TÉCNICO LISBOA

UNIVERSIDADE DE LISBOA
INSTITUTO SUPERIOR TÉCNICO

Illuminating the Dark Proteome

Nelson Ricardo Perdigão Pereira

Supervisor: Doctor Agostinho Cláudio da Rosa
Co-Supervisors: Doctor Seán Ignatius O'Donoghue
Doctor Andrea Schafferhans-Fuhrmann

Thesis approved in public session to obtain the PhD Degree in
Information Systems and Computer Engineering
Jury final classification: Pass with Distinction

Jury
Chairperson: President of the Instituto Superior Técnico
Members of the Committee:
Doctor Jorge Manuel Santos Pacheco
Doctor Arlindo Manuel Limede de Oliveira
Doctor Cláudio Emanuel Moreira Gomes
Doctor Agostinho Cláudio da Rosa
Doctor Francisco José Moreira Couto
Doctor Seán Ignatius O'Donoghue
Doctor James Benedict Procter

2017

Cover Photo Credits: Photo created by Christopher Hammang, Sean O'Donoghue, and Julian Heinrich for Dark Proteome paper at Proceedings of National Academy of Sciences of the United States of America. doi: 10.1073/pnas.1508380112
Awarded at ISMB2016, Orlando, Florida, USA with the ISMB2016 Art & Science Award.

TÉCNICO LISBOA

UNIVERSIDADE DE LISBOA
INSTITUTO SUPERIOR TÉCNICO

Illuminating the Dark Proteome

Nelson Ricardo Perdigão Pereira

Supervisor: Doctor Agostinho Cláudio da Rosa
Co-Supervisors: Doctor Seán Ignatius O'Donoghue
Doctor Andrea Schafferhans-Fuhrmann

Thesis approved in public session to obtain the PhD Degree in
Information Systems and Computer Engineering
Jury final classification: Pass with Distinction

Jury
Chairperson: President of the Instituto Superior Técnico
Members of the Committee:
Doctor Jorge Manuel Santos Pacheco, Full Professor, Escola de Ciências, Universidade do Minho;
Doctor Arlindo Manuel Limede de Oliveira, Full Professor, Instituto Superior Técnico, Universidade de Lisboa;
Doctor Cláudio Emanuel Moreira Gomes, Coordinator Researcher, Faculdade de Ciências, Universidade de Lisboa;
Doctor Agostinho Cláudio da Rosa, Associate Professor with Habilitation, Instituto Superior Técnico, Universidade de Lisboa;
Doctor Francisco José Moreira Couto, Associate Professor, Faculdade de Ciências, Universidade de Lisboa;
Doctor Seán Ignatius O'Donoghue, OCE Science Leader, Commonwealth Scientific and Industrial Research Organization, Australia;
Doctor James Benedict Procter, Senior Post-Doctoral Researcher, School of Life Sciences, University of Dundee, UK.

Funding Institutions
Fundação para a Ciência e Tecnologia (FCT)
European Molecular Biology Organization (EMBO)

2017

Resumo

Modelos moleculares estruturais das proteínas ajudam a dar uma visão detalhada sobre as suas funções, especialmente quando combinadas com as características das sequências. Atualmente modelos 3D estão disponíveis para diversas proteínas, contudo na prática é complexo encontrar os modelos apropriados e visualizá-los conjuntamente com as características da sequência.

Tendo isto em mente foi desenvolvido o Aquária, um novo recurso que fornece 46 milhões modelos estruturais pré-calculados através da utilização de homologia entre sequências e estruturas – 10 vezes mais do que atualmente disponibilizado por outros recursos. Fornece também, pelo menos um modelo para 87% de todas as proteínas da Swiss-Prot com uma média de 35 modelos por proteína. Através do Aquária foi analisado o proteoma conhecido ou "visível". Contudo, o seu complementar, o proteoma *dark* ou desconhecido, - i.e., regiões das proteínas que permanecem teimosamente inacessíveis quer por determinação estrutural experimental quer por modelação – também foi pesquisado, armazenado e indexado na *Dark Proteome Database*.

Usando os dois sistemas indicados acima, foi feito o estudo mais exaustivo sobre modelação estrutural cobrindo 546,000 proteínas pertencentes a vários organismos, onde se concluiu que 44–54% do proteoma das eucarióticas e dos vírus são *dark*, comparado com 14% do proteoma das arqueas e das bactérias. Surpreendentemente mais de metade do proteoma *dark* não pode ser explicado pelos argumentos habituais tais como desordem intrínseca, regiões trans-membranares ou viés composicional. Aproximadamente metade do proteoma *dark* é composto por proteínas dark, em que a sequência total não tem semelhança com as estruturas conhecidas. As proteínas *dark* possuem uma série de funções, mas um subconjunto largo e distinto destas mostraram características inesperadas tais como associação com secreções, presença em tecidos específicos, como o retículo endoplasmático e clivagem proteolítica. As proteínas *dark* são também mais curtas ao nível da sequência, têm pouca reutilização evolucionária e poucas interações conhecidas com outras proteínas. Esta tese sugere ainda, a existência de regiões trans-menbranares que são indetectáveis pelos métodos correntes de predição.

Desta forma, esta tese sugere novas direções de investigação em biologia estrutural e computacional. Este trabalho vai ajudar, certamente, a concentrar esforços futuros sobre a investigação do restante proteoma *dark*, potencialmente revelando processos moleculares da vida que são atualmente desconhecidos.

Palavras-chave: Big Data; Bases de Dados; Homologia; Proteínas; Estrutura.

Abstract

Molecular models of a protein's structure can give detailed insight into mechanisms underlying its function, especially when viewed in combination with sequence features. In theory, 3D structural models are now available for many proteins, however in practice it is often complex to find all appropriate models and view them with sequence features.

Thus, we developed Aquaria, a new web resource that provides 46 million pre-calculated structural models using homology from sequence to structure – 10 times more than currently available from other resources, resulting in at least one matching structure for 87% of Swiss-Prot proteins and a median of 35 structures per protein Using Aquaria, we surveyed the known or visible proteome. Its complement, the 'unknown' or 'dark' proteome, i.e., regions of proteins that remain stubbornly inaccessible to both experimental structure determination and modeling, was scanned, stored and indexed into the Dark Proteome Database.

Using the above systems, it was performed the most recent structural modeling study covering 546,000 proteins across many organisms, where it was found 44–54% of the proteome in eukaryotes and viruses is dark, compared with only 14% for archaea and bacteria. Surprisingly, most of the dark proteome could not be accounted for by conventional explanations, such as intrinsic disorder, transmembrane regions or compositional bias. Nearly half of the dark proteome comprised dark proteins, in which the entire sequence lacked similarity to any known structure. Dark proteins fulfill a wide variety of functions, but a subset showed distinct and largely unexpected features, such as association with secretion, specific tissues, the endoplasmic reticulum, disulfide bonding, and proteolytic cleavage. Dark proteins also had short sequence length, low evolutionary reuse, and few known interactions with other proteins. This thesis also suggests the existence of transmembrane regions undetected by current prediction methods.

Therefore, our work suggests several new directions for research in structural and computational biology. This work surely will help focus the efforts of future research to shed light on the remaining dark proteome thus potentially revealing molecular processes of life that are currently unknown.

Key-words: Big Data; Databases; Homology; Proteins; Structure.

Para a minha querida mãe.

Acknowledgements

I would like to thank to my supervisors
Prof. Agostinho Rosa, Doctor Andrea Schafferhans and Doctor Seán O'Donoghue,

I also want to thank the co-authors from CSIRO, Garvan and University of Sydney
Kenny Sabir, Christian Stolte, Julian Heinrich, Vivian Ho, Manfred Ross,
Fabian Buske, Michael Buckley, Bruce Tabor,
Beth Signal, Brian Gloss
and Christopher Hammang.

Also a thanks to the co-authors of RostLab at Munich,
Maria Kalemanov and Benjamin Wellmann.

A thank you also to
Doctor Theodoros Soldatos,
Doctor Buckhard Rost,
Doctor Reinhard Schneider and
Venkata Satagopam.

Also a thank you to
Professor Lawrence Hunter and
Professor Des Higgins
for the enlighten and coffees offered
in Heidelberg and Dublin.

Last but not least, a final thanks to all the technical staff of post-graduation secretary of IST, specially to Ana Rosa, Carla Amaral, Joaquim Naia, Paula Cunha, Paula Simões, Drª. Josefina Miranda, Dr. Nuno Riscado and Drª Julia Oliveira for all the kindness and professionalism.

To EMBO/EMBL through
EMBO ASTF 263-2010/Award

To Fundação para a Ciência e Tecnologia through
SFRH/BD/29967/2006 PhD Grant

Contents

FIGURE INDEX

TABLE INDEX

PUBLICATIONS

Journals

[1] The Dark Proteome Database
N. Perdigão, A. C. Rosa, S. I. O'Donoghue, **BioData Mining, Springer Nature. (Minor Revision)**

[2] Unexpected Features of the 'Dark' Proteome
N. Perdigão, J. Heinrich , C. Stolte , K. S. Sabir , M. Buckley , B. Tabor , B. Signal , B. S. Gloss , C. J. Hammang , B. Rost, A. Schafferhans, S. I. O'Donoghue, **Proceedings of the National Academy of Sciences**, vol. 112 no. 52, pp.15898–15903, 2015.
doi: 10.1073/pnas.1508380112
http://www.pnas.org/content/112/52/15898

[3] Aquaria: simplifying discovery and insight from protein structures, S. O'Donoghue, K. Sabir, M. Kalemanov, C. Stolte, B. Wellmann, V. Ho, M. Roos, **N. Perdigão**, F. Buske, J. Heinrich, B. Rost, A. Schafferhans, **Nature Methods**, Vol. 12, pp. 98–99,
doi:10.1038/nmeth.3258, 2015
http://www.nature.com/nmeth/journal/v12/n2/full/nmeth.3258.html

Conferences

[1] Visual analytics of gene set comparison
N. Perdigão, T.G. Soldatos, K.S. Sabir, S.I. O'Donoghue, **IEEE Symposium on Big Data Visual Analytics**, pp. 1-2, Hobart, Australia, 2015
doi: 10.1109/BDVA.2015.7314304
http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=7314304

[2] Unexpected Features of the 'Dark' Proteome of structural biology, **N. Perdigão,** J. Heinrich , C. Stolte , K. S. Sabir , M. Buckley , B. Tabor , B. Signal , B. S. Gloss , C. J. Hammang , B. Rost, A. Schafferhans, S. I. O'Donoghue, Proc. of Structural Bioinformatics and Computational Biophysics, pp. 54-55, 3Dsig 2016 – Orlando, Florida, United States of America, 2016.

[3] Aquaria: simplifying discovery and insight from protein structures, S. O'Donoghue, K. Sabir, M. Kalemanov, C. Stolte, B. Wellmann, V. Ho, M. Roos, **N. Perdigão**, F. Buske, J. Heinrich, B. Rost, A. Schafferhans, Proc. of Structural Bioinformatics and Computational Biophysics, pp. 76, 3Dsig 2015 – Dublin, Ireland, 2015.

[4] From SRS 3D to Aquaria, making protein structures discoverable, S.O'Donoghue, K. Sabir, C. Stolte, **N. Perdigão**, V. Ho, V. P. Satagopam, M. Kalemanov, M. Roos, D. Ma, B. Rost, A. Schafferhans, Proc. of Structural Bioinformatics and Computational Biophysics, pp. 97, 3Dsig 2013 – Berlin, Germany, 2013.

Posters with oral presentation

[1] The Dark Proteome, N. Perdigão, J. Heinrich, C. Stolte, K. Sabir, M. J. Buckley, B. Taylor, B. Signal**,** B. S. Gloss, C. J. Hammang, B. Rost, A. Schafferhans, S.O'Donoghue, Poster C04, Vizbi, EMBL, Heidelberg, Germany, 2016.

[2] Aquaria – Simplifying Insight from Protein Structures, S.O'Donoghue, K. Sabir, C. Stolte, M. Kalemanov, B. Wellmann, V. Ho, F. Buske, M. Ross, **N. Perdigão,** J. Heinrich, B. Rost, A. Schafferhans, Poster A14, Vizbi, EMBL, Heidelberg, Germany, 2015.

[3] Aquaria – the 3D Viewer, K. Sabir, S.O'Donoghue, M. Kalemanov, C. Stolte, B. Wellmann, V. Ho, F. Buske, M. Ross, **N. Perdigão,** J. Heinrich, B. Rost, A. Schafferhans, Poster A13, Vizbi, Broad Institute of MIT and Harvard, Cambridge MA, USA, 2014.

[4] Aquaria – the 2D parts, S.O'Donoghue, K. Sabir, M. Kalemanov, C. Stolte, B. Wellmann, V. Ho, F. Buske, M. Ross, **N. Perdigão,** J. Heinrich, B. Rost, A. Schafferhans, Poster A13, Vizbi, Broad Institute of MIT and Harvard, Cambridge MA, USA, 2014.

[5] From SRS 3D to Aquaria, making protein structures discoverable, S.O'Donoghue, K. Sabir, C. Stolte, **N. Perdigão**, V. Ho, V. P. Satagopam, M. Kalemanov, M. Roos, D. Ma, B. Rost, A. Schafferhans, Poster D05, Vizbi, EMBL, Heidelberg, Germany, 2013.

AWARDS

2015 National iAwards Merit
Aquaria: simplifying discovery and insight from protein structures, S. O'Donoghue, K. Sabir, M. Kalemanov, C. Stolte, B. Wellmann, V. Ho, M. Roos, **N. Perdigão**, F. Buske, J. Heinrich, B. Rost, A. Schafferhans.

2015 New South Wales iAwards Winner
Aquaria: simplifying discovery and insight from protein structures, S. O'Donoghue, K. Sabir, M. Kalemanov, C. Stolte, B. Wellmann, V. Ho, M. Roos, **N. Perdigão**, F. Buske, J. Heinrich, B. Rost, A. Schafferhans.

ABREVIATIONS

BLAST	Basic Local Alignment Search Tool
CASP	Critical Assessment of Structure Prediction
COPS	Classification Of Protein Structures
DNA	Deoxyribonucleic Acid
DPD	Dark Proteome Database
DSSP	Dictionary of Protein Secondary Structure: Pattern recognition
EBI	European Bioinformatics Institute
EMBL	European Molecular Biology Laboratory
EMBO	European Molecular Biology Organization
HSSP	Homology-derived Secondary Structure of Proteins
HHblits	Homology detection by iterative HMM-HMM comparison
IDP	Intrinsically Disordered Proteins
mRNA	messenger Ribonucleic Acid
NIH	National Institutes of Health
NMR	Nuclear Magnetic Resonance
PDB	Protein Databank
PIR	Protein Identification Resource
PSD	Protein Sequence Database
PSI-BLAST	Position-Specific Iterative Basic Local Alignment Search Tool
PSSH	Protein Sequence-to-Structure Homologies
RNA	Ribonucleic Acid
rRNA	ribosomal RNA
SIB	Swiss Institute of Bioinformatics
SRS	Sequence Retrieval System
SRS3D	Sequence Retrieval System 3D
SNP	Single Nucleotide Polymorphisms
TrEMBL	Translated EMBL
tRNA	transport RNA
UniProt	United Protein Databases

PART I

MOTIVATION

1. Introduction

The determination of biomolecular 3D structures at atomic-resolution has provided fundamental insights that have revolutionized our understanding of the molecular machinery of life. Probably the most spectacular example so far has been the elucidation of DNA's structure and the insight this provided into the mechanism through which genetic information is stored and inherited. There are thousands of other examples, including broad categories such as the use of structural insight in rational drug design, as well as in antibody engineering.

1.1. The Visible Protein 3D Structures (The Bright Side)

Methods for the experimental determination of structure have improved continuously, as evidenced by the steady growth of PDB (Berman et al., 2000), which collects essentially all known biomolecular structures and that has reached more than 120,250 total of entries in 2016. As often noted, there is a rapidly increasing gap between this and the rate at which DNA and protein sequence information is being acquired – less than 0.1% of UniProt (Consortium, 2014) proteins contain a matching PDB structure for part of their sequence (statistics from UniProt website). However, evolution tends to conserve structure more than sequence, e.g., the human proteome has well over 100,000 distinct protein sequences believed to adopt only around a few thousand distinct folds. This understanding has led to several large-scale computational modelling initiatives e.g., ModBase (Pieper et al., 2014), SWISS-MODEL (Kiefer et al, 2009) and CSPM (Stroud et al., 2009). The models from many of these initiatives are consolidated in the Protein Model Portal (PMP) (Haas et al., 2013). Together, these currently provide some structural information for about 9.2% of all UniProt proteins; in most cases however, this does not cover the full length sequence, so in total structural information can be inferred for 6.7% of all UniProt (statistics from PMP website). Since protein sequence can be predicted quite accurately from genomic sequences, these advances mean that structural biology now scales with the very rapid advance of genomic sequencing (Mardis, 2011) by first predicting protein sequence from genomes.

While decades ago atomic resolution structures were relatively rare and were not available for most of the proteins, RNAs, or protein-DNA complexes studied by biologists. In the present, there is a considerable amount of structural information for the majority of known protein sequences, which provides a wealth of detailed insight into biological functions - far beyond what is accessible from sequence alone.

However, members of the structural biology community have expressed the concern that structures are under-utilized by life scientists (O'Donoghue et al., 2015). Why would such potentially useful and insightful data be underutilized? There are several contributing factors:

(1) Data volume. Now that millions of sequences and thousands of structural models are available, we are facing the classic problem of Big Data – where knowledge can easily be lost in the sea of data. Overcoming this requires careful designed strategies to enable effective navigation and use of such large databases.

(2) Data complexity. The amount of information conveyed in a macromolecular structure is intrinsically much more complex than, for example, the information conveyed in the corresponding protein, RNA or DNA sequences. Thus, relatively complex software user interfaces are required for both in order to find this information and to use it for deriving insight into biological function. Using such interfaces and interpreting the data requires some specialization in structural biology. In addition, the PDB website itself addresses the needs of the structural biologists, as they are, after all, the people who created the database. However, for many of the remaining biologists who are not experts in macromolecular structures, PDB's organization and websites can be confusing. Thus, other websites that provide a different view of this data have emerged.

