, Wei R, Leggewie C, et al. New Insights into the Function and Global Distribution of Polyethylene Terephthalate (PET)-Degrading Bacteria and Enzymes in Marine and Terrestrial Metagenomes. Appl Environ Microbiol. 2018;84(8).

[56] Herrero Acero E, Ribitsch D, Steinkellner G, Gruber K, Greimel K, Eiteljoerg I, et al. Enzymatic surface hydrolysis of PET: Effect of structural diversity on kinetic properties of cutinases from Thermobifida. Macromolecules. 2011;44(12):4632–4640.

[57] Ronkvist ÅM, Xie W, Lu W, Gross RA. Cutinase-catalyzed hydrolysis of poly(ethylene terephthalate). Macromolecules. 2009;42(14):5128–5138.

[58] Masaki K, Kamini NR, Ikeda H, Iefuji H. Cutinase-like enzyme from the yeast Cryptococcus sp. strain S-2 hydrolyzes polylactic acid and other biodegradable plastics. Appl Environ Microbiol. 2005;71(11):7548–7550.

[59] Roth C, Wei R, Oeser T, Then J, Föllner C, Zimmermann W, et al. Structural and functional studies on a thermostable polyethylene terephthalate degrading hydrolase from Thermobifida fusca. Appl Microbiol Biotechnol. 2014;98(18):7815–7823.

[60] Dyson MR, Shadbolt SP, Vincent KJ, Perera RL, McCafferty J. Production of soluble mammalian proteins in Escherichia coli: identification of protein features that correlate with successful expression. BMC Biotechnol. 2004;4:32.

[61] Kirkwood J, Hargreaves D, O'Keefe S, Wilson J. Using isoelectric point to determine the pH for initial protein crystallization trials. Bioinformatics. 2015;31(9):1444–1451.

[62] Shaw KL, Grimsley GR, Yakovlev GI, Makarov AA, Pace CN. The effect of net charge on the solubility, activity, and stability of ribonuclease Sa. Protein Sci. 2001;10(6):1206–1215.

[63] Gamage DG, Gunaratne A, Periyannan GR, Russell TG. Applicability of Instability Index for In vitro Protein Stability Prediction. Protein Pept Lett. 2019;26(5):339–347.

Chapter 5

General Discussion and Prospective Work

5.1 “What did we create?”

This thesis was composed of three core objectives. The first was to quantify the protein sequences of prokaryotic origin with unknown molecular function from public databases, and create a repository populated with this data. The second objective was to classify these protein sequences, using both well established, as well as state-of-the-art function prediction tools. The third objective, although manifold, was primarily concerned with validating the classification that had taken place.

5.1.1 Quantification

Apart from quantifying the amount of protein sequences of unknown function, we also sought to store all data in a in-house repository. This step was necessary so that the data shared an unified format and stood under a common architecture. On a technical note, we opted for a NoSQL data-framework (i.e., Apache Solr) in order to store, and later retrieve, our data. We chose to do so in order to expedite the creation of our repository. This way we needed to invest less time than that required for the design of a relational data model. In hindsight however, we reckon that a future version of this repository should be stored in a relational database (e.g., MySQL). By creating a relational model *a priori*, one would be facilitating downstream querying, sample-data retrieval, and thus scientific hypothesis testing.

Upon creating this repository, we had gathered 134,894,520 protein sequences whose molecular function could not be inferred by public databases. As a preliminary step, we had also considered including sequence data from publicly-available metagenomes. However, the space and computation time needed in order to store and process such an amount of sequence data promptly led us to discard this possibility. Nevertheless, with decreasing computation and data storage costs, we envision that a future version of this repository might be able to encompass metagenomic sequence data as well. Additionally, it would be interesting to expand the foregoing repository to include uncharacterized protein sequences of fungal and viral origin.

To our knowledge, the repository we created constitutes the first worldwide centralized collection of prokaryotic protein sequences of unknown molecular function. This repository is therefore the first contribution of this thesis. We regard it as a contribution in its own right for the reasons that follow. By creating this repository we curtailed the need for a forthcoming researcher to go through the same obstacles and pitfalls as those we experienced while gathering

this data. Likewise, the sequences comprised therein can be further scrutinized by the scientific community, even if via bulk retrieval of the entire dataset.