(3) Tailored views for non-specialists. Many of the web resources that disseminate structure data have been created by and for the structural biology community. Now that structural models are available for a significant fraction of all protein sequences, this data becomes more interesting to a broader group of life scientists, many of which have less experience in molecular graphic's methods or in concepts required to interpret macromolecular structures. Related to this point, we believe there is one important and useful structural data view that is currently missing, one that would provide a concise visual summary of all related structural information for any given protein.

In the attempt of finding answers to the above points, Aquaria is developed (Chapter 4). Besides visualizing all the information concerning a protein, Aquaria also reorganizes all PDB/Swiss-Prot. Using Protein Sequence-to-Structure Homologies (PSSH2) (O'Donoghue et al., 2015) by systematically comparing 546,000 Swiss-Prot sequences against 100,326 PDB proteins structures, i.e., an alignment between all well-described protein sequences across all range of known structures is established. This comparison resulted in 46 million sequence-to-structure alignments (O'Donoghue et al.,

2015), which represents a depth not available from other resources, increasing the structural knowledge for sequences that didn't possess this information by default.

1.2. The Dark Side

The Dark Side of the Proteome is the core of this thesis. Aquaria was used to map and integrate all the structural information from PDB using homology into the Swiss-Prot sequences (The 'Bright' Side). This was the purpose of Aquaria, but after observing so many dark regions through homology, it was decided to map and characterize these dark regions, i.e., regions of protein sequence, or whole sequences, stubbornly inaccessible to either experimental structure determination or modeling, and hence where 3D conformation is completely unknown (Chapters 5 and 6). The dark proteome has often been overlooked so far, but after this mapping and the unexpected features obtained (Perdigão et al., 2015) that raised so many questions, that dark proteome initiatives started to appear.

Scientists have long speculated about the nature of the dark proteome, the area of proteins that are completely unknown. Having mapped the boundaries of these dark regions, bring us one step closer to discovering the complete structure and function of all proteins, because knowing what we do not know has provided to the scientific community a new roadmap to focus future research.

From the reactions of the scientific community (scientific media) as well as, private communications of a couple of noted scientists, this survey of the unknown using computational methodology will set future research directions, as dark matter has done in physics (Bertone et al., 2005).

1.3. Contributions

This thesis has four main contributions using Big Data. Big data is a term for data sets that are so large or complex that traditional data processing applications are inadequate. Challenges include analysis, capture, data curation, search, sharing, storage, transfer, visualization, querying, updating and information privacy (Hilbert & López, 2011).

Aquaria

A new, powerful and publicly available web resource providing quick and extensive insight into the 3D structure of proteins is developed to help life scientists better understand diseases and develop new medicines.

When scientists discuss proteins, they are talking about the many thousands of molecules that act as the essential building blocks of life as we know it. Because proteins are so important to constructing life, researchers need a way to visualize the 3D structures of proteins and the exact ways in which they fit together, so as fully understand their functions – in our bodies and elsewhere in nature.

In the past, the search for protein structures was very tedious and required expert knowledge. In Aquaria, all data is already processed for 46 million models using homology, because Aquaria calculates the structure of most proteins based on Protein Data Bank (PDB) an online resource that houses more than 100.326 known protein structures for 546.000 Swiss-Prot protein sequences. The PDB is a fantastic resource containing a wealth of details about the molecular processes of life but we are aware that few biologists take full advantages of it. Therefore, we created Aquaria to make this valuable information more accessible and easier to use for discovery purposes.

Freely and publicly accessible, Aquaria will be useful to a broad range of life scientists, from medical researchers to those studying agriculture, biosecurity, ecology and nutrition. It can help them streamline their discovery process and gain new insight into protein structures. Aquaria is available at http://aquaria.ws.

Dark Proteome Database

The complete map of the Dark Proteome lives in this database and it was made for the first time. This database was the main resource for the work published at the Proceedings of the National Academy of Sciences of United States of America (Perdigão et al., 2015).

A valuable resource for scientists that wish to join us in the discovering of the Dark Proteome and its frontiers. The Dark Proteome Database is available at http://darkproteome.ws.

Dark Proteome

The full and complete map of the Dark Proteome and it’s characterization is made for the first time. As knowledge of three-dimensional protein structures continues to expand (also with the help of the first contribution above), we can identify regions within each protein that are different to any region where structure has been determined experimentally, coining the 'dark proteome' term.

These dark regions are unlike any known structure, so they cannot be predicted, so identifying these areas is very exciting, as we now have a map to focus research efforts. Our map defined the boundaries right at the edge of protein knowledge.

The research has yielded some surprising results, including that nearly half of the proteome in eukaryotes is dark and has unexpected features, including an association with secretory tissues, low evolutionary conservation, and very few known interactions with other proteins.

This work will help future research by shedding light on the remaining dark proteome, revealing molecular processes of life that are currently unknown. It will also provide insight into dark proteins based illnesses like cancer, type 2 diabetes, and many neurodegenerative diseases, such as Parkinson's disease and Alzheimer's. The dark proteome undoubtedly plays a key role in human health, as well as many other areas of life science. We believe that studying the dark proteome will clarify future research directions, as studies of dark matter have done in physics.

Dark Proteome Autonomy

Through the build of the Dark Autonomy Database**,** another unexpected result comes out, that was the fact that dark proteins have less protein-protein iterations, and therefore we could conclude that they are mostly autonomous.

Therefore, I can clearly state that this thesis has identified new and exciting scientific mysteries that will set directions for future research.

1.4. Organization

This dissertation has five Parts holding eight Chapters. The first Part (Motivation) holds Chapter 1 consisting of an introduction, a light overview and text outlook.

The second Part (Context) of the thesis contains two introductory Chapters that describe its application spectrum. The second Chapter consist of a brief explanation about Life, describes the existent cell types, as well as its organelles, and how genes generate proteins that are necessary for living organisms to operate and live. The third Chapter presents a small fraction of the Protein Universe where protein structure levels will be introduced, followed by the protein types and some actual protein databases

together with the type of information they hold. To conclude the chapter, it will be briefly described some protein homology concepts, and its corresponding databases.

The third Part (Methods) contains the chapters of the tools used to obtain the results of this thesis. Starting with Aquaria (Chapter 4) it will be explained why it is currently the best tool to explore the 'Bright Side' (i.e., the visible protein universe) as well as its advantages in regards to others systems. Chapter 5 represents the entry into the 'Dark Side', i.e., where the 'Dark Matter' of the Proteome or the 'Dark' Proteome will be mapped using the Dark Proteome Database tool.

The fourth Part (Results) exposes the results obtained with the previous described tools. This Part holds Chapter 6 and it's the core of the thesis and can be summarized as "the unstructured proteome that Aquaria didn't detected but we know that exists". This "unknown" structural proteome is the reason for our use of the Dark Matter metaphor in Physics. This chapter is the most important because mapped the complete Dark Proteome based on Swiss-Prot data of 2014, which is something that no one ever achieved before. Besides the mapping, unexpected results arose from this work caused considerable discussion in the scientific community. It also motivated the creation of Dark Proteome initiatives as consequence of the published results (Perdigão et al., 2015) that contradicted the mind-set of the structural biology community until then. The last results chapter presents the Dark Proteome Autonomy (Chapter 7) where it shows that dark proteins appear to be more autonomous than non-dark proteins.

Finally, the fifth Part (Conclusions) holds Chapter 8 (General Discussion), which assembles the different strands of this thesis into a discussion involving the Dark Proteome and its impact on biology and computer science.

PART II

CONTEXT

2. Life

What is life? This is probably the most common inner question that people asks themselves through their existence. Through history, civilizations and cultures the above question is made, but also about life origins, where it lays and how it works. Preceding the philosophers of ancient Greece, the concept of *vital forces* that were mysterious, divine, and which keep organisms functional and alive. Still, despite the numerous scientific breakthroughs and revelations of the last century in several scientific areas, the above concept advocating the metaphysical uniqueness of living matter enjoyed wide acceptance one century ago in the scientific community (Lagerkvist, 2005).

The above shift started around 200 years ago, through the dissemination of mechanical theories concerning physiology and nature (Nurse, 2003). Those theories suggested that all life-related phenomena could be explained by the same physical and chemical principles that were applied to the non-living world (Komdeur et al., 2009). Louis Pasteur (1822–1895), made an important breakthrough on this approach by demonstrating that the fermentation (i.e., a chemical process) of converting sugar to alcohol resulted in the growth of microorganisms. By doing this, Pasteur discovered the missing link between life processes and chemical reactions. That is, fermentation could be reproduced in the absence of the microorganism, by using substances extracted from it. Although at first the chemical nature of these substances was unclear, it was proved later on to be proteins (Chapter 3). These proteins therefore accelerated chemical reactions within cells without changing their main nature i.e., they acted as catalysts.

From this moment on, life was no longer seen as a mysterious and/or divine phenomena acting on organisms, but instead the result of several chemical processes performed by the proteins. This conscience became the basilar rock for modern biochemistry and molecular biology. Besides protein catalysts, (called later on enzymes), many other functional and relevant proteins have been found since then, where the most well-known example is hemoglobin, a protein that carry the oxygen from the lungs to body organs and tissues, and carry carbon oxide back to the lungs.

The genetic revolution that occurred in the second half of the previous century with the decoding of DNA (Deoxyribonucleic Acid) structure, and its genetic code, has proved that proteins are more than just molecular machinery that lives within the cells; they are also genes primary products, among other things.

2.1. Domains of Life

Our planet holds an enormous variety of organisms; This variety is manifested in the way they look, how they behave, their diet, reproduction mode, and how long they live, but there is one universal characteristic common to all of them; they are all made of cells (Nurse, 2003; Koshland, 2002). It is usual to separate the population of biological cells, as well as the organisms they form, into two principal types: cells without a nucleus and cells with a nucleus. Cells with no nucleus are *prokaryotic*. Prokaryotic cells are further classified into two groups: *bacteria* and *archaea.* Cells that have a nucleus are called *eukaryotic*. All the cells above have cell membranes, organelles, cytoplasm, and DNA.

2.1.1. Prokaryotic Cells

Prokaryotes are small ($\sim 10^{-6}$ m), and lacks any visible internal organization. Prokaryotic cells are made of a lipid membrane (the plasma membrane) engulfing an inner aqueous environment (the cytoplasm). The cytoplasm is where all life processes take place and it is separated from the external environment of the cell by the plasma membrane. *Escherichia coli,* the biochemically most well-characterized organism, is a typical prokaryote. Bacteria and archaea are prokaryotes which are single-celled organisms that do not have a nucleus or membrane-bound organelles.

2.1.1.1. Bacteria

Bacteria are the most common prokaryotes, and they are the smallest cells known. These tiny organisms are present in almost everywhere. They do not have a nucleus, but they do have DNA (nucleoid region). A bacteria's DNA is a long, circular molecule, shaped like a twisted rubber band. Bacteria have no membrane-covered organelles, but they do have ribosomes. Ribosomes are tiny, round organelles made of proteins and other material.

Bacteria also have a strong, web-like exterior cell wall. This wall helps the cell retain its shape. A bacterium's cell membrane is just inside the cell wall. Together, the cell wall and cell membrane allow materials into and out of the cell.

Some bacteria live in the soil and water. Others live in, or on, other organisms. For example, you have bacteria living on your skin and teeth. You also have bacteria living in your digestive system. These bacteria help the process of digestion. A typical bacterial cell is shown in Figure 2.1.

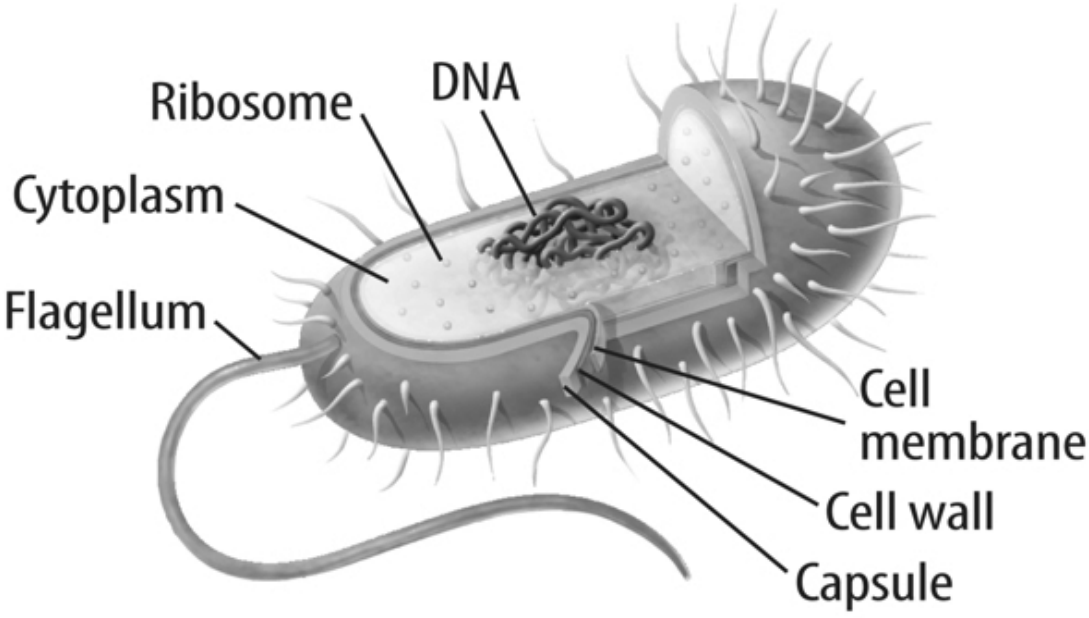


Figure 2.1: Bacterial structure. Bacterial cells lack membrane-bound organelles but have a variety of cell structures (Wikipedia site).

2.1.1.2. Archaea

The second kind of prokaryote are the archaea. Archaea are similar to bacteria in some ways. For example, both are single-celled organisms. Both have ribosomes, a cell membrane, and circular DNA. Both lack a nucleus and membrane-bound organelles, but archaea differ from bacteria in some way, too. For example, archaeal ribosomes are different from bacterial ribosomes.

Archaea are similar to eukaryotic cells in some ways, too. For example, archaeal ribosomes are more like the ribosomes of eukaryotic cells, but archaea also have some features that no other cells have. For example, the cell walls and cell membranes of archaea are different from the cell walls of other organisms. Some archaea live in places where no other organisms could live.

Three types of archaea are heat-loving, salt-loving, and methane-making. Heat-loving and salt-loving archaea are sometimes called extremophiles. Extremophiles live in places where conditions are extreme. They live in very hot water, such as in hot springs, or where the water is extremely salty.

2.1.2. Eukaryotic Cells

Eukaryotic cells are the largest cells. Most eukaryotic cells are still microscopic, but they are about 10 times larger than most bacterial cells.

All living things that are not bacteria or archaea are made of one or more eukaryotic cells. Organisms made of eukaryotic cells are called eukaryotes. Eukaryotes are multicellular. Multicellular means “many cells”. Multicellular organisms are usually larger than single-cell organisms. Most organisms you see with your naked eye are eukaryotes. Eukaryotes have enormous morphological diversity on the cellular as well as on the organismal level. They are classified into four kingdoms: Protista, Plantae, Fungi, and Animalia (including Human). Fungi are organisms such as mushrooms or yeasts. Mushrooms are multicellular eukaryotes. Yeasts are single-celled eukaryotes.

Unlike bacteria and archaea, eukaryotic cells have a nucleus. The nucleus is one kind of membrane-bound organelle. A cell’s nucleus holds the cell’s DNA. Eukaryotic cells have other membrane-bound organelles as well. Organelles are like the different organs in your body. Each kind of organelle has a specific job in the cell. Together, organelles perform all the processes necessary for life (Fig. 2.2).

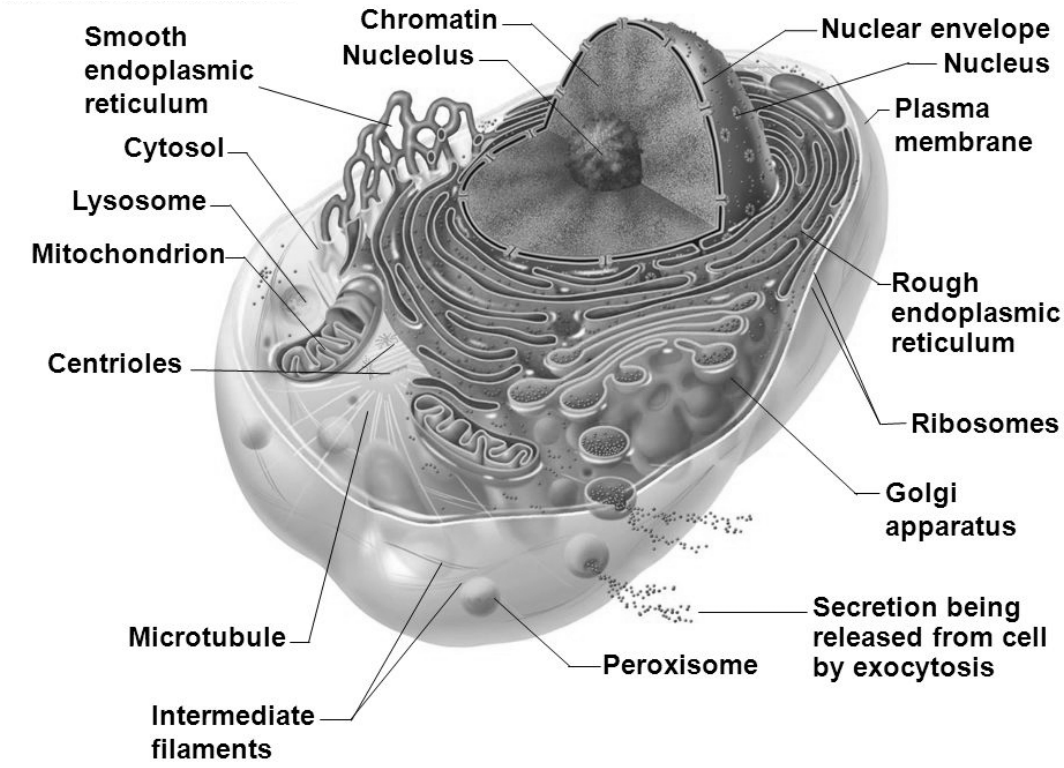


Figure 2.2: Eukaryote organelle schema (Pearson Education site).