On the same note, one should severely emphasize that many so-called "routine" tasks are extremely labor-intensive, and time-consuming. These tasks include, but are not limited to: (i) dissecting unconventional APIs and FTP server directories; (ii) relation-mapping between data-sources; (iii) managing access policies; (iv) troubleshooting broken server connections; (v) creating special rules for data with missing fields; (vi) writing custom scripts to manage downloads programmatically; or to (vii) parse a myriad data formats; among countless others. Practical examples of these tasks are hinted at in the Methods section of Chapters 2 and 4.

Not only are these tasks cumbersome, but the inconsistencies between (and within) data-sources hinder large-scale, cross-platform, computational analyses. Similar predicaments were reported by Schwartz et al [1], back in 2001. In agreement with what these authors proposed, we also urge for the adoption of ubiquitous standards and formats—especially for sequence meta-data—by whomever oughts to provide open sequence data to the public. Reaching a consensus in this matter would greatly expedite future endeavours worldwide.

The last goal concerning this objective was to reduce the dimension of our repository. Choosing the appropriate clustering algorithm was critical, as it would dictate the time spent processing these sequences thereafter. Creating a pairwise distance matrix for $>$134 million protein sequences was unfeasible. Thus, several clustering algorithms had to be disregarded (e.g., distance-based, centroid-based, hierarchical). Our best options lied with "greedy" incremental clustering algorithms. After pondering on a few options (see [2–6]), we chose CD-HIT because it was fairly fast, and because it had become well-established by that time.

A prospective, larger repository might greatly benefit from a two-step clustering approach. In the first step, a very fast clustering algorithm (e.g., Linclust [6]) might be used to partition the data into clusters at a predefined sequence identity threshold. As a second step, a MSA might ensue for each cluster in order to assess the conservation of the sequences within. A MSA reliability measure—e.g., transitive consistency score (TCS)[7]—might also be calculated. This score should suggest whether the clusters are conserved at the given sequence identity threshold. Additionally it might also indicate whether it is reasonable to extend the annotation of the representative to be inferred downstream, to the remaining cluster members.

Following the steps of previous authors [8–10], we clustered our sequences throughout 8

global sequence identity thresholds. This proved useful for four reasons. It allowed us to: (i) create a non-redundant version of the sequence data (i.e., at 95% identity); (ii) assess the diversity of these sequences; (iii) create clusters of putative protein families at 7 different resolutions; and (iv) reduce the size of our repository, without forgoing the relations among these sequences.

The first clustering step enabled us to reduce the size of our repository in 79.51%—i.e., from $>$134 million protein sequences to $\sim$27 million cluster representatives at 95% minimum identity. This result showed two things. First, that it is possible to render our repository by a smaller, non-redundant set of proteins. Second, that there might be a high level of redundancy among (or within) independent data-sources, for protein sequences of unknown function.

The last clustering step allowed us to achieve an additional 54.2% size reduction from the non-redundant set—i.e., from $\sim$27 million cluster representatives at 95% minimum identity, to $\sim$12 million cluster representatives at 60% minimum identity. This implies a $\sim$90.62% size reduction relative to the initial number of $>$134 million sequences. This was an encouraging result. Assuming that the proportion of representative sequences remains at a steady $\sim$10% for larger dataset sizes, together with ever-improving clustering algorithms that provide speedier computation times, future endeavors aiming at the annotation of billions of sequences might be feasible.

One can argue that a size reduction of this magnitude comes at the cost of cluster conservation. However, given that throughout all thresholds there are more singleton representatives than cluster representatives per se; this decrease in conservation only affects the representative sequences issuing from clusters—e.g., 29.28% of representative sequences at 60% minimum identity. Nevertheless, this potential decrease in conservation might be mitigated in the future by implementing the MSA-based clustering routine proposed above.

By meeting this dimension reduction goal we showed that high percentages of dataset size reduction can be achieved. Thus, the computational load required for downstream tasks can be greatly mitigated as a consequence. In addition, all metadata regarding the cluster membership for each sequence throughout all thresholds was appended to our repository. This output might be useful for future researchers aiming to create a sequence similarity network for these unfunctionalized proteins throughout the aforementioned thresholds.