Eukaryotic cells, which are far more complex than those of prokaryotes, are characterized by having numerous membrane-enclosed organelles. The most conspicuous of these is the nucleus, which contains the cell’s chromosomes, and the nucleolus, where ribosomes are assembled. The endoplasmic reticulum is the site of

synthesis of lipids and of proteins that are destined for secretion. Further processing of these products occurs in the Golgi apparatus. The mitochondria, wherein oxidative metabolism occurs, are thought to have evolved from a symbiotic relationship between an aerobic bacterium and a primitive eukaryote. Other eukaryotic organelles include the lysosome, which functions as an intracellular digestive chamber, and the peroxisome, which contains a variety of oxidative enzymes. The eukaryotic cytoplasm is pervaded by a cytoskeleton whose components include microtubules, which consist of tubulin; microfilaments, which are composed of actin; and intermediate filaments, which are made of different proteins in different types of cells.

2.2. Central Dogma of Molecular Biology

The central dogma of molecular biology explains that DNA codes for RNA, which codes for proteins and it was first stated by Francis Crick in 1956 (Crick, 1956, 1958) and re-stated in a Nature paper in 1970 (Crick, 1970).

In this thesis, I will only focus in the simplified version of Central Dogma i.e., "DNA → RNA → Protein" (Fig. 2.3).

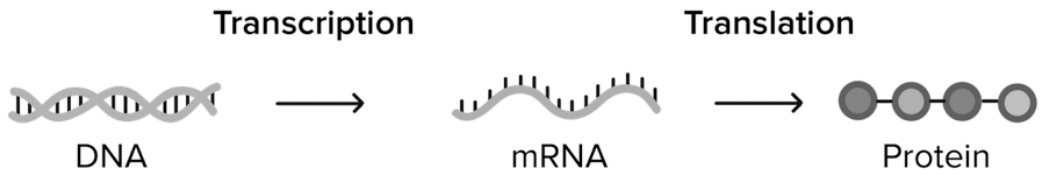


Figure 2.3: The Central Dogma of Molecular Biology (Wikipedia site).

DNA is the molecule of heredity that passes from parents to offspring. It contains the instructions for building RNA and proteins, which make up the structure of the body and carry out most of its functions.

Inside the cells, tiny molecular machines are constantly reading the information in DNA and using it to build proteins. In Figure 2.4 you can see all the three types of RNA that are essential to this process: messenger RNA (mRNA), transfer RNA (tRNA), and ribosomal RNA (rRNA).

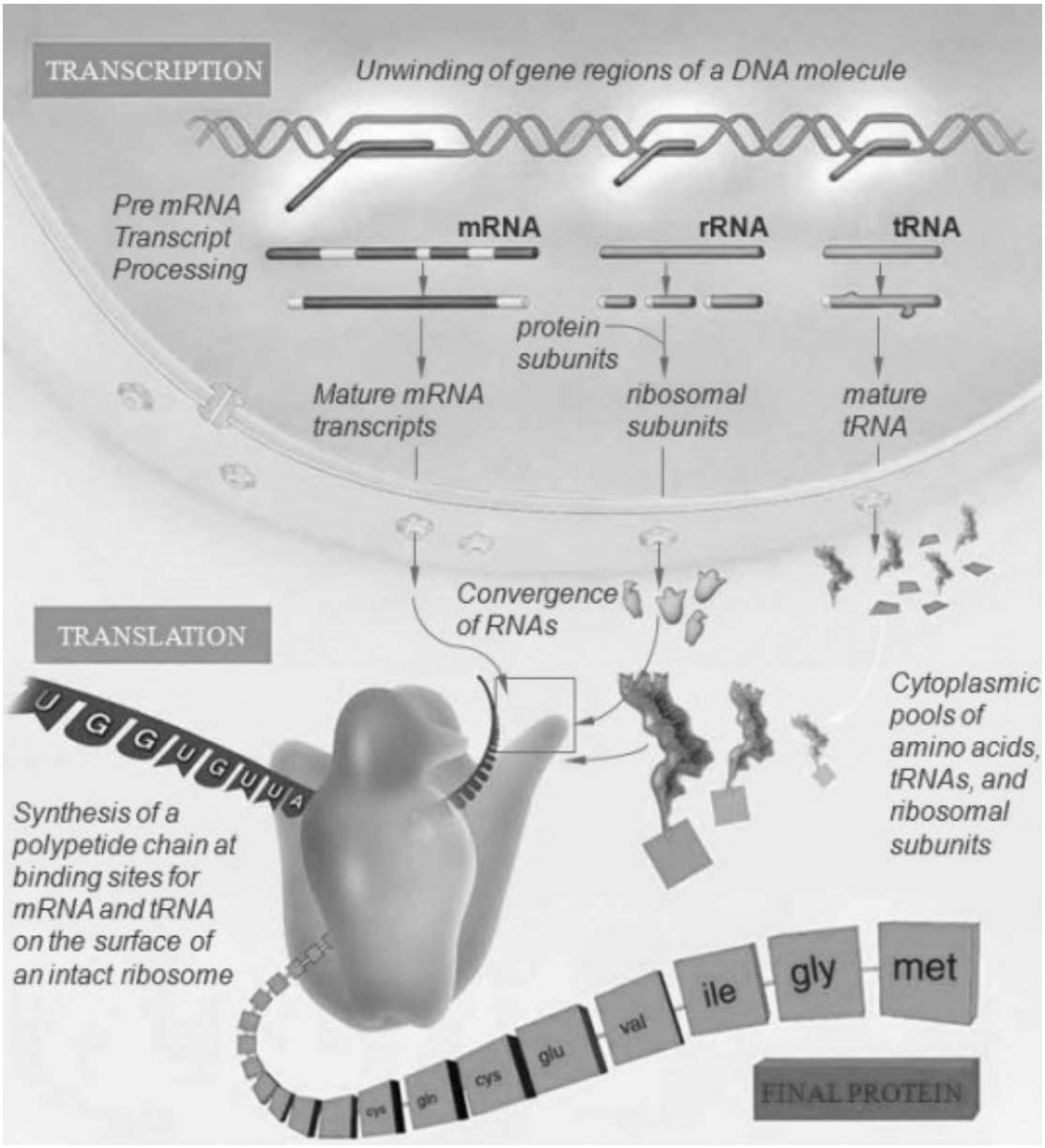


Figure 2.4: Simplified Central Dogma of Molecular Biology that can be described as "DNA makes RNA and RNA makes Protein" (Thomson Learning site).

Transcription is the process (Fig. 2.5) by which the information contained in a section of DNA is replicated in the form of a newly assembled piece of mRNA. In eukaryotic cells, the primary transcript is (pre-mRNA). Pre-mRNA must be processed for translation to proceed. Processing includes the addition of a 5' cap and a 3' tail to the pre-mRNA chain, followed by splicing. Alternative Splicing may occur broaden the diversity of the proteins that any single mRNA can generate. The product of the entire transcription process that began with the production of the pre-mRNA chain, is a mature mRNA chain, and they carry information from the genome to the ribosome (the cell's

protein synthesis instrument). The tRNA molecules are untranslated RNA that transport amino acids, the building blocks of proteins, to the ribosome. Last, but not least rRNA molecules are the untranslated RNA components of ribosomes, which are complexes of protein and RNA. The rRNAs play a role in anchoring the mRNA to the ribosomes.

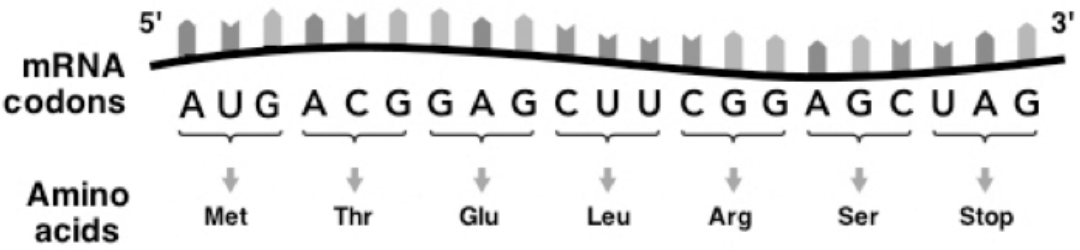


Figure 2.5: Transcription process (Wikipedia site).

Translation is the process by which the information contained in a section of all the three RNA's is used to form proteins (Figs. 2.6 and 2.7).

The mature mRNA finds its way to a ribosome, where it is translated. In prokaryotic cells, which have no nuclear compartment, the processes of transcription and translation may be linked together without clear separation. In eukaryotic cells, the site of transcription (the cell nucleus) is usually separated from the site of translation (the cytoplasm), so the mRNA must be transported out of the nucleus into the cytoplasm, where it can be bound by ribosomes. The ribosome reads the mRNA triplet codons, usually beginning with an AUG (adenine−uracil−guanine), or initiator methionine (Met) codon downstream of the ribosome binding site. Complexes of initiation factors and elongation factors bring aminoacylated transfer RNAs (tRNAs) into the ribosome-mRNA complex, matching the codon in the mRNA to the anti-codon on the tRNA. Each tRNA bears the appropriate amino acid residue to add to the polypeptide chain being synthesized. As the amino acids get linked into the growing peptide chain, the chain begins folding into the correct conformation. Translation ends with a stop codon, which may be a UAA, UGA, or UAG triplet.

The mRNA does not contain all the information for specifying the nature of the mature protein. The nascent polypeptide chain released from the ribosome commonly requires additional processing before the final product emerges. The correct folding process is complex and vitally important but is beyond the scope of this thesis.

First letter	Second letter: U	Second letter: C	Second letter: A	Second letter: G	Third letter
U	UUU Phe UUC Phe UUA Leu UUG Leu	UCU Ser UCC Ser UCA Ser UCG Ser	UAU Tyr UAC Tyr **UAA Stop** **UAG Stop**	UGU Cys UGC Cys **UGA Stop** UGG Trp	U C A G
C	CUU Leu CUC Leu CUA Leu CUG Leu	CCU Pro CCC Pro CCA Pro CCG Pro	CAU His CAC His CAA Gln CAG Gln	CGU Arg CGC Arg CGA Arg CGG Arg	U C A G
A	AUU Ile AUC Ile AUA Ile AUG Met	ACU Thr ACC Thr ACA Thr ACG Thr	AAU Asn AAC Asn AAA Lys AAG Lys	AGU Ser AGC Ser AGA Arg AGG Arg	U C A G
G	GUU Val GUC Val GUA Val GUG Val	GCU Ala GCC Ala GCA Ala GCG Ala	GAU Asp GAC Asp GAA Glu GAG Glu	GGU Gly GGC Gly GGA Gly GGG Gly	U C A G

Figure 2.6: Genetic code for translating each nucleotide triplet in mRNA into an amino acid or a termination signal in a nascent protein (National Institutes of Health site).

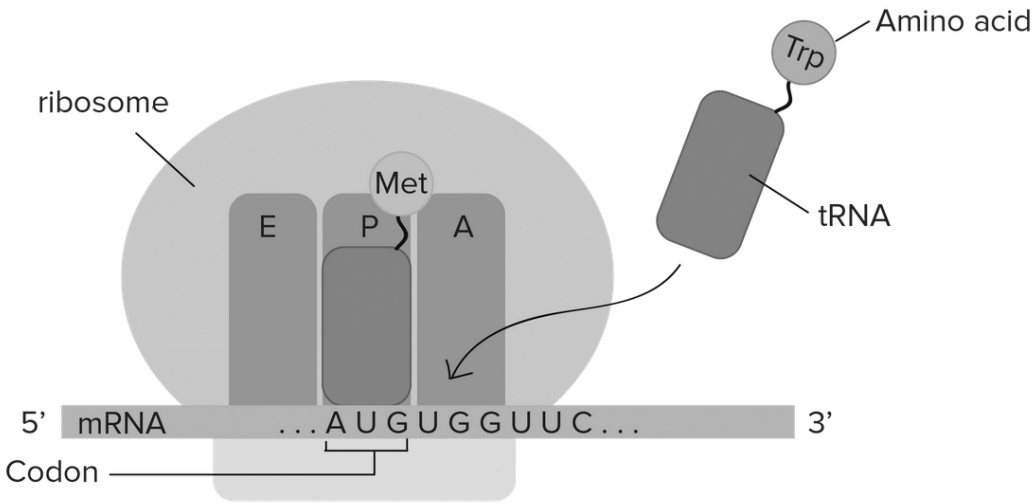


Figure 2.7: Translation process (OpenStax College site).

3. Proteins

The main idea of this chapter is to stress the importance of protein level structures that exist, as well as protein types, keeping in mind the sequence → structure → function paradigm (Fig. 3.1). In short, we need to know the structure of a protein to understand its function. Therefore, the ideal case would be to obtain all the proteins structures to understand all their functions consequently.

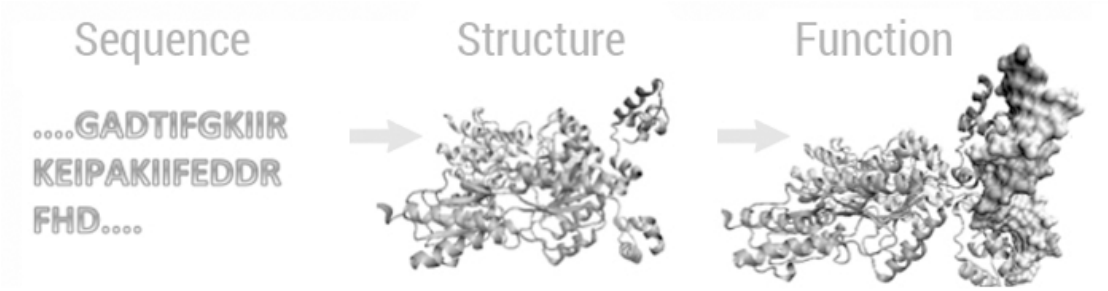


Figure 3.1: Sequence → Structure → Function paradigm (Griffith University site).

I start with the introduction of some very basic knowledge about the physics and chemistry of protein structures followed by its types. There are highly recommendable textbooks in molecular biology that give introduction to protein science from many different perspectives, in this chapter we followed the 'Encyclopedia of Molecular Biology' (Meyers, 2005). Concerning the buildup to protein backbone by elementary atomic constituents there are only a few rules to learn and therefore it is very easy to acquire the basic knowledge about the assembly of a realistic, plastic toy model. These rules are derived from quantum chemistry.

Proteins are long chain polymers of amino acids. They are linear, non-branched similar to polyethylene or polystyrene but with a much more versatile nature than the latter due to the many different types of amino acids involved.

3.1. Proteins Structure

3.1.1. Primary Structure

The sequence of the different amino acids is called the primary structure of the peptide or protein. Counting of residues always starts at the N-terminal end (NH_2-group), which is the end where the amino group is not involved in a peptide bond. The gene corresponding to the protein determines its primary structure. A specific sequence of nucleotides in DNA is transcribed into mRNA, which is read by the ribosome in a

process called translation. The sequence of a protein is unique to that protein, and defines the structure and function of the protein. The sequence of a protein can be determined by methods such as tandem mass spectrometry. Often however, it is read directly from the sequence of the gene using the genetic code.

The 20 different amino acids (Fig. 3.2) consists of four chemical groups: a nitrogen containing amino group, a carboxyl group, a central or "alpha" carbon in common, but each with a different/variable radical (the side-chain) attached to a carbon atom termed the C_α atom (Fig. 3.3).

	One-letter-code	Three-letter-code	Name	Hydrophobic
1	**A**	**Ala**	**Alanine**	**yes**
2	**C**	**Cys**	**Cysteine**	**yes**
3	**D**	**Asp**	**Aspartic Acid**	**no**
4	**E**	**Glu**	**Glutamic Acid**	**yes**
5	**F**	**Phe**	**Phenylalaline**	**yes**
6	**G**	**Gly**	**Glycine**	**no**
7	**H**	**His**	**Histine**	**no**
8	**I**	**Ile**	**Isoleucine**	**yes**
9	**K**	**Lys**	**Lysine**	**no**
10	**L**	**Leu**	**Leucine**	**yes**
11	**M**	**Met**	**Methionine**	**yes**
12	**N**	**Asn**	**Asparagine**	**no**
13	**P**	**Pro**	**Proline**	**yes**
14	**Q**	**Gln**	**Glutamine**	**no**
15	**R**	**Arg**	**Arginine**	**no**
16	**S**	**Ser**	**Serine**	**no**
17	**T**	**Thr**	**Threonine**	**no**
18	**V**	**Val**	**Valine**	**yes**
19	**W**	**Trp**	**Tryptophan**	**yes**
20	**Y**	**Tyr**	**Tyrosine**	**no**

Figure 3.2: The twenty amino acids commonly found in proteins.

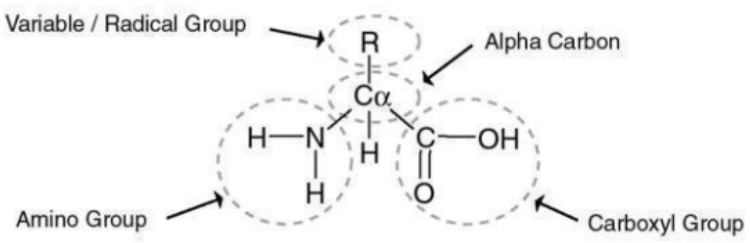


Figure 3.3: Schematic of amino acids (Hunter, 2009).

The amino, or more often called the peptide, links are connected to each other in a linear fashion such that the carbonyl end of one link is connected to the amino end of the next link and so that the resulting polypeptide chain (the protein without the side-chains) has a clear orientation.

Thus a protein molecule has a fairly simple structure with respect to its atomic constituents being first a nitrogen atom followed by a carbon atom with a side-chain (1 out of 20) attached to it and then finally followed by another carbon atom with an oxygen attached to it (Fig. 3.4)

The remaining sites are occupied by hydrogen atoms. This peptide unit is repeated typically several hundred times (for an average size protein) but mostly with a different side-chain attached to the C_α atom. The link between each amino acid connecting the carbonyl end with the next amino end has a partial double bonded nature that makes the peptide chain fairly rigid.

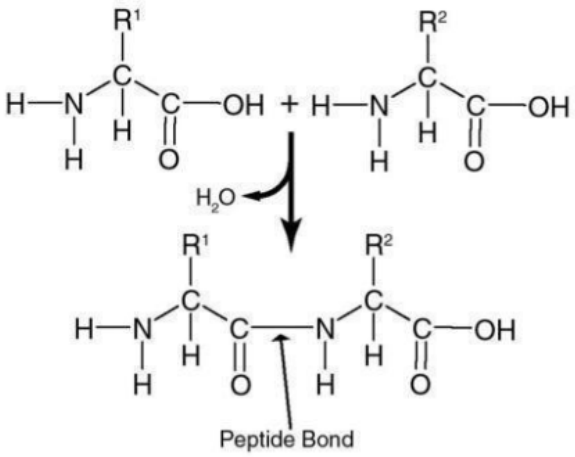


Figure 3.4: The polypeptide chains of proteins have a main chain of constant structure and sidechains that vary in sequence. Here R^1 and R^2 represent side chains. The side chains may be chosen, independently, from the set of 20 standard amino acids (Hunter, 2009).

Some parts of this polypeptide chain are flexible. Most bonds have a narrow range of angles that are energetically favorable, so the shapes of molecules containing them are effectively fixed, but there are three places where the bond angles are free to rotate. Most important, the two dihedral angles between adjacent amino acids (Φ and Ψ) can rotate freely, as shown in Figure 3.5.

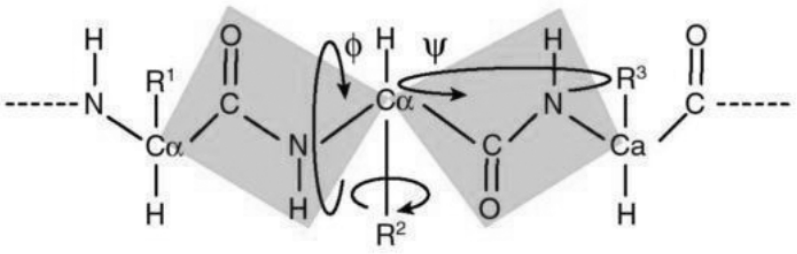


Figure 3.5: A schematic showing the three places in a polypeptide where the bonds are free to rotate (Hunter, 2009).

The chemical activity of this polypeptide chain is for the most parts controlled by the electrostatic nature of the different side-chains. These 20 common amino acids can be derived into polar and non-polar where the polar ones can be either charged positive (basic hydrophilic) or negative (acidic hydrophilic) or neutral. The non-polar amino acids are to a higher or lesser degree hydrophobic. The role of being hydrophilic or hydrophobic (turning towards into or away from water molecules) becomes, an important factor in the folding process when the protein is attaining its "native" active structure.

3.1.2. Secondary Structure

As we saw from the last section there appears a universal pattern in the 'local' structure of almost all proteins known up to now. The fact is that there appear distinct substructures in each protein that can be classified to be either helical, sheet or coil (this last class include single loops or turns). These distinct substructures are stabilized by hydrogen bonds giving rise to the usual classifying criteria for the substructures.

The most frequently occurring helical structure is the α-helix with 3.8 residues per turn and that is the mostly found to be right handed. The fractional number of residues per winding is because it provides the helical element with maximal stability since the hydrogen bonds appear asymmetrical in that case (with respect to the cylindrical symmetry).

In Fig. 3.6 the α-helix and the β-sheets are shown. The last ones can occur both as parallel or anti-parallel patterns and are the dominant substructures in immuneglobin and most proteases. These substructures are called the secondary structures because they occur on the second hierarchical level of organization, the first level being the sequence and the third level being the tertiary structures, the end product of the folding process. There are been an extensive effort in the field to produce prediction schemes

that could determine the occurrence of these structures from sequence information. These secondary structures will again arrange themselves into tertiary, or sometimes even into quaternary structures (protein-protein interactions).

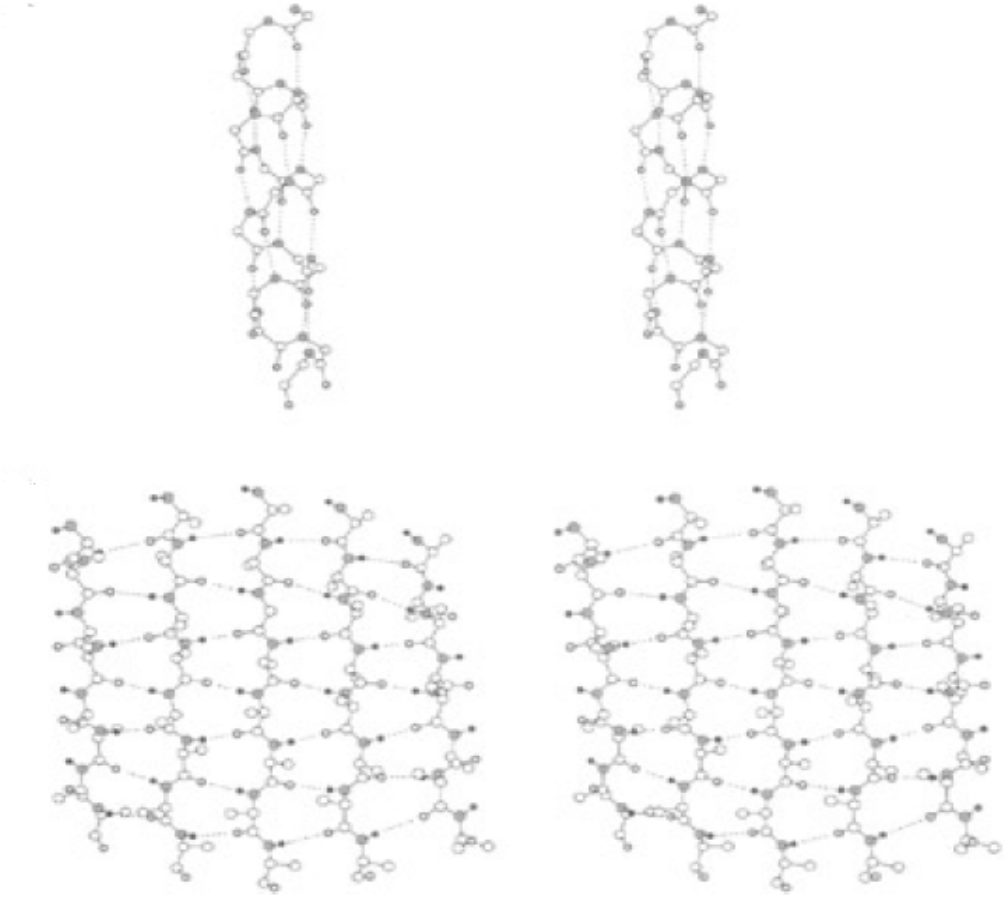


Figure 3.6: Standard secondary structures of proteins α-helix (above) β-sheet (below) (Lesk, 2002).

3.1.3. Tertiary Structure

The elements of secondary structure are usually folded into a compact shape using a variety of loops and turns, i.e., the information about the precise position of every atom (or, equivalently, every Φ/Ψ and sidechain angles) is called the tertiary structure of the protein. The formation of tertiary structure is usually driven by the burial of hydrophobic residues, but other interactions such as hydrogen bonding, ionic interactions can also stabilize the tertiary structure (Fig. 3.7). The tertiary structure encompasses all the non-covalent interactions that are not considered secondary structure, and is what defines the overall fold of the protein, and is usually indispensable for the function of the protein.

3.1.4. Quaternary Structure

The quaternary structure is the interaction between several chains of peptide bonds (Figs. 3.7). The individual chains are called subunits. The individual subunits are usually not covalently connected, but might be connected by a disulfide bond. Not all proteins have quaternary structure, since they might be functional as monomers. The quaternary structure is stabilized by the same range of interactions as the tertiary structure. Complexes of two or more polypeptides (i.e. multiple subunits or chains) are called multimers. Specifically, it would be called a dimer if it contains two subunits, a trimer if it contains three subunits, and a tetramer if it contains four subunits. The subunits are usually related to one another by symmetry axes, such as a 2-fold axis in a dimer. Multimers made up of identical subunits may be referred to with a prefix of "homo-" (e.g. a homotetramer) and those made up of different subunits may be referred to with a prefix of "hetero-" (e.g. a heterotetramer, such as the two α and two β chains of hemoglobin) (Fig. 3.9).

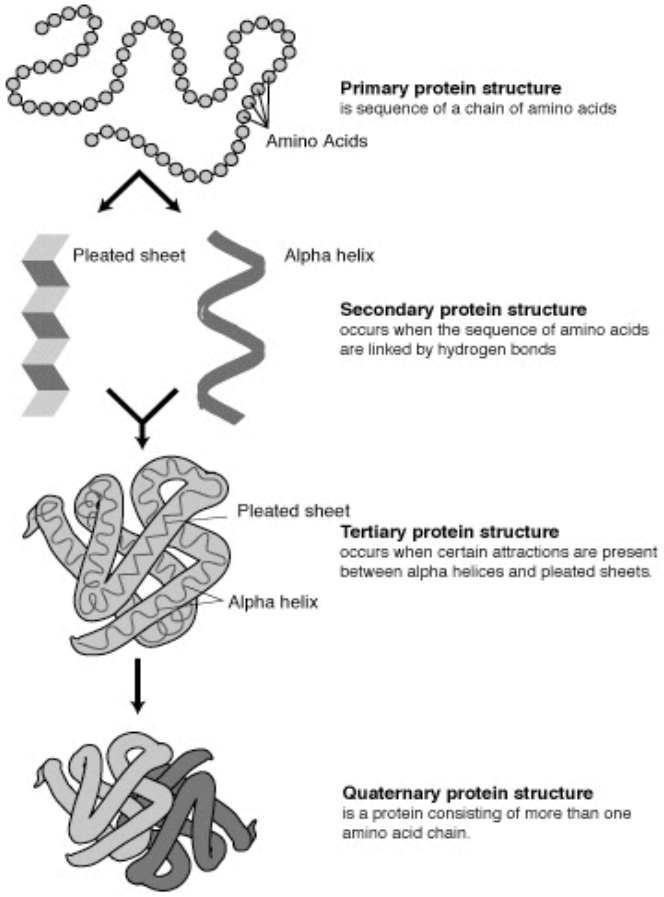


Figure 3.7: Primary, secondary, tertiary and quaternary structures of proteins (Lesk, 2002).

3.2. Proteins Types

Exist four main types of proteins: fibrous, globular, membrane and disordered (Meyers, 2005). These types will be briefly introduced in this chapter.

3.2.1. Fibrous Proteins

Fibrous proteins are extremely elongated molecules whose secondary structures are their dominant structural motifs such as α-helixes and β-sheets (Fig. 3.8). There are many fibrous proteins such as those of skin, tendon, and bone, function as structural materials having a connective, supportive or protective role in organisms. Others have motive functions, such as muscle. The structural simplicity of these proteins relative to globular proteins makes them particularly conformable to understanding how their structures suit them to their biological roles.

Fibrous molecules rarely crystallize and hence are usually not subject to structural determination by single-crystal X-ray structure analysis. Therefore, solid state Nuclear Magnetic Resonance (NMR) became a very useful as an alternative means to fiber X-ray diffraction (Parry & Squire, 1998).

Figure 3.8: Fibrous protein structure – triple α-helix collagen (Wikipedia site).

3.2.2. Globular Proteins

Globular proteins are somewhat water-soluble, unlike the fibrous or membrane proteins (Figs. 3.9).

Comprise a highly diverse group of substances that, in their native states, exist as compact spheroidal molecules (globular or coiled shape). The spherical structure of these proteins is induced by the protein's tertiary structure or quaternary (when chains are involved). Enzymes are globular proteins, as are receptor and transport proteins.

Most of their detailed structural knowledge of these proteins, and also a large extent their function, has resulted from X-ray crystal structure determinations of globular

proteins and, more recently, from their nuclear magnetic resonance (NMR) structure determination (Travaglini-Allocatelli et al., 2009).

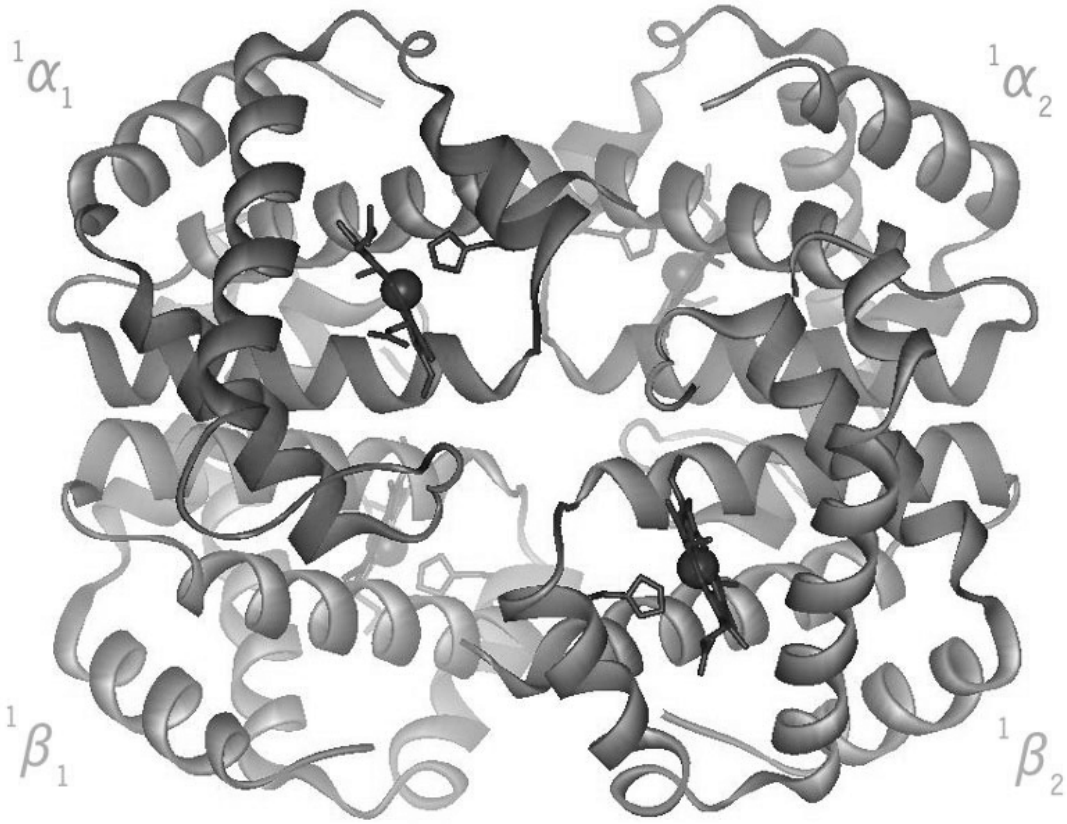


Figure 3.9: Globular protein structure – human hemoglobin heterotetramer (Wikipedia site).

3.2.3. Membrane Proteins

Membrane proteins are operationally classified according to how tightly they are associated with membranes (Fig. 3.10):

a) Integral or intrinsic proteins are permanently bound to membranes and can only be separated from them by treatment with agents that disrupt membranes. These include organic solvents and detergents. It has been shown that some integral proteins are exposed only to a specific surface of a membrane, whereas others, known as transmembrane proteins, span the membrane. Transmembrane proteins may have different transmembrane topology such as: 1) a single pass α-helix protein; 2) a multi pass transmembrane α-helical protein; 3) and finally multi pass transmembrane β-sheet protein.

b) Peripheral or extrinsic proteins are dissociated from membranes by relatively soft procedures that leave the membrane intact, such as exposure to high salt solutions, or high pH values. Peripheral proteins are stable in aqueous solution and do not bind to lipids. They associate with a membrane by binding at its 1) surface to its lipid head groups and/or 2) its integral proteins through hydrogen bonding and electrostatic interactions.

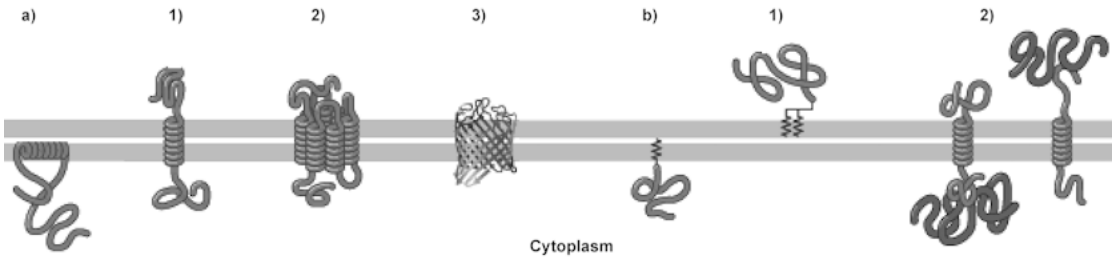


Figure 3.10: Membrane proteins types: a) Integral or intrinsic proteins 1) Single pass α-helix 2) a multi pass transmembrane α-helical 3) multi pass transmembrane β-sheet; b) Peripheral or extrinsic proteins associate with a membrane by binding at its 1) surface to its lipid head groups and/or 2) integral proteins through hydrogen bonding and electrostatic interactions (University of Tokyo site).

Membrane proteins represent between 20 and 30% of the proteomes of most organisms (Krogh et al., 2001), and more than 40% of these transmembrane proteins are the target for modern drugs (Overington et al., 2006), and yet very few structures of these molecules have been solved by X-ray crystallography or NMR (Carpenter et al., 2008). Therefore, the determination of structures for this this type of protein remains a challenge in large part due to the hardness in establishing experimental conditions where the correct conformation of the protein in isolation from its native environment is preserved.

3.2.4. Intrinsically Disordered Proteins

Intrinsically Disordered Proteins (IDP's) lack fixed or ordered tertiary structure (Fig. 3.11) and are therefore composed by ensembles of conformations (Dunker et al., 2001).