5.1.2 Classification

The second objective of this thesis was to classify the protein sequences in our repository. Classification was done on the dataset clustered at 60% minimum identity, containing 12,654,843 protein sequences of unknown function. We estimated several physico-chemical properties *in silico*. The absence of distinctive patterns among the distributions for the these properties precluded the grouping of proteins according to these attributes. The underlying cause for this occurrence might have been the sheer size of the dataset. Such a large sample size might have contributed to the cluttering of data and the exacerbation of noise from outliers.

In the future, one might benefit from using a different approach. Instead of using physico-chemical properties predicted *in silico* to discriminate between different groups of proteins within the entire dataset; an alternative approach would be to use these properties to differentiate among distinct protein groups within a given molecular function annotation (e.g., a 4th-digit EC). This could be achieved for instance via Principal Component Analysis (PCA).

Nonetheless, the pI appeared to be an exception to this patternless norm. Indeed, our sequences were arranged into two major density curves in relation to their pI, for the pH of 5 and 10, respectively. Even though this is not a original result (see [1]), it nevertheless indicates two things. First, that a forthcoming dataset might be partitioned into two main groups at a very early stage—e.g., upstream from the molecular function imputation step. This dataset might be partitioned according to the theoretical pI of its sequences, property which is likely associated with different protein subcellular locations; namely cytosolic (pI $\approx$ 5) and that of integral membrane proteins (pI $\approx$ 9) [1]. Secondly, the information regarding predicted subcellular locations, by itself, allows one to target each of these main groups with distinct classification approaches accordingly. Therefore, we posit that the computational load and time required to identify enzymes of biotechnological and/or industrial interest can be drastically shortened if one prioritizes the classification of sequences whose theoretical pI is associated with subcellular locations of interest (e.g., cytosolic).

To ascertain the molecular function of these $\sim$12 million proteins, we chose to use both a gold-standard (i.e., RPS-BLAST [11]) as well as state-of-the-art function prediction tools (i.e., DeepGO [12], and DeepEC [13]). We decided to use these three tools, and thus three Classification Systems (i.e., EC, GO and CD). Our reasoning was that they provided non-overlapping and highly complementary annotations to one another.

We were unaware of how would DeepGO and DeepEC perform (or if they would perform at all). Our concerns were threefold. The first concern lied with the fact that these were cutting-edge tools at the time, with the article for DeepGO being published in 2018, and that of DeepEC in 2019. This meant that they had not been benchmarked nor validated by third parties yet. Thus, their reliability lied solely with the account of the authors in each publication.

Secondly, DeepGO was trained on manually-curated protein sequence data [12] from SwissProt [14]. DeepEC was also partially trained using SwissProt (see [13], Supplementary Information). Hence, if the sequences in our dataset were too dissimilar from those in the training sets, there was a chance that these classifiers would not be able to predict their function.

Thirdly, to our knowledge, this work was the first attempt at classifying a large-scale dataset of proteins, whose databases of origin were not able to characterize in the first place. Therefore there was a chance this exploratory endeavor would not bear any fruit. We reasoned that if DeepGO and/or DeepEC were to fail at annotating these proteins, our fallback strategy lied with RPS-BLAST; from which we could gather the CDs for these proteins.

Notwithstanding, we managed to annotate 99.97% (12,651,624) of the dataset with at least one term, from at least one of the Classification Systems used. This was a significant development, given that we started off with unannotated sequence data to begin with. Moreover, 99.91% (12,644,075) of the dataset was annotated with at least one GO term.

This portrayed an extensive annotation coverage by DeepGO. However, the deepest GO term annotation for most of these proteins is either at the first, or second level of the DAG. Only a small fraction of proteins annotated with GO possessed an annotation term whose depth was greater than the third level of the DAG. These results hinted at the vagueness of these GO annotations, as it also suggests permissiveness from DeepGO. Therefore, we might conclude that the annotation coverage provided by DeepGO came at the expense of decreased annotation specificity.

Moreover, we found proteins that had more than one GO term annotation at the first level of the DAG. This result suggests one of two possible scenarios. Either that these proteins perform distinct molecular functions; or alternatively, that DeepGO’s predictions are prone to ambiguity. The fact that the proteins with a single GO assignment appear to have higher prediction scores than those with multiple assignments suggests that the second scenario might be taking place.