In the first half of the 20th century, protein structures were solved by protein crystallography. These initial structures suggested that a fixed 3D structure might be required to establish biological functions of proteins (Mirsky & Paulin, 1936; Pauling

& Coryell, 1936). These publications solidified the central dogma "sequence → structure → function" in proteins.

During the subsequent decades, however, many large protein regions could not be assigned in x-ray datasets, indicating that they occupy multiple positions, which average out in electron density maps. Additional techniques for determining protein structures, such as NMR, demonstrated the presence of large flexible linkers and termini in many solved structural ensembles. This leaded to a theory that stated that "proteins must be properly folded in order to perform their functions" (Anfinsen, 1973) achieving the Nobel in 1972.

Around year 2000 it was recognized that not all proteins function in a folded state (Wright & Dyson, 1999; Dunker et al., 2001). Some proteins must be unfolded or disordered in order to perform their functions, and others bind to some other molecule such as a protein, a nucleic acid, or a membrane component (targets), and in doing so fold into stable tertiary structures (Bright et al., 2001). These are termed intrinsically disordered protein (IDP).

IDPs therefore, defied the protein structure paradigm, where protein function depends on a stable 3D structure. Their most common function appears to be binding to specific DNA sequences to facilitate processes like replication, transcription, transposition and repair. However, they are also refereed in several other functions including intracellular signaling and in aiding other proteins and RNAs to fold to their native conformations (Dyson & Wright, 2005; Dunker et al., 2008).

IDPs have been implicated recently in a number of diseases (Uversky et al., 2008).

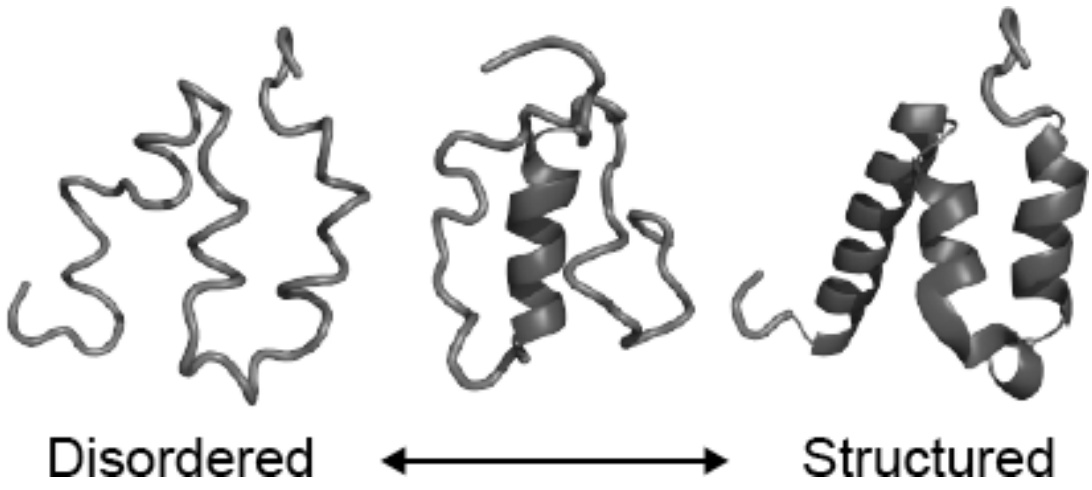


Figure 3.11: Intrinsically disorder protein (University of Kansas site).

3.3. Proteins Databases

In this section it will be described the most relevant protein databases of sequences (1D) and structures (3D) for this thesis. The repositories of proteins detailed here are the most commonly used and are a reference by themselves. There are others repositories but they are not relevant for the work that will be presented later.

3.3.1. Sequence Databases

Swiss-Prot

The Swiss-Prot (Bairoch & Apweiler, 2000) protein knowledge base is an annotated protein sequence database that is maintained collaboratively by the European Bioinformatics Institute (EBI) and the Swiss Institute of Bioinformatics (SIB) since 1986. The database is non-redundant, meaning that many pages of scientific literature are condensed in a single entry, being aimed to provide a reliable protein sequences associated with high level of annotation through a process of literature-based manual curation. This includes descriptions of the function(s) of the protein, post-translational modifications, domains, similarities to other proteins (secondary and quaternary structure), developmental stages in which the protein is expressed, tissues locations, pathways, sequence conflicts and variants.

Swiss-Prot contains data from a wide variety of organisms: as of July 2016, release 2016_06 of 06-Jul-16 contained 551,705 annotated sequence entries (Fig. 3.12) from almost 13328 different species. Figure 3.13 shows a sequence entry from Swiss-Prot.

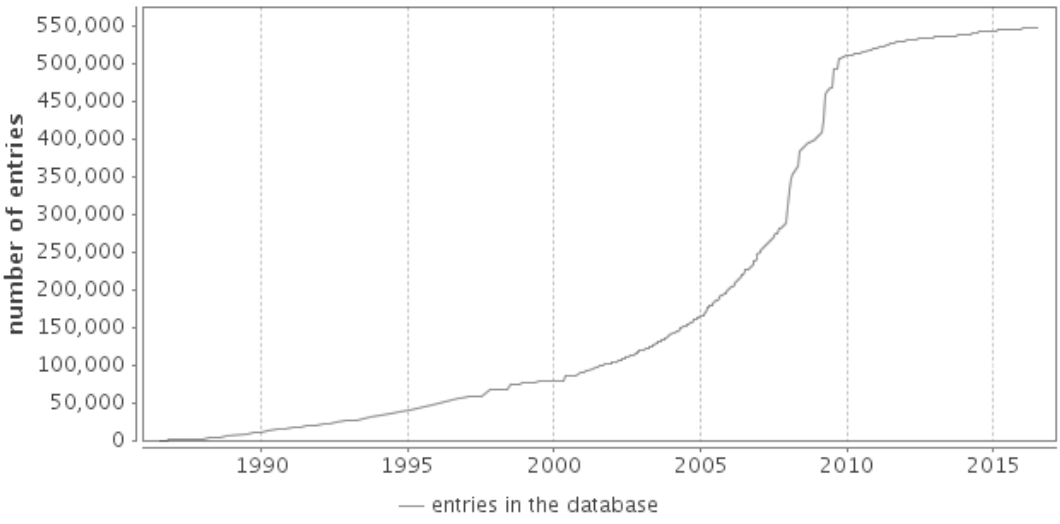


Figure 3.12: Number of entries in Swiss-Prot over time (UniProt site)

```
ID   4EBP2_HUMAN             Reviewed;         120 AA.
AC   Q13542;
DT   19-SEP-2003, integrated into UniProtKB/Swiss-Prot.
DT   01-NOV-1996, sequence version 1.
DT   06-JUL-2016, entry version 132.
DE   RecName: Full=Eukaryotic translation initiation factor 4E-binding protein 2 {PubMed:7935836};
DE            Short=4E-BP2 {ECO:0000303|PubMed:7935836};
DE            Short=eIF4E-binding protein 2 {ECO:0000303|PubMed:7935836};
GN   Name=EIF4EBP2 {ECO:0000312|HGNC:HGNC:3289};
OS   Homo sapiens (Human).
OC   Eukaryota; Metazoa; Chordata; Craniata; Vertebrata; Euteleostomi;
OC   Mammalia; Eutheria; Euarchontoglires; Primates; Haplorrhini;
OC   Catarrhini; Hominidae; Homo.
OX   NCBI_TaxID=9606 {ECO:0000312|EMBL:AAH05057.1};
RN   [1]
RP   NUCLEOTIDE SEQUENCE [MRNA], AND INTERACTION WITH EIF4E.
RC   TISSUE=Placenta;
RX   PubMed=7935836; DOI=10.1038/371762a0;
RA   Pause A., Belsham G.J., Gingras A.-C., Donze O., Lin T.-A.,
RA   Lawrence J.C. Jr., Sonenberg N.;
RT   "Insulin-dependent stimulation of protein synthesis by phosphorylation
RT   of a regulator of 5'-cap function.";
RL   Nature 371:762-767(1994).
RN   [2] {ECO:0000312|EMBL:AAP35981.1}
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE MRNA].
RA   Kalnine N., Chen X., Rolfs A., Halleck A., Hines L., Eisenstein S.,
RA   Koundinya M., Raphael J., Moreira D., Kelley T., LaBaer J., Lin Y.,
RA   Phelan M., Farmer A.;
RT   "Cloning of human full-length CDSs in BD Creator(TM) system donor
RT   vector.";
RL   Submitted (MAY-2003) to the EMBL/GenBank/DDBJ databases.
RN   [3] {ECO:0000312|EMBL:AAH05057.1}
RP   NUCLEOTIDE SEQUENCE [LARGE SCALE MRNA].
RC   TISSUE=Lung {ECO:0000312|EMBL:AAH05057.1}, and
RC   Uterus {ECO:0000312|EMBL:AAH50633.1};
RX   PubMed=15489334; DOI=10.1101/gr.2596504;
RG   The MGC Project Team;
RT   "The status, quality, and expansion of the NIH full-length cDNA
RT   project: the Mammalian Gene Collection (MGC).";
RL   Genome Res. 14:2121-2127(2004).
RN   [4]
RP   PHOSPHORYLATION [LARGE SCALE ANALYSIS] AT THR-37 AND THR-46, AND
RP   IDENTIFICATION BY MASS SPECTROMETRY [LARGE SCALE ANALYSIS].
RC   TISSUE=Cervix carcinoma;
RX   PubMed=18669648; DOI=10.1073/pnas.0805139105;
RA   Dephoure N., Zhou C., Villen J., Beausoleil S.A., Bakalarski C.E.,
RA   Elledge S.J., Gygi S.P.;
RT   "A quantitative atlas of mitotic phosphorylation.";
RL   Proc. Natl. Acad. Sci. U.S.A. 105:10762-10767(2008).
RN   [5]
RP   IDENTIFICATION BY MASS SPECTROMETRY [LARGE SCALE ANALYSIS].
RX   PubMed=19413330; DOI=10.1021/ac9004309;
RA   Gauci S., Helbig A.O., Slijper M., Krijgsveld J., Heck A.J.,
RA   Mohammed S.;
RT   "Lys-N and trypsin cover complementary parts of the phosphoproteome in
RT   a refined SCX-based approach.";
RL   Anal. Chem. 81:4493-4501(2009).
RN   [6]
RP   PHOSPHORYLATION [LARGE SCALE ANALYSIS] AT THR-37, AND IDENTIFICATION
RP   BY MASS SPECTROMETRY [LARGE SCALE ANALYSIS].
RC   TISSUE=Leukemic T-cell;
RX   PubMed=19690332; DOI=10.1126/scisignal.2000007;
RA   Mayya V., Lundgren D.H., Hwang S.-I., Rezaul K., Wu L., Eng J.K.,
RA   Rodionov V., Han D.K.;
RT   "Quantitative phosphoproteomic analysis of T cell receptor signaling
RT   reveals system-wide modulation of protein-protein interactions.";
RL   Sci. Signal. 2:RA46-RA46(2009).
RN   [7]
RP   IDENTIFICATION BY MASS SPECTROMETRY [LARGE SCALE ANALYSIS].
RC   TISSUE=Erythroleukemia;
RX   PubMed=23186163; DOI=10.1021/pr300630k;
RA   Zhou H., Di Palma S., Preisinger C., Peng M., Polat A.N., Heck A.J.,
RA   Mohammed S.;
RT   "Toward a comprehensive characterization of a human cancer cell
RT   phosphoproteome.";
RL   J. Proteome Res. 12:260-271(2013).
RN   [8]
RP   DOMAIN, INTERACTION WITH EIF4E, AND MUTAGENESIS OF 54-TYR--LEU-59.
RX   PubMed=24207126; DOI=10.1016/j.str.2013.08.030;
RA   Lukhele S., Bah A., Lin H., Sonenberg N., Forman-Kay J.D.;
RT   "Interaction of the eukaryotic initiation factor 4E with 4E-BP2 at a
RT   dynamic bipartite interface.";
RL   Structure 21:2186-2196(2013).
RN   [9]
RP   X-RAY CRYSTALLOGRAPHY (2.2 ANGSTROMS) OF 47-65 IN COMPLEX WITH EIF4E,
RP   INTERACTION WITH EIF4E, AND PHOSPHORYLATION.
RX   PubMed=21661078; DOI=10.1002/psc.1384;
```

Figure 3.13: Swiss-Prot text file (partial) for protein Q13542 (to see the full file please consult Supplementary File I)

TrEMBL

The production of a fully curated Swiss-Prot entry is a highly labor-intensive process and is the rate-limiting step in the growth of the database. This is because, with the increased data flow from genome projects, new sequences are submitted more quickly than they can be manually annotated and integrated into the database. To address this, a supplement to Swiss-Prot was created in 1996 to fulfill the vital role of making new sequences available as quickly as possible while preventing the dilution of the high quality annotation found in Swiss-Prot. This supplement, TrEMBL (Translation of EMBL nucleotide sequence database) (Bairoch & Apweiler, 2000), consists of computer-annotated entries derived from the translation of all coding sequences of nucleotide sequence in EMBL (Cochrane et al., 2008) / GenBank (Clark et al., 2016) /DNA Data Bank of Japan (DDBJ) (Mashima et al., 2016) databases that are not yet included in Swiss-Prot. To ensure completeness, it also contains a number of protein sequences extracted from the literature or submitted directly by the user community. TrEMBL follows the Swiss-Prot format and conventions as closely as possible.

In July 2016 TrEMBL holds 65.378.749 (Fig 3.14) proteins automatically annotated and not reviewed.

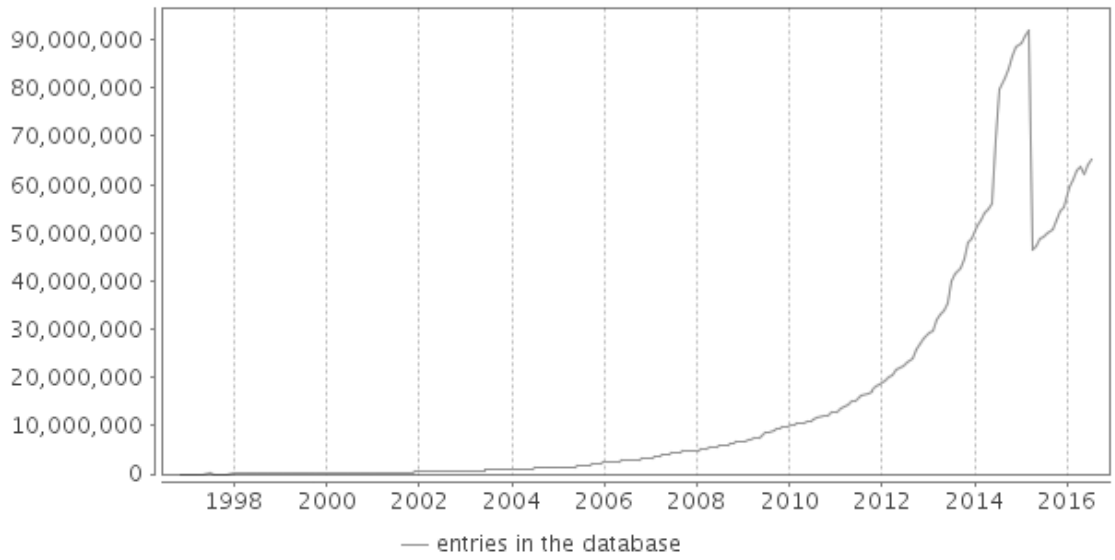


Figure 3.14: Number of entries in TrEMBL over time (UniProt site)

PIR-PSD

The Protein Identification Resource (PIR) Protein Sequence Database (PSD) (Wu et al., 2003) is a protein information resource initiated in 1961 by the National Biomedical Research Foundation (in USA), the Martinsried Institute for Proteins Sequence (in Europe), and the Japan International Protein Information Database (in Japan). It compiles comprehensive, non-redundant protein sequence data, organized by superfamily and family, and annotated with functional, structural, bibliographic and genetic data. The database also contains the name and classification of the protein, the name of the organism in which it naturally occurs, literature references, function and general characteristics of the protein, sites and regions at the sequence of biological interest. In 2002, PIR along with its international partners, EBI and SIB, were awarded a grant from National Institutes of Health (NIH) and other five institutions to create UniProt, a single worldwide database of protein sequence and function, by unifying the PIR-PSD, Swiss-Prot, and TrEMBL databases.

UniProt

In 2004 the Swiss-Prot, TrEMBL and PIR databases have joined forces to form the United Protein Databases (UniProt) (Apweiler et al., 2004). UniProt is funded by the US National Human Genome Research Institute, National Institutes of Health (NIH), European Commission, Swiss Federal Government, cancer Biomedical Informatics Grid (caBIG), and the Department of Defense (DOD).

Generation of genome sequences for many organisms is at peak actually, most notably human sequences, attention is now turning to the identification and function of proteins encoded by these genomes. A complete and up-to-date protein database is essential for the increasingly information dependent biological and biotechnological research, especially in proteomics. The long term objective of UniProt is therefore, to create, maintain and provide a stable, comprehensive, fully classified, and accurately annotated protein sequence knowledge base, with extensive cross-references.

3.3.2. Structure Databases

PDB

Protein Databank (PDB) (Bernstein, 1977) started by the late 1971 at Brookhaven National Laboratories, New York, USA. This database keeps experimentally derived three-dimensional structures of proteins determined by both X-Ray, Nuclear Magnetic

Resonance (NMR) and (more recently) Cryo-Electron Microscopy (cryo-EM) since then.

In 2005 the Research Collaboratory for Structural Bioinformatics (RCSB), the Protein Databank (PDB), the European Bioinformatics Institute (EBI) – PDB Europe (PDBe), and the Protein Data Bank Japan (PDBj) have formed the worldwide Protein Data Bank (wwPDB) (Berman et al., 2007), with the goal of producing a unified archive.

The wwPDB is now managed by RCSB, a distributed organization based at Rutgers University, in New Jersey; The San Diego Supercomputer Center in California; and the National Institute of Standards and Technology in Maryland, all in the USA. Some depositions are done to sites in the UK (PDBe) and Japan (PDBj). A single archive of annotated data is managed at RCSB. Mirror sites are available worldwide. Collectively, the RCSB provides the central PDB data repository. The other groups contribute to the deposition and annotation effort, but the RCSB is the only with "write privileges" for the publicly distributed archive, in order to help reduce the chances of creation of divergent versions.

PDB contains in July 2016 the total amount of 120262 biological structures (Fig. 3.15) and for each protein, a general header is provided followed by a list of all ATOMS present in the structure, with three spatial coordinates to indicate their position. Figure 3.16 displays a small sample of such a file.

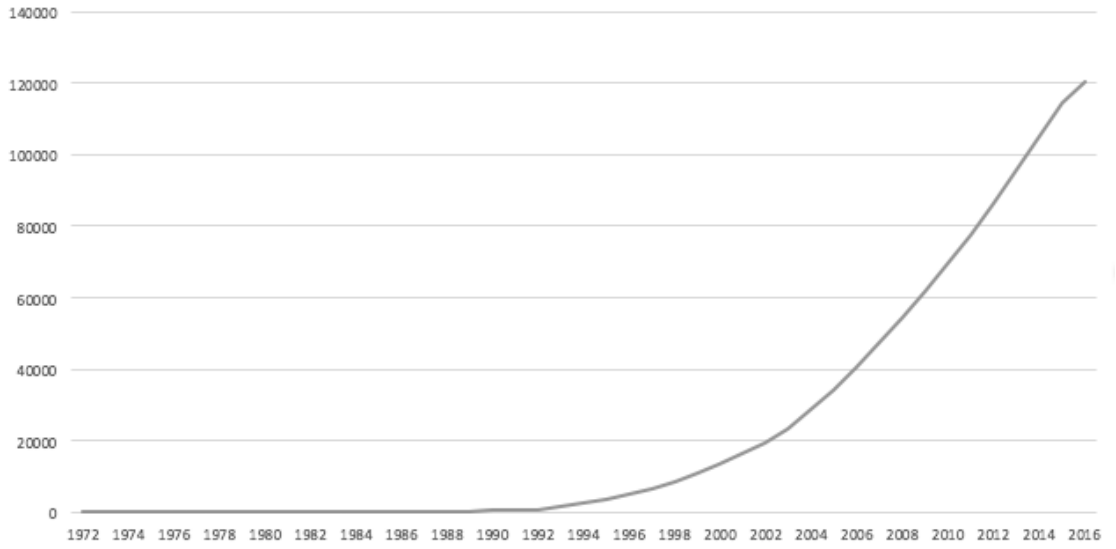


Figure 3.15: Number of entries in PDB over time (PDB site data)

```
ATOM 1     N N   . PRO A 1 1  ? 0.592  -0.732 -0.717  1.00 0.00 ? ? ? ? ? ? 18 PRO A N   1
ATOM 2     C CA  . PRO A 1 1  ? 1.885  -0.592 -0.059  1.00 0.00 ? ? ? ? ? ? 18 PRO A CA  1
ATOM 3     C C   . PRO A 1 1  ? 2.362  0.854  -0.085  1.00 0.00 ? ? ? ? ? ? 18 PRO A C   1
ATOM 4     O O   . PRO A 1 1  ? 1.950  1.639  -0.940  1.00 0.00 ? ? ? ? ? ? 18 PRO A O   1
ATOM 5     C CB  . PRO A 1 1  ? 2.811  -1.519 -0.853  1.00 0.00 ? ? ? ? ? ? 18 PRO A CB  1
ATOM 6     C CG  . PRO A 1 1  ? 2.241  -1.522 -2.230  1.00 0.00 ? ? ? ? ? ? 18 PRO A CG  1
ATOM 7     C CD  . PRO A 1 1  ? 0.751  -1.431 -2.046  1.00 0.00 ? ? ? ? ? ? 18 PRO A CD  1
ATOM 8     H H2  . PRO A 1 1  ? 0.497  -1.331 -1.512  1.00 0.00 ? ? ? ? ? ? 18 PRO A H2  1
ATOM 9     H H3  . PRO A 1 1  ? -0.194 -1.080 -0.205  1.00 0.00 ? ? ? ? ? ? 18 PRO A H3  1
ATOM 10    H HA  . PRO A 1 1  ? 1.853  -0.860 1.007   1.00 0.00 ? ? ? ? ? ? 18 PRO A HA  1
ATOM 11    H HB2 . PRO A 1 1  ? 3.848  -1.150 -0.852  1.00 0.00 ? ? ? ? ? ? 18 PRO A HB2 1
ATOM 12    H HB3 . PRO A 1 1  ? 2.829  -2.533 -0.427  1.00 0.00 ? ? ? ? ? ? 18 PRO A HB3 1
ATOM 13    H HG2 . PRO A 1 1  ? 2.620  -0.673 -2.819  1.00 0.00 ? ? ? ? ? ? 18 PRO A HG2 1
ATOM 14    H HG3 . PRO A 1 1  ? 2.518  -2.438 -2.773  1.00 0.00 ? ? ? ? ? ? 18 PRO A HG3 1
ATOM 15    H HD2 . PRO A 1 1  ? 0.266  -0.857 -2.849  1.00 0.00 ? ? ? ? ? ? 18 PRO A HD2 1
ATOM 16    H HD3 . PRO A 1 1  ? 0.273  -2.422 -2.027  1.00 0.00 ? ? ? ? ? ? 18 PRO A HD3 1
ATOM 17    N N   . THR A 1 2  ? 3.234  1.202  0.856   1.00 0.00 ? ? ? ? ? ? 19 THR A N   1
ATOM 18    C CA  . THR A 1 2  ? 3.762  2.558  0.948   1.00 0.00 ? ? ? ? ? ? 19 THR A CA  1
ATOM 19    C C   . THR A 1 2  ? 5.252  2.592  0.634   1.00 0.00 ? ? ? ? ? ? 19 THR A C   1
ATOM 20    O O   . THR A 1 2  ? 6.030  1.814  1.188   1.00 0.00 ? ? ? ? ? ? 19 THR A O   1
ATOM 21    C CB  . THR A 1 2  ? 3.528  3.161  2.345   1.00 0.00 ? ? ? ? ? ? 19 THR A CB  1
ATOM 22    O OG1 . THR A 1 2  ? 2.121  3.213  2.618   1.00 0.00 ? ? ? ? ? ? 19 THR A OG1 1
ATOM 23    C CG2 . THR A 1 2  ? 4.107  4.566  2.424   1.00 0.00 ? ? ? ? ? ? 19 THR A CG2 1
ATOM 24    H H   . THR A 1 2  ? 3.538  0.509  1.525   1.00 0.00 ? ? ? ? ? ? 19 THR A H   1
ATOM 25    H HA  . THR A 1 2  ? 3.276  3.194  0.207   1.00 0.00 ? ? ? ? ? ? 19 THR A HA  1
ATOM 26    H HB  . THR A 1 2  ? 4.010  2.528  3.090   1.00 0.00 ? ? ? ? ? ? 19 THR A HB  1
ATOM 27    H HG1 . THR A 1 2  ? 1.978  3.588  3.490   1.00 0.00 ? ? ? ? ? ? 19 THR A HG1 1
ATOM 28    H HG21 . THR A 1 2 ? 3.625  5.199  1.680   1.00 0.00 ? ? ? ? ? ? 19 THR A HG21 1
ATOM 29    H HG22 . THR A 1 2 ? 3.932  4.975  3.418   1.00 0.00 ? ? ? ? ? ? 19 THR A HG22 1
ATOM 30    H HG23 . THR A 1 2 ? 5.179  4.528  2.228   1.00 0.00 ? ? ? ? ? ? 19 THR A HG23 1
ATOM 31    N N   . ARG A 1 3  ? 5.644  3.496  -0.256  1.00 0.00 ? ? ? ? ? ? 20 ARG A N   1
ATOM 32    C CA  . ARG A 1 3  ? 7.042  3.634  -0.643  1.00 0.00 ? ? ? ? ? ? 20 ARG A CA  1
ATOM 33    C C   . ARG A 1 3  ? 7.554  5.042  -0.369  1.00 0.00 ? ? ? ? ? ? 20 ARG A C   1
ATOM 34    O O   . ARG A 1 3  ? 6.926  6.027  -0.758  1.00 0.00 ? ? ? ? ? ? 20 ARG A O   1
ATOM 35    C CB  . ARG A 1 3  ? 7.281  3.227  -2.090  1.00 0.00 ? ? ? ? ? ? 20 ARG A CB  1
ATOM 36    C CG  . ARG A 1 3  ? 8.713  3.393  -2.574  1.00 0.00 ? ? ? ? ? ? 20 ARG A CG  1
ATOM 37    C CD  . ARG A 1 3  ? 8.939  2.961  -3.976  1.00 0.00 ? ? ? ? ? ? 20 ARG A CD  1
ATOM 38    N NE  . ARG A 1 3  ? 10.298 3.154  -4.455  1.00 0.00 ? ? ? ? ? ? 20 ARG A NE  1
ATOM 39    C CZ  . ARG A 1 3  ? 10.745 2.770  -5.666  1.00 0.00 ? ? ? ? ? ? 20 ARG A CZ  1
ATOM 40    N NH1 . ARG A 1 3  ? 9.958  2.137  -6.509  1.00 0.00 ? ? ? ? ? ? 20 ARG A NH1 1
ATOM 41    N NH2 . ARG A 1 3  ? 12.005 3.024  -5.977  1.00 0.00 ? ? ? ? ? ? 20 ARG A NH2 1
ATOM 42    H H   . ARG A 1 3  ? 4.955  4.103  -0.676  1.00 0.00 ? ? ? ? ? ? 20 ARG A H   1
ATOM 43    H HA  . ARG A 1 3  ? 7.659  2.959  -0.048  1.00 0.00 ? ? ? ? ? ? 20 ARG A HA  1
ATOM 44    H HB2 . ARG A 1 3  ? 6.991  2.181  -2.178  1.00 0.00 ? ? ? ? ? ? 20 ARG A HB2 1
ATOM 45    H HB3 . ARG A 1 3  ? 6.621  3.839  -2.705  1.00 0.00 ? ? ? ? ? ? 20 ARG A HB3 1
ATOM 46    H HG2 . ARG A 1 3  ? 8.984  4.446  -2.500  1.00 0.00 ? ? ? ? ? ? 20 ARG A HG2 1
ATOM 47    H HG3 . ARG A 1 3  ? 9.366  2.804  -1.930  1.00 0.00 ? ? ? ? ? ? 20 ARG A HG3 1
ATOM 48    H HD2 . ARG A 1 3  ? 8.713  1.898  -4.059  1.00 0.00 ? ? ? ? ? ? 20 ARG A HD2 1
ATOM 49    H HD3 . ARG A 1 3  ? 8.277  3.525  -4.633  1.00 0.00 ? ? ? ? ? ? 20 ARG A HD3 1
ATOM 50    H HE  . ARG A 1 3  ? 11.089 3.583  -3.993  1.00 0.00 ? ? ? ? ? ? 20 ARG A HE  1
```

Figure 3.16: PDB file for intrinsically disordered protein 2mx4.pdb first 50 ATOM 3D coordinates (to see the full file please consult Supplementary File II)

3.3.3. Homology Databases

A Multiple Sequence Alignment (MSA) normally is used to detect homology structure among sequences based on database homologues search methods like BLAST and PSI-BLAST. In this section I'm going to present how these methods can be used to reduce the gap size between 3D databases like PDB (too small) in comparison with the databases size of known sequence information (1D) that have 1000 times more entries altogether like UniProt.

Usually, many proteins in the database of known sequences are similar in sequence to a protein of known structure and this fact can be exploited to close the size gap between the two databases. The technique used to close the gap is homology derived.

HSSP

Homology-derived StructureS of Proteins (HSSP) was a derived database merging structural (2D/ 3D) and sequence (1D) information. For each protein of known 3D structure from the PDB, the database had a text file with a Multiple Sequence Alignment (MSA) of all homologues, properly aligned to the PDB protein. Where homologues were very likely to have the same 3D structure as the PDB protein to which they have been aligned. As a result, the database was not only a database of sequence families aligned, but it was also a database of implied secondary and tertiary structures.

It is confirmed that structural homology can be inferred from the level of sequence identity, and that structural homology depends strongly on the length of the alignment (Sander and Schneider, 1991). The selection of sequence alignments with significant sequence identity (homology) to proteins of know structure leaded to a database of homology-derived protein structures several times larger than PDB. The intent of HSSP was to reduce the gap size between 3D protein databases like PDB (too small, only 694 entries in 1991) compared with the 1D database size of known sequences (Swiss-Prot) (around 10.000 sequences also in 1991).

Homology Thresholds

The transfer of structure information to a potentially homologous protein is straightforward when the sequence identity is high and extended in length, but the assessment of the structural significance can be difficult when sequence identity is weak or restricted to a short region. This is the key problem and in short what it says is, the shorter the length of the alignment, the higher the level of identity required for structural significance.

To solve this problem, it was needed to calibrate the length dependence of structural and sequence identity. Empirically, this can be done by deriving from a database of known structures a quantitative description of the relationship between sequence identity, structural identity and alignment length. The resulting definition of a length-dependent homology threshold can provide the basis for reliably deducing the structure of globular proteins likelihood down to the size of domains and fragments. Previously, the relation between the sequence identity and a three dimensional structure for the entire globular proteins was quantified (Chothia & Lesk, 1986).

With a clear idea of structural homology importance, we are now in a position to define a sequence identity cutoff above which structure homology can be inferred. For each alignment length, the cutoff is determined by the inspection of the Structure_Identity/Sequence_Identity scatter plot (Fig. 3.17) or histograms (Fig. 3.18), where above that sequence identity value (arrow in Fig. 3.18) the alignments are structurally homologous.

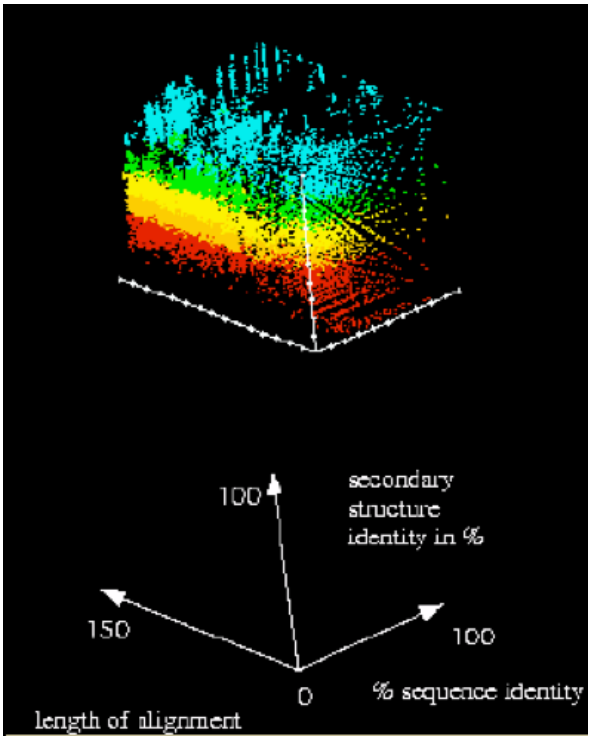


Figure 3.17: Color view. Calibration of the homology threshold is based on this 3D scatter plot of sequence identity (Y, range 0-100%), structure identity (Z, range 0-100%) and alignment length (X, range 0-150 residues) for pairwise protein sequence alignments (Schneider, 1994).

In Fig. 3.17 each point represents the alignment of two protein fragments, each one from a protein of known 3-D structure produced by FASTA (Pearson & Lipman, 1988). Red points are not identical pairs in structure (bad pairs), blue points are pairs identical in structure (good pairs), and intermediate colors for other values of structural agreement. The rectangular blue slice represents good pairs; they occur for almost all sequence identity and length values. The absence of (yellow and red) points in the top left and front shows that no pairs with sufficiently high sequence identity have low structure identity. Sequence identical oligopeptides (5-10 residues long) without identical local structure are red points at the front top right. Homologous protein pairs with about 150 residues of length are blue points at the back top left (Schneider, 1994).

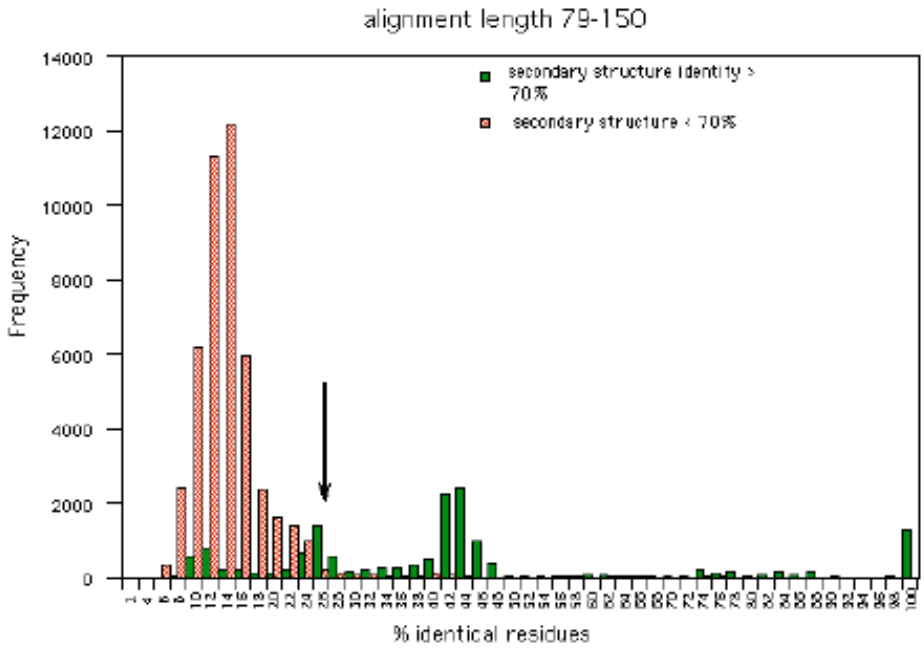


Figure 3.18: Structure identity implied by sequence identity, with dissimilar (red)/similar (green) dividing line at 70% identity of secondary structure symbols (A,R,N,...) (Schneider, 1994).