Furthermore, nearly all proteins annotated with GO (99.39%) possessed the GO term “bind-

ing" (41.86%), "catalytic activity" (4.7%), or both (52.82%). These results imply that 57.52% of these proteins might be enzymes. Interestingly, the group of proteins with the highest prediction scores are those with "catalytic activity" as their sole GO term at the first level of the DAG. This was a curious result. It suggested that DeepGO's predictions for putative enzymes ranked best. Yet, this account contrasts with the number of proteins annotated by DeepEC—from 12,654,843 representatives, only 2.78% were annotated as possible enzymes. The disparity between the results of both function prediction tools might be explained by DeepEC's conservativeness. Therefore, protein sequences that were too dissimilar from those in its training set might have been disregarded as false negatives. Nonetheless, additional work would be necessary to confirm whether these DeepGO assignments actually relate to highly divergent enzymes, or if they are merely false positives.

Since we gathered these results, the authors of DeepGO have commented on its limitations [15]. These limitations include: (i) the inability to predict the full extent of molecular functions in the GO ontology; (ii) the use of interaction network features, which are absent from uncharacterized proteins; and (iii) the potential overfitting to specific features in the training set, which might lead to inadequate results in real prediction scenarios [15]. Thus, a prospective version of our ensemble framework might benefit from using GO prediction tools that were published since DeepGO (e.g., [15–18]), and that achieved better predictive performance than the latter [15].

RPS-BLAST assigned at least one CD to 48.53% (6,142,383) of these sequences. This annotation coverage fell short of our expectations. We posit that this unexpected coverage might have been the product of a bit-score cutoff that was too stringent. A stringent bit-score cutoff was necessary to mitigate false positives, and ensure the quality of CD annotations. Nonetheless, it might have come at the price of reduced annotation coverage.

It is also possible that the sequences lacking CD annotation might not be proteins at all. Upon gathering our sequence data we only kept sequences from 200 to 10,000 amino-acids in length. Therefore it is unlikely for these sequences to be peptide signaling molecules like auto-inducing peptides (AIPs); or ribosomally synthesized and post-translationally modified peptides (RiPPs), as these are typically < 35 amino-acids [19], and 20 to 110 amino-acids [20] in length, respectively.

As a prospective outlook, we would like to emphasize that there is evidence that small uncharacterized proteins perform a multitude of significant physiological roles yet to be uncharted

[21]. Likewise, small peptides, such as RiPPs, have attracted increasing interest from academic and industrial sectors alike [20, 22]. These natural products can be unearthed with the aid of ever-improving deep learning algorithms, like DeepRiPP [23]. Therefore, we propose that upon gathering novel protein sequences from data-sources aiming at the creation of a new version of this repository, one should disregard any lower or upper limits for protein sequence length.

Another hypothesis concerning the sequences lacking CD annotation is that these might correspond to false positives from the open reading frame (ORF)-calling step, presumably undertaken upstream from our work. If none of these possibilities are verified, we speculate that these sequences with no CD might be at a prime for discovering new molecular functions altogether. RPS-BLAST results (or absence thereof) should also be cross-validated with more sensitive procedures in the future, like querying hidden Markov model (HMM) profiles of protein domains.

DeepEC assigned at least one 4th-digit EC to 2.78% (351,917) of proteins, thus suggesting that these are potentially enzymes. Conversely, this result indicates that 97.22% (12,302,926) of these uncharacterized proteins lack catalytic activity—or at least one recognizable by DeepEC. One should emphasize that 304,293 (86%) of these putative enzymes also have GO and CD annotation. Once again, this suggests that DeepEC might be the most conservative prediction tool out of the three we used.

DeepEC outputs a score for each prediction. DeepEC results without prediction scores issue from an homology analysis sub-routine. We required the prediction scores for downstream work. This meant excluding 15,333 putative enzymes that were product of homology analysis. By doing so, a new total of 336,584 putative enzymes was gathered. From these 336,584 putative enzymes, 326,962 had a single EC annotation. We found that the most abundant enzymatic classes were Transferases (182,797) and Hydrolases (100,475). This was a promising result, as the majority of enzymes of biotechnological interest are of hydrolytic nature [24].

We also found 9,622 putative enzymes that had more than one predicted 4th-digit EC. Most of these enzymes have EC numbers that share the first 3 digits (8,387); or the first 2 (365). We hypothesize that these enzymes are catalytically promiscuous. Potential catalytic promiscuity, as suggested by multiple EC assignments, might be assessed in the future by quantitatively comparing the multiple potential reactions presumably catalyzed by a given enzyme—e.g., as achieved by the late EC-BLAST [25].

The remaining putative enzymes had multiple ECs that either share the 1st-digit (121), or span multiple enzymatic groups altogether (749). These might be enzymes that perform multiple functions. In theory, these might have attained multiple active sites via gene fusion events. Prospective endeavors might determine multifunctionality by analyzing the domains of these enzymes for the presence of active sites commonly associated with those catalytic activities. Nonetheless, additional work is necessary to determine the promiscuity, or multifunctionality, of these enzymes.

Addressing this particular subset of enzymes is of great significance for future biotechnological developments, as it might offer innovative solutions for the enzymes industry. For instance, by using an enzyme capable of catalysing several key reactions in a given industrial process, the need for multiple enzymes, and thus multiple physico-chemical requirements (e.g., temperature, pH, solvent) might, in theory, be disregarded altogether.

The metadata enrichment process described in this section represents the second set of contributions from this thesis. The molecular function annotations we have gathered from our ensemble framework provide a valuable accretion to the repository created in the first objective, and offer a glimpse into the potential functions of these elusive proteins. In a future version of this ensemble framework, one should also strive to include more classifiers and predicted features, such as genomic context, protein-protein interactions, subcellular location, and tertiary structure.

5.1.3 Validation

The third and last objective of this thesis was manifold. First, we sought to validate the annotations gathered for our $\sim$12 million representative protein dataset. Afterwards, we wanted to determine whether this dataset enclosed any putative enzymes of known biotechnological or industrial interest. Lastly, we aimed to select a putative enzyme subclass of pressing societal significance, and provided a proof-of-concept for the possible real-world usefulness of the work developed hitherto.

The first step towards this validation, was to develop a method that would allow us to numerically qualify each protein according to its annotation. This was required so that the annotations became quantitatively comparable to one another. Minding that some annotation terms are more specific than others, this meant that our method needed to tap into the amount of information

conveyed by a given term—i.e., its Information Content (*IC*).

IC-based metrics have been extensively used for biological ontologies ever since the work of Lord et al. [26]. Yet, to our knowledge, these types of metrics had never been applied to EC numbers, nor CD PSSM identifiers. Moreover, the *IC* of a term does not represent the full extent of information in an annotated protein. There are two main reasons for this. First, a protein might have multiple annotations. Second, these annotations might issue from multiple Classification Systems.

To accommodate these circumstances, we developed a new family of *IC*-based measures, tailored for the computation of the informational value within an annotated protein. This computation takes into account a set of multiple terms from a given Classification System. These measures were the Protein Information Content (*PIC*), and its prediction-score-weighted version (PIC_S). Given that the Classification Systems used in this work belong to one of two main categories—with and without subsumption relations—we developed two equations for each measure. This way, each measure could be calculated for all Classification Systems we used.

By calculating the PIC/PIC_S for all proteins annotated with each Classification System, we have shown that it is possible to represent our entire dataset as a collection of distributions. These distributions depict spectra of protein annotation, and their values provide insight into the extent of these annotations; and even whether these annotations relate to multiple molecular functions.

We ultimately sought to combine the PIC_S measures calculated for each Classification System into an index. This index would illustrate the extent to which a protein was annotated in relation to the three Classification Systems we used. We named it Protein Annotation Index (*PAI*). To our knowledge, this metric is the first attempt at creating a unifying measure for protein annotation content and quality.

The *PAI* allows to systematize protein annotations according to the full extent of their informational value. Consequently, it enables the creation of a global distribution that characterizes different degrees of protein annotation within a given dataset. Moreover, the *PAI* is both scalable and modular. It can be extended to include other Classification Systems, provided the predicted terms have an associated confidence score. The weights it comprises (i.e., w_C) might also be replaced by other constants, according to specific needs. We propose that the *PAI* measure should be improved in order to include modular weight constants according to different

objectives. For instance, including the p-value of the hypergeometric test in order to ascertain how overrepresented a term is in comparison to a gold-standard dataset. This might also be used to assess the rarity of a given molecular function.