Detailed justification for the particular values of the homology threshold (arrow) is provided by histograms projections of Fig. 3.17 data: frequency of structurally identical/not identical alignments as a function of identical residues percentage in the alignment, for alignments of length 79-150 residues (Fig. 3.18). The threshold is perfect if all fragments pairs to the right of the threshold arrow are similar in structure (green bars) without intrusion by structurally dissimilar pairs (red bars). The strong mixture of red and green bars to the left of the arrow indicates that below the threshold one cannot use sequence identity percentage as indicator of structure identity. The particular choice of threshold represents an attempt to divide the range of sequence identity values into a 'do not know' region (left) and a 'sequence identity implies structure identity' region (right).

The resulting homology cutoff curve (Fig. 3.19) is a strongly varying function of alignment length up to 70-80 residues. For example, for alignment length 30, sequence identity has to be at least 43% to infer structural homology. For very long alignment lengths 25% sequence identity structural homology cannot be asserted nor excluded – the region of weaker sequence identity is a "don't know" region (mixture of squares and crosses under the curve in Fig. 3.19)

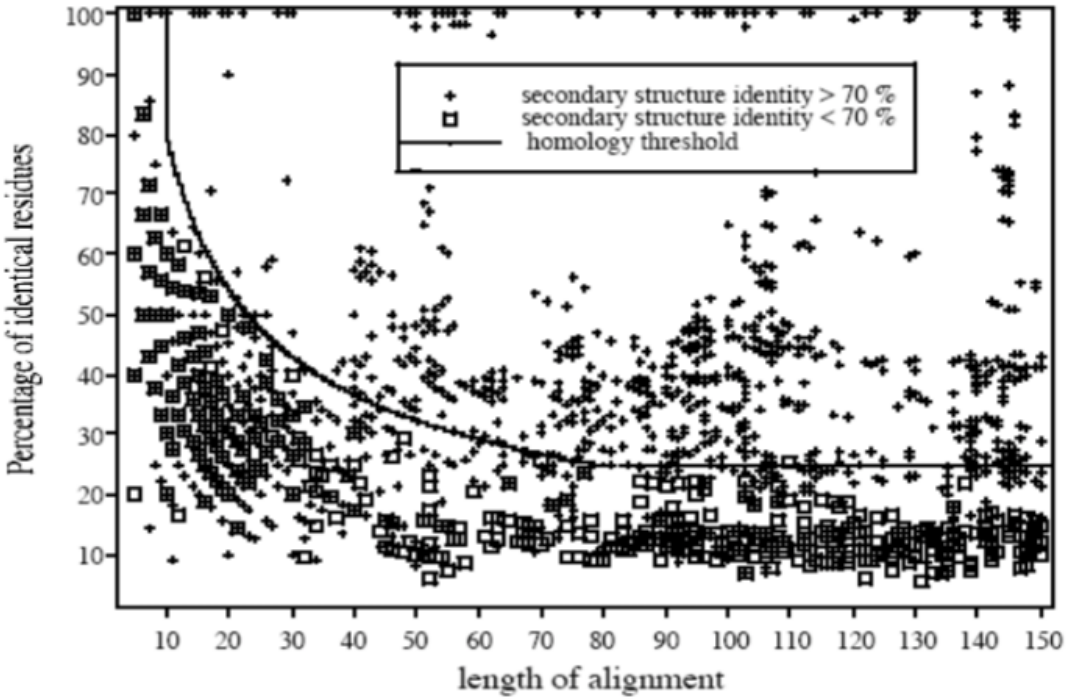


Figure 3.19: Homology threshold for structurally reliable alignments as a function of alignment length. Each data point represents an alignment between two fragments from points of known structure. The homology threshold (curved line) divides the graph into a region of safe structural homology (upper right) where essentially all fragment pairs are observed to have good structural identity (crosses, secondary structure identity above 70%) and a region of homology unknown or unlikely (lower left) where some fragment pairs are structurally similar (crosses) and some are not (squares, secondary structure identity below 70%) (Sander & Schneider, 1991).

In Homology threshold for structurally reliable alignments as a function of the alignment length, each data point represents an alignment between two fragments from proteins of known structure. The graph of Fig. 3.19 is a two dimensional projection of figure 3.17 onto the plane of sequence identity/alignment length, with structural identity collapsed to a one bit yes/no description (crosses/squares). The data points are a subset of the data in Figure 3.17. The homology threshold (curved line) divides the graph into a region of safe structural homology (upper right) where essentially all fragment pairs are observed to have good structural identity (crosses, secondary structure identity above 70 %) and a region of unknown or unlikely homology (lower left) where some fragment pairs are structurally similar (crosses) and some are not (squares, secondary structure identity below 70 %). The histogram of figure 3.18 corresponds to a vertical slice of this graph in the length range 79-150 residues, summing all available data points in that length range.

A sequence alignment between two proteins is considered to imply structure homology if the sequence identity is equal to or above the homology threshold *t* in a

sequence region of given length L. For example, an alignment with 30% sequence identity over a length of 60 residues implies homology while one with 30% sequence identity over a length of 40 residues does not. The threshold values *t(L)* were derived from an analysis of thousands of aligned fragment pairs from the PDB and can be represented by the formula

$$t(L) = \begin{cases} -, & L < 10 \\ 290.15 * L - 0.562 & 10 \leq L \leq 80 \\ 24.8, & L > 80 \end{cases} \quad \text{(Equation 3.1)}$$

where L is in the range 10-80 residues. For alignments shorter than 10 residues any value of sequence identity appears to be consistent with any degree of structure identity. Alignments longer than 80 residues have the asymptotic threshold of about 25 % identical residues.

Given a safe structural homology threshold, the database of homology derived protein structures production can occur, where for each protein of known structure in PDB it is performed a search in the sequence database for structurally significant alignments.

MaxHom Algorithm

The two main requirements for a MSA algorithm are 1) the calculation of an optimal alignment; 2) with the lowest possible computational expense. In theory, as well as in practice, making alignments with these two requirements that are mutually exclusive requires a compromise to find the optimal alignment, especially in the computation of a MSA of several hundreds or thousands of sequences, a fast alignment algorithm is required.

Next it will be described the MaxHom algorithm for multiple sequence comparison (Sander & Schneider, 1991; Schneider, 1994). It is a dynamic programming algorithm extended with position-dependent weights applied to identity matrix cell for the amino acids in question.

The commonly used substitution matrices essentially mirror the exchange of amino acids. These substitution matrices (Dayhoff et al., 1978) thus give an average probability in the "universe" of known protein sequences. The result of a MSA is

position dependent information of a protein family. This can be either in the form of a consensus sequence or as a measure of preservation. In the conventional methods, this information is usually calculated after the alignment. The method described here uses this protein family-specific information to build the MSA. The position-dependent conservation weights *cw(i)* are defined as follows:

$$cw(i) = \frac{\sum_{k,l=1}^{N} w(k,l).s(k_i,l_i)}{\sum_{k,l=1}^{N} w_{k,l}} \qquad \text{(Equation 3.2)}$$

$$w_{k,l} = \left(1 - \frac{1}{100}.\%id_{k,l}\right) \qquad \text{(Equation 3.3)}$$

Where:

cw(i): conservation weight at position *i*

N: number of alignments

k,l: index of sequences in multiple alignment.

$w_{k,l}$: weighing factor of a sequence pair to correct the unequal distribution in "sequence space".

$s(k_i,l_i)$: similarity value of the amino acid sequences of the pair k and l, at position *i*

In the calculation of the conservation weights are only those alignments that are above the derived homology curve, where an additional security area of 5% below is used. These conservation weights are updated after each pairwise alignment is performed. The formulation of the dynamic programming algorithm is as follows (Fig. 3.20):

$$F(i,j) = \max[L(i,j), F(i-1,j-1) + cw(i)*s(x_i,y_j), U(i,j), 0] \qquad \text{(Equation 3.4)}$$

When calculating a multiple sequence alignment, these weights at the beginning of the procedure all have the value of 1.0, after each pairwise alignment, they are recalculated and they are used for each subsequent pairwise comparison. In the case of the HSSP database, there is a sequence whose 3D structure is known. This sequence therefore has more information content and it is used as the reference sequence.

In MaxHom the similarity matrix used was the identity matrix.

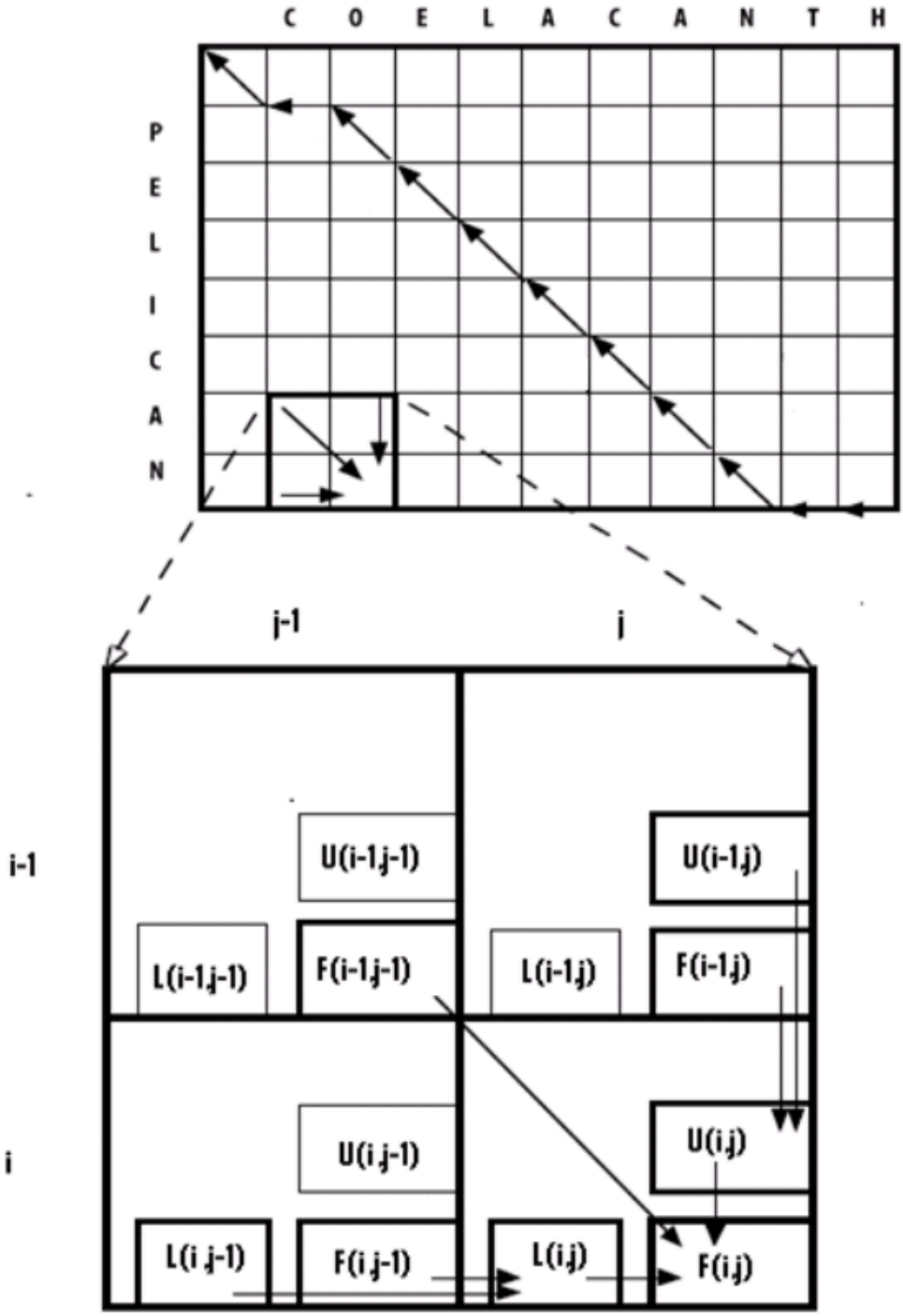


Figure 3.20: Schematic representation of the MaxHom dynamic programming algorithm for pairwise alignment. The upper part is a two sequence comparison using an alignment matrix, the bottom is a magnification and shows a detailed description of an iteration. In any step of the alignment calculation exists 6 values to be compared (a comparison is represented through an arrow). In the figure it is displayed 5 comparisons, the sixth is the comparison with zero. The best value of the alignments is set into cell F(i,j). The best values for either horizontal or vertical alignments are stored into temporary auxiliary fields L(i,j) and U(i,j) (Schneider, 1994).

Figure 3.21 is a schematic representation of the algorithm, where in a first run, sequences are sorted by similarity with the reference sequence, and weights are updated. To avoid bias in the pairwise alignments, after processing the whole list of homologues and fixing weights, a second run is made in which all pairwise alignments are compared with the same conservation weights. It is therefore a three-step algorithm:

- Pairwise alignments of potentially homologous sequences, using the conservation weights of the previous pairwise alignments;
- Fix of the conservation weights and normalize;
- Repeat all the pairwise alignments with the fixed conservation weights.

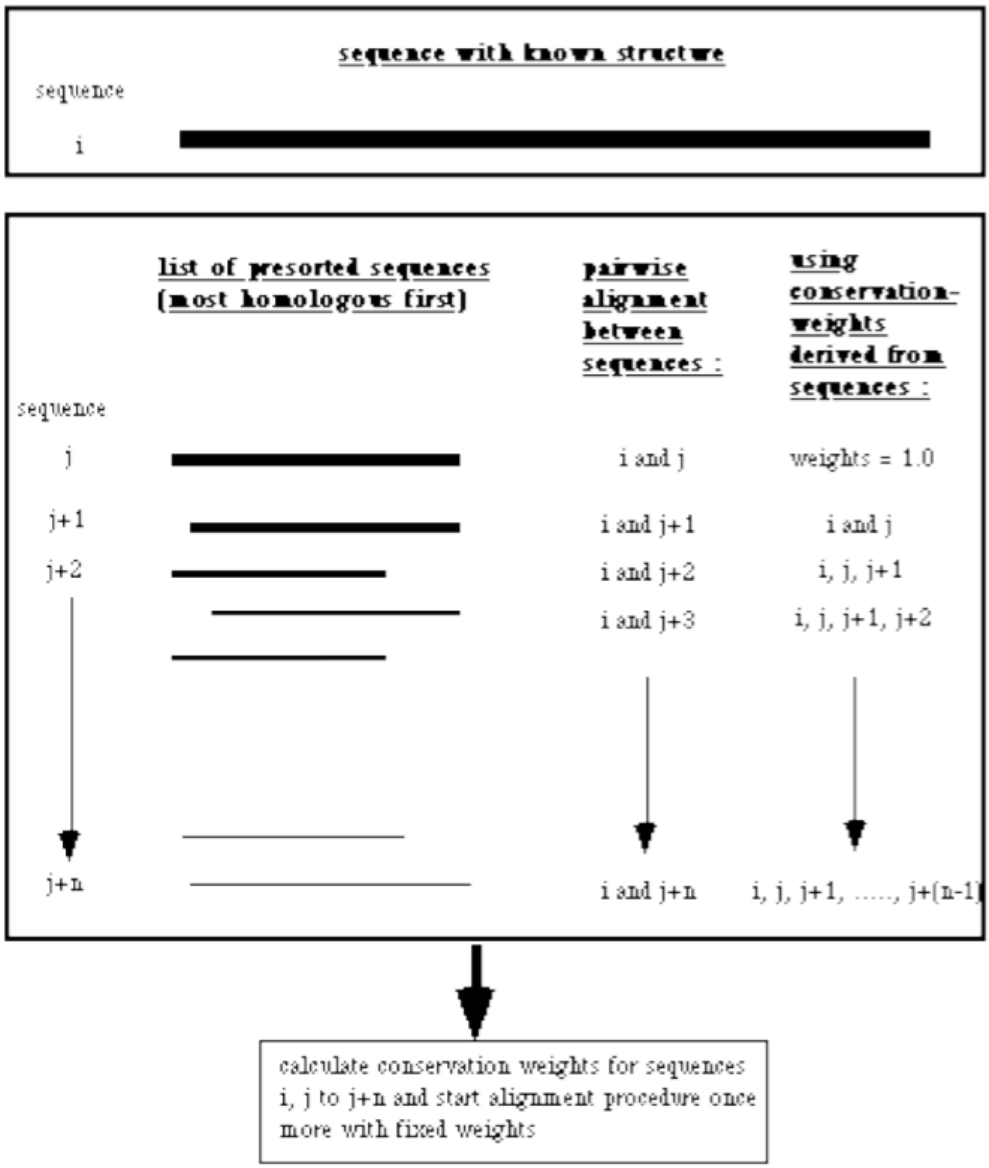


Figure 3.21: Schematic representation of the MaxHom extended dynamic programming algorithm with conservation weights (Schneider, 1994).

Figure 3.22 shows the development of position-dependent conservation weights during the alignment procedure. It was clear that after about 10 to 15 pairwise alignments, the conservation weights stabilization was achieved and positions with a high conservation showed high conservation values (weights).

An important side effect of this procedure was the relative insensitivity concerning the order of the sequence list at the beginning. If the list order of a protein family that contains around 15-20 sequences was reversed, no major changes were observed in the values of the final conservation weights (Schneider, 1994).