We found that by calculating the value of *PAI* for all proteins in a dataset, we can unravel distinct sub-distributions under the guise of density peaks. These conformed to proteins with differing levels of annotation. This development allowed us to distinguish which proteins had been more thoroughly annotated. After applying a thresholding algorithm, we were further able to select a value of *PAI* that allowed for the subsection of our dataset among two major protein groups—i.e., those that were effectively annotated, and those that were not. By creating this dichotomy we expedite the targeting of protein sequences that were the most pliant to *in silico* annotation. In addition, we facilitate the selection of the most promising protein sequences for further analyses. Additionally, by building upon the family of measures that was created, we envision that a novel metric aimed at quantifying the consensus of semantic term annotations throughout different ontologies should take place.

Likewise, we also propose the creation of a unified measure of societal and economic potential for the enzymes whose biotechnological interest has been established *a priori*. This could be achieved by developing cost evaluation models to dynamically estimate market value. These models should be updatable through time, allowing for the recalculation of an enzyme's value, according to internal factors such as their annotation quality (i.e., *PAI*); and external ones like market fluctuations. This measure should also allow to predict potential domains of application—e.g., those established by the United Nations 2030 Agenda for Sustainable Development [27]. In doing so, it should allow to assign a tangible value for the most promising enzyme candidates, and establish a priority hierarchy for future biochemical characterization endeavors.

From a subset containing the most extensively annotated proteins, we selected the enzyme subclasses that were of known biotechnological and/or industrial interest (see Table 1.1). This amounted to 64 enzyme subclasses, totaling on 13,734 protein cluster representatives. From these 64 enzyme subclasses, 34 belonged to the Hydrolase class (i.e., EC:3.-.-.-), amounting to 11,757 protein cluster representatives. Overall, these 13,734 protein sequences represent prime candidates of industrial interest for biochemical characterization.

Among 34 Hydrolase subclasses, we chose that of Cutinases (EC:3.1.1.74) as a case study. We chose to investigate the Cutinase subclass based on the catalytic versatility of its members [28], industrial pluripotentiality [29], and reported ability to hydrolize high molecular weight

polyesters, like polyethylene terephthalate (PET) [30]. This case study provided a proof-of-concept for the entire workflow developed throughout this thesis.

We have identified 48 novel putative cutinases that possessed the characteristic cutinase CD. These held the most promise as their annotation depicted a consensus between the EC and CD Classification Systems. Afterwards, we compared the foregoing 48 putative cutinases with 30 bona fide PET hydrolases. Surprisingly, we found that the PET hydrolases that shared the greatest amino-acid identity to these 48 putative cutinases were of fungal origin. These were the *Fusarium solani pisi* cutinase (FsC), and the *Humicola insolens* cutinase (HiC). Additional work is needed in order to understand how do these putative cutinases relate to a broader spectrum of serine esterases across the three domains of life.

Upon performing a Multiple Sequence Alignment (MSA) among HiC, FsC, and the 48 putative cutinases, we were able to confirm that the active site is present throughout all sequences. Moreover, a characteristic pentapeptide motif, and four cysteine residues involved in disulfide bond formation, were also highly conserved. The conservation of these motifs suggested three things. First, that these 48 putative cutinases are likely functional. Second, that they might be thermostable to some extent. Third, that the tertiary structure of these 48 putative cutinases might resemble that of FsC and/or HiC.

After calculating several physico-chemical properties for these 48 cutinases, we tentatively suggest that these might withstanding harsher industrial conditions than those allowed for FsC or HiC. However, subsequent analyses are needed in order to confirm these assumptions. This process might start by predicting the tertiary structure for these enzymes, and by comparing it against thermostable PET hydrolases (e.g., HiC) whose structure has already been resolved. If the results from the previous procedure are promising, heterologous expression within a compatible host might ensue, followed by protein purification along with thermostability and pH stability assays.

We conclude that these putative cutinases might be promising candidates for biotechnological applications, and further disclose that these cluster representatives relate to a total of 254 cluster members. These 254 protein sequences might be at a prime for further analysis in the search for novel PET hydrolases.

References

[1] Schwartz R, Ting CS, King J. Whole proteome pI values correlate with subcellular localizations of proteins for organisms within the three domains of life. Genome Res. 2001;11(5):703–709.

[2] Hauser M, Mayer CE, Söding J. kClust: fast and sensitive clustering of large protein sequence databases. BMC Bioinformatics. 2013;14:248.

[3] Edgar RC. Search and clustering orders of magnitude faster than BLAST. Bioinformatics. 2010;26(19):2460–2461.

[4] Fu L, Niu B, Zhu Z, Wu S, Li W. CD-HIT: accelerated for clustering the next-generation sequencing data. Bioinformatics. 2012;28(23):3150–3152.

[5] Hauser M, Steinegger M, Söding J. MMseqs software suite for fast and deep clustering and searching of large protein sequence sets. Bioinformatics. 2016;32(9):1323–1330.

[6] Steinegger M, Söding J. Clustering huge protein sequence sets in linear time. Nat Commun. 2018;9(1):2542.

[7] Chang JM, Di Tommaso P, Notredame C. TCS: a new multiple sequence alignment reliability measure to estimate alignment accuracy and improve phylogenetic tree reconstruction. Mol Biol Evol. 2014;31(6):1625–1637.

[8] Suzek BE, Huang H, McGarvey P, Mazumder R, Wu CH. UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics. 2007;23(10):1282–1288.

[9] Li W, Wooley JC, Godzik A. Probing metagenomics by rapid cluster analysis of very large datasets. PLoS One. 2008;3(10):e3375.

[10] Li W. Analysis and comparison of very large metagenomes with fast clustering and functional annotation. BMC Bioinformatics. 2009;10:359.

[11] Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 1997;25(17):3389–3402.

[12] Kulmanov M, Khan MA, Hoehndorf R, Wren J. DeepGO: predicting protein functions from sequence and interactions using a deep ontology-aware classifier. Bioinformatics. 2018;34(4):660–668.

[13] Ryu JY, Kim HU, Lee SY. Deep learning enables high-quality and high-throughput prediction of enzyme commission numbers. Proc Natl Acad Sci U S A. 2019;116(28):13996–14001.

[14] Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, et al. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. 2003;31(1):365–370.

[15] Kulmanov M, Hoehndorf R. DeepGOPlus: improved protein function prediction from sequence. Bioinformatics. 2020;36(2):422–429.

[16] You R, Yao S, Mamitsuka H, Zhu S. DeepGraphGO: graph neural network for large-scale, multispecies protein function prediction. Bioinformatics. 2021;37(Suppl_1):i262–i271.

[17] You R, Huang X, Zhu S. DeepText2GO: Improving large-scale protein function prediction with deep semantic text representation. Methods. 2018;145:82–90.

[18] You R, Zhang Z, Xiong Y, Sun F, Mamitsuka H, Zhu S. GOLabeler: improving sequence-based large-scale protein function prediction by learning to rank. Bioinformatics. 2018;34(14):2465–2473.

[19] Aggarwal C, Jimenez JC, Nanavati D, Federle MJ. Multiple length peptide-pheromone variants produced by Streptococcus pyogenes directly bind Rgg proteins to confer transcriptional regulation. J Biol Chem. 2014;289(32):22427–22436.

[20] Arnison PG, Bibb MJ, Bierbaum G, Bowers AA, Bugni TS, Bulaj G, et al. Ribosomally synthesized and post-translationally modified peptide natural products: overview and recommendations for a universal nomenclature. Nat Prod Rep. 2013;30(1):108–160.

[21] Sberro H, Fremin BJ, Zlitni S, Edfors F, Greenfield N, Snyder MP, et al. Large-Scale Analyses of Human Microbiomes Reveal Thousands of Small, Novel Genes. Cell. 2019;178(5):1245–1259.e14.

[22] Hetrick KJ, van der Donk WA. Ribosomally synthesized and post-translationally modified peptide natural product discovery in the genomic era. Curr Opin Chem Biol. 2017;38:36–44.

[23] Merwin NJ, Mousa WK, Dejong CA, Skinnider MA, Cannon MJ, Li H, et al. DeepRiPP integrates multiomics data to automate discovery of novel ribosomally synthesized natural products. Proc Natl Acad Sci U S A. 2020;117(1):371–380.