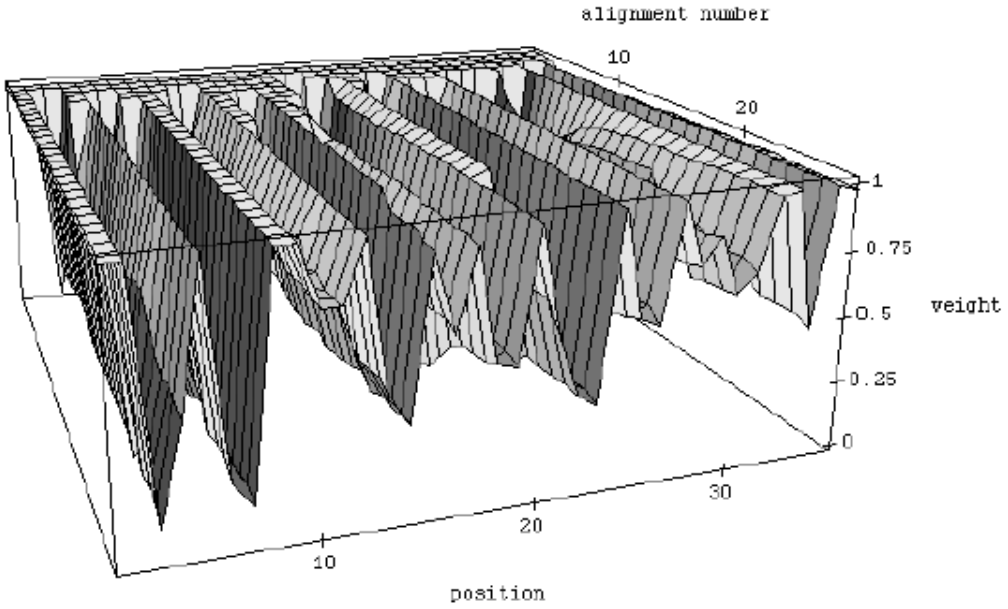


Figure 3.22: Evolution of conservation weights. The position-dependent conservation weights changes during the alignment procedure for the *Crambin* protein and its homologous sequences. At the beginning of the algorithm, each sequence position has a weight of 1.0. After a pairwise alignment, new weights are re-calculated and used for the next pairwise alignments. In this example, after approximately 10-15 pairwise alignments stable values for the weights were achieved. Position weights close to 1.0 indicate protein family positions that are conserved, while the ones with low values indicate positions that aren't conserved (Schneider, 1994).

A measure of sequence similarity, called *weighted similarity* was also introduced. This measure was calculated by using the exchange matrix for amino acids (in this case identity matrixes, but PAM250 (Dayhoff et al., 1978) or similar were also possible) multiplied by the obtained conservation weights (Fig. 3.22).

$$wsim = \frac{\sum_{p=i}^{j} cw(p).s(t_p,l_p)}{\sum_{p=i}^{j} cw(p).s(t_p,,t_p)} \quad \text{(Equation 3.5)}$$

Where:

wsim: *weighted similarity* (similar to a weighted alignment);
p: position in the sequence alignment;
i: start position alignment with respect to the test sequence;
j: end position alignment in relation to the test sequence;
cw(p): conservation weight at the position p;
t,l: sequences index. The *t* denotes the test sequence being *l* the comparison sequence;
$s(t_p,l_p)$: similarity value of the amino acid pair at the position p in the sequences *t* and *l*;
$s(t_p,t_p)$: similarity value of the amino acid pair at the position *p* in the test sequence *t* himself

For example, if we have two sequences with an identity of 25% on a length of 100 alignment positions with as test sequence, and a weighted identity of 35% for the first and 14% or lower for the second, it can be assumed that the first sequence belongs to the test sequence protein family, while the second sequence for sure is a non-related family sequence with the test sequence. The second sequence would have the same number of identical amino acid pairs, but these are majority in positions where the protein family shows high variability or low conservation.

Technically the algorithm MaxHom was a kind of a profile implementation. The values were calculated in recursive steps of the dynamic programming algorithm described with position-dependent values or weights. According with the authors this was a very flexible and adaptive algorithm, not only because of the information that was used (just sequence information), but where structural information could also be included. Widening the concept above it opened the possibility for a novel alignment algorithm for comparing two profiles (but was never implemented – Schneider personal communication). A schematic representation of the alignment program developed is shown in Figure 3.23.

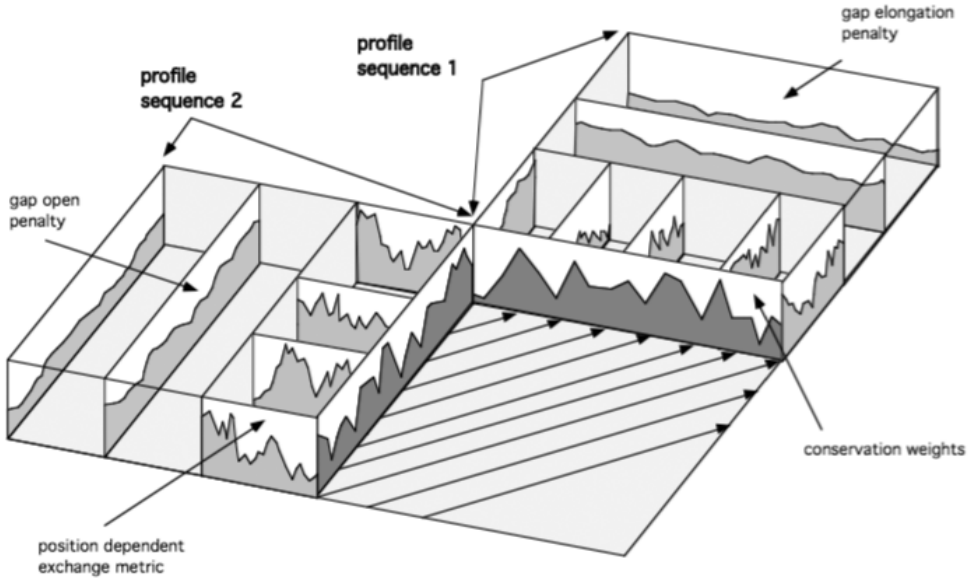


Figure 3.23: Schematic representation of the alignment algorithm. The simplest version of a program collapses sequence profile of pure sequence information, while other merely has a position independent exchange matrix (e.g. identity matrix). All other values are constant, such as conservation weights. In this form, simple pairwise sequence comparisons can be carried out as well as, profile comparison to a sequence or profile-profile comparisons (never implemented). The actual alignment as a square matrix that is shown in the foreground (Schneider, 1994).

Database Search with MaxHom

The database search using MaxHom can be described as follows: Each sequence alignment is the result of a pairwise comparison; The end result is a multiple sequence alignment; The search performed on protein sequence database can be synthesized in several steps.

(1) Rapid scan of the database using FASTA (Pearson & Lipman, 1988) with sufficiently low identity score cutoff yields a list containing all proteins potentially homologous to the reference PDB protein.

(2) A more refined proteins comparison to the ones present in the above list using MaxHom and retaining the 5 best distinctly different alignments for each pairwise comparison, yields an improved list of candidate alignments (Schneider, 1994).

(3) Only alignments with identity scores above the significance threshold (Eq. 3.1) are retained.

(4) All alignments are reported and registered relative to a single instance of the PDB reference protein.

Comment

In MaxHom/HSSP (Sander & Schneider, 1991; Schneider, 1994) it was performed an empirical determination of homology thresholds by studying thousands of sequence alignments within the PDB database. Each protein from a selected set of high and low-resolution protein structures is compared with all others from the set. The threshold for structural homology it was used to improve the selection of potential homologues in sequence database searches. Search methods like FASTA or BLAST sort the best hits on total similarity, careless of length. The suggested homology threshold curve presented in (Sander & Schneider, 1991; Schneider, 1994) can be used to order the database matches by the extent to which their score exceeds the threshold, in appropriate units, introducing more diversity and improving sensitivity in homologue selection.

HSSP2

In the HSSP2 (Rost, 1999) the following main questions were investigated: Do false positives increase more rapidly in the twilight zone (20-35% sequence identity)? Was the curve defined by (Sander & Schneider, 1991) still valid in 1999 with the increase size of the databases? Would using sequence similarity rather than identity improve accuracy (as speculated by (Sander & Schneider, 1991))?

The results of (Rost, 1999) verify, partially, earlier work based on a 1000-fold larger data set (Sander & Schneider, 1991). The main novel aspects were:

(i) A refinement of the threshold for identity (Fig. 3.24);

(ii) A definition of the threshold for similarity (Fig. 3.25);

Homology Thresholds

Protein databases are biased towards particular protein families. To reduce this bias, analyses are usually restricted to representative data sets (Hobohm et al., 1992). Rost chose the maximal set of sequence-unique proteins (792 in total) of known structure available in early 1997 (Holm & Sander, 1996). ‘Sequence unique’ was defined as ‘no pair in the set falls below the HSSP-curve (Eq. 3.1) by (Sander & Schneider, 1991; Schneider, 1994). As a rule-of-thumb, no pair had more than 25% pairwise sequence identity. Each of these proteins was aligned against the subset of PDB contained in the early 1997 release of the FSSP database of protein structure alignments (Holm & Sander, 1997). This subset amounted in total to about 5646 protein chains. Obviously,

the second step (792 versus 5646) reintroduced bias into the results. However, aligning the 792 sequence-unique pairs against themselves would not have yielded any result for most of the twilight zone analyzed there. Thus, 792 versus 5646 was the best compromise in reducing bias and monitoring the biased region. The resulting test set was the largest possible set of proteins for which structural information was available (and thus false and correct hits could be automatically distinguished).

The problems of the original HSSP-curve (Equation 3.1) considering the new and larger dataset were:

i) A threshold of 25% was not reasonable for an alignment length below 150-200 residues;

ii) Above, an alignment length of about 100 residues the derivative of the curve separating true and false positives should be lower than at lengths below 80.

Rost attempted to solve these problems by defining a new curve for separating true and false positives (Eq. 3.6).

$$t^l(L) = 480 * L \text{ -}0.32(1+\exp(\text{-}L/1000)) \text{ (Equation 3.6)}$$

where L gave the number of residues aligned between two proteins; t^l defined the cutoff percentage of identical residues over the L aligned residues. The constraints in visually selecting the final curve were:

i) to maintain the functional form defined by equation 3.1;

ii) to hit the 100% mark at alignments that are too short to reveal anything about structural similarity (11 residues);

iii) to saturate at levels around 20% sequence identity (reached for lengths of 300);

iv) to roughly reflect the observed gradient. Saturation for long alignments was realized by the functional form of the exponent (the term exp(–L/1000) resulted in an exponential decay). This 'saturation' constraint also afflicted the particular value of the factor (0.32 rather than 0.562 (Eq. 3.1) as suggested by the distribution of the data, Fig. 3.24).

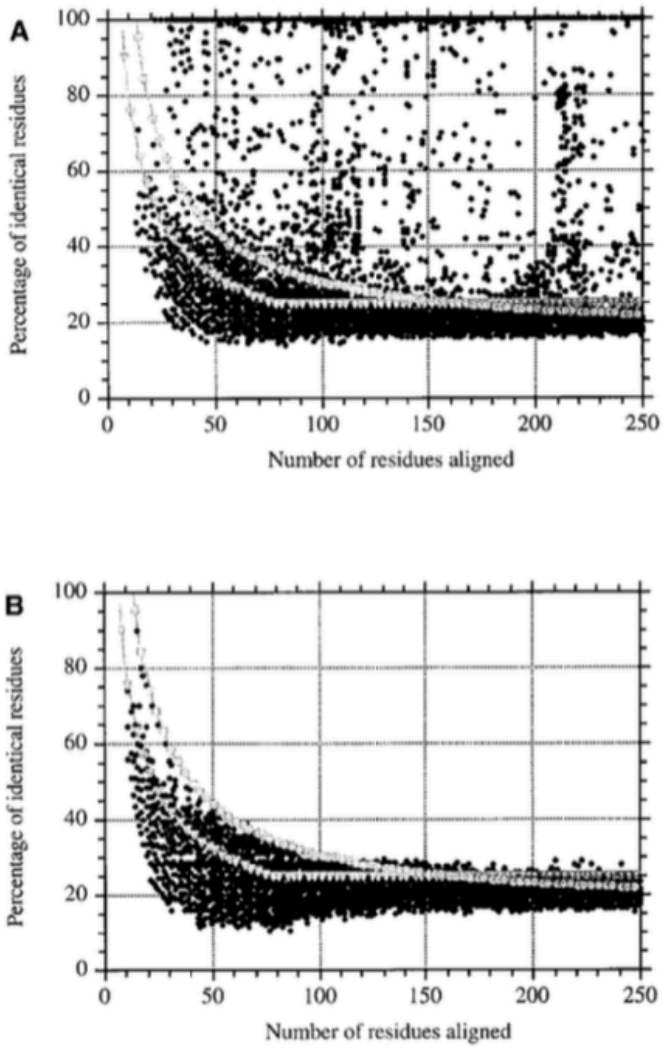


Figure 3.24: A) Pairwise sequence identity versus alignment length for true positives. The original HSSP-curve (Sander & Schneider, 1991) (triangles, Eq. 3.1) appeared to fit the true positives (homologues) better than the false positives B). In contrast, the curve proposed by Rost (circles, Eq. 3.6) was more conservative in excluding false positives (Rost, 1999).

Pairwise sequence identity was defined by the percentage of residues identical between two aligned sequences, and pairwise sequence similarity was defined by the percentage of residues similar between two sequences. Similarity scores depend on the particular metric used to capture physicochemical properties of amino acids. Consequently, levels of similarity are not directly comparable between different matrices. In (Rost, 1999) it was used the McLachlan metric (Gribskov et al., 1987; McLachlan, 1971).

The original HSSP-curve was derived for sequence identity, not for sequence similarity (Sander & Schneider, 1991). The functional dependence between similarity and length appeared comparable to the one between identity and length (Rost, 1999).

This prompted a similar definition for the separation between true and false positives based on similarity:

$$t^S(L) = 420 * L^{-0.335}(1+\exp(-L/2000)) \quad \text{(Equation 3.7)}$$

where L gave the number of residues aligned between two proteins; t^S defined the cutoff for the percentage of residue similarity over the L aligned residues (Figure 3.25).

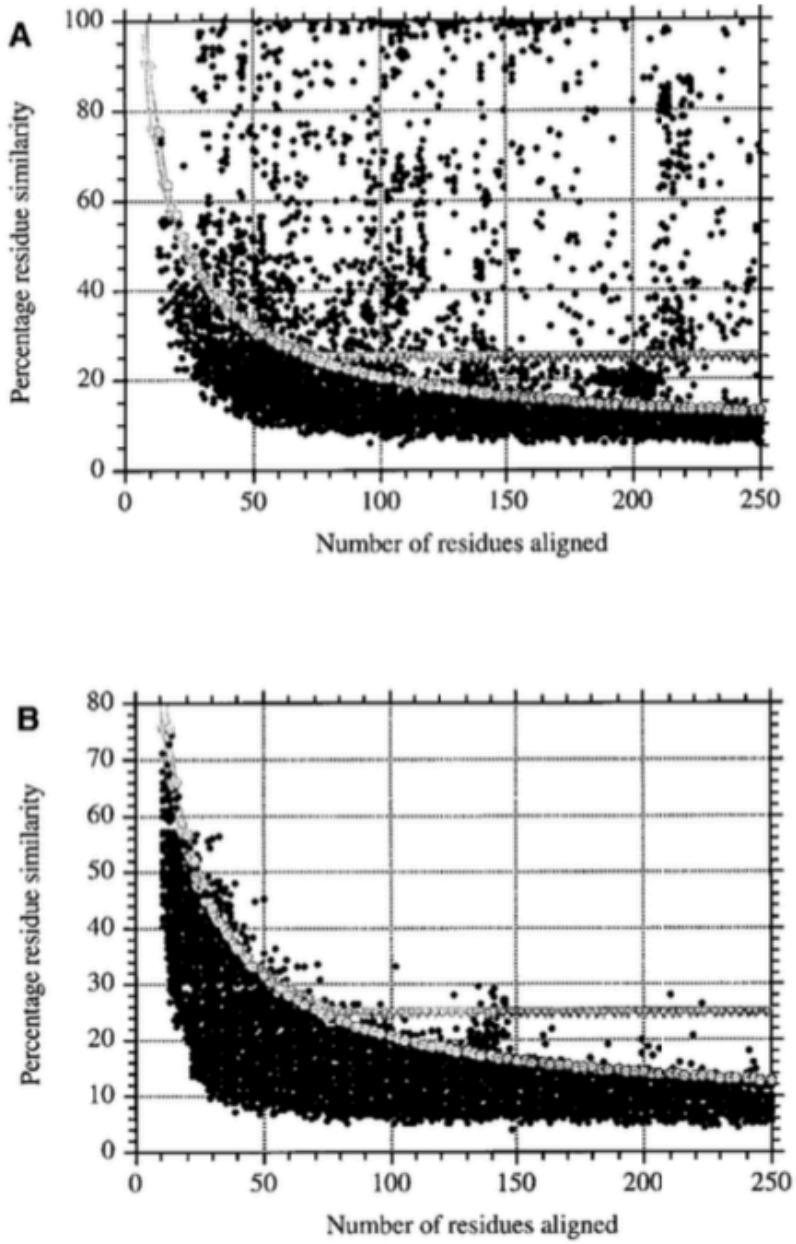


Figure 3.25: Pairwise sequence similarity versus alignment length. A) Correctly detected structural homologues; B) false positives. Open circles, original HSSP-curve (Sander & Schneider, 1991); triangles, Rost-curve (Rost, 1999).

The new curves for length-dependent cutoffs in sequence identity (Eq. 3.6) and similarity (Eq. 3.7) resulted in clearly lower false positive rates (higher accuracy) than the original HSSP curve. Furthermore, at any level of true positives detected, the number of false positives was smaller for the new curves (Eqs. 3.6 and 3.7) than for the original HSSP curve (Eq. 3.1).

MaxHom2 Algorithm

MaxHom2 algorithm is the previous MaxHom arranged with the threshold curve defined in equation 3.7.

Database Search Algorithm

MaxHom2 basically uses the multiple alignment method MaxHom (Sander & Schneider, 1991; Schneider, 1994) , together with the new homology threshold curves proposed by (Rost, 1999) (Figs. 3.24 and 3.25), and other minor improvements. A database search using it is outlined below:

1) For each structure in the PDB, an initial list of sequence hits is generated by running BLASTP (protein-protein BLAST) (Altschul, 1990; Altschul et al., 1997) against SWALL (RostLab internal sequences database). Initially, a relatively unrestrictive threshold identity is used, so that even relatively poor matches are retrieved.
2) All matching sequences are aligned with the structure using MaxHom alignment, based on the Smith-Waterman algorithm (Smith & Waterman, 1981), and modified as described by (Sander & Schneider, 1991). Similarity is measured using the McLachlan matrix (McLachlan, 1971).
3) The similarity-based homology threshold of (Rost, 1999) is used to determine the sequences that can safely be assumed (with 95% confidence) to have the same fold as the structure, within the aligned regions.
4) Sequences that fall inside the threshold