[24] Verma S, Meghwanshi GK, Kumar R. Current perspectives for microbial lipases from extremophiles and metagenomics. Biochimie. 2021;182:23–36.

[25] Rahman SA, Cuesta SM, Furnham N, Holliday GL, Thornton JM. EC-BLAST: a tool to automatically search and compare enzyme reactions. Nat Methods. 2014;11(2):171–174.

[26] Lord PW, Stevens RD, Brass A, Goble CA. Investigating semantic similarity measures across the Gene Ontology: the relationship between sequence and annotation. Bioinformatics. 2003;19(10):1275–1283.

[27] Transforming our world: the 2030 Agenda for Sustainable Development. Sustainable Development Knowledge Platform;. https://sustainabledevelopment.un.org/post2015/transformingourworld.

[28] Danso D, Chow J, Streit WR. Plastics: Environmental and Biotechnological Perspectives on Microbial Degradation. Appl Environ Microbiol. 2019;85(19).

[29] Pio TF, Macedo GA. Cutinases: properties and industrial applications. Adv Appl Microbiol. 2009;66:77–95.

[30] Carr CM, Clarke DJ, Dobson ADW. Microbial Polyethylene Terephthalate Hydrolases: Current and Future Perspectives. Front Microbiol. 2020;11:571265.

Chapter 6

Concluding Remarks

Protein function imputation has stood as one of the most challenging tasks in computational biology to date. This challenge has been tackled by numerous research groups worldwide, and substantial progress has been made in the past years. Apart from identifying the molecular function of a protein sequence, another challenge is that of unearthing its potential use for mankind.

Notwithstanding, gene products from the FDM keep accumulating throughout public sequence databases and repositories. One can envision this predicament leading to several nefarious consequences in due time (if not already). For instance, the ever-expanding volume of FDM might give rise to a deluge of intangible and disorganized information in a near future. This scenario might make data prospection an exceedingly laborious task, as the volume of unannotated data to be mined rises exponentially. One can also foresee database queries taking increasingly longer to comply with a given request, due to a cumulative magnification of sequence search space.

Additionally, if left unaddressed, FDM renders an incomplete portrait of knowledge regarding prokaryotic genomes, and therefore biological systems as a whole. This circumstance in turn perpetuates the existence of unanswered questions, and thus unmet explanations. By neglecting the potential for groundbreaking advances in biotechnological solutions for current (and future) society, we risk being kept in the dark, while the solutions to a fair amount of our problems might very well lay hidden right in front of us. We also take great risk in trying to reinvent "biological wheels". As an example, one might try to synthetically engineer new enzymes with enhanced catalytic performance, or chemically synthesize new compounds of pharmacological interest; when even better extremozymes lay unnoticed, and BGCs coding for the very compounds one is striving to create lay unseen.

This thesis took the first small steps towards the identification of biotechnological potential within the prokaryotic FDM. We present evidence corroborating the outstanding source of novelty among these uncharacterized gene products. In doing so, we have laid the foundations for a prospective systematic interrogation of the FDM, possibly allowing to extricate new biotechnological solutions from these elusive elements in the near future. The work developed throughout this dissertation might be automated hereafter, paving the way for the functional inference of FDM from "Big-Sequence-Data". It is our opinion that true, accurate, *in silico* protein function imputation is presently attainable. The attainment of this goal might very well be a matter of exhaustiveness, stringency, and connecting the information provided by a multitude of differing

sources, and distinct annotations issuing from orthogonal molecular function imputation tools. It is of utmost importance to bring this uncharted knowledge to light, as these gene products can provide remarkable contributions to overcome contemporary societal issues.

Garcia de Orta, a XVI-century Portuguese naturalist and physician, triggered a paradigm shift in medicine and science during the Portuguese Age of Discovery. Among his many writings he states that "we now learn more in one day by the Portuguese than what has been known for a hundred years with the Romans". This quote could not resonate more with the current circumstances provided by the wealth and breadth of information in this post-genomic era. Undoubtedly the horizon goes far and wide, and the potential treasures that can be unearthed by prospecting this information are hidden in plain sight. It is incumbent upon the current, and the following, generation of microbiologists and bioinformaticians alike to relentlessly seek light amidst the microbial dark matter. A call to arms beckons from the abyss, and as Garcia de Orta once declared: "What we do not know today, we shall know tomorrow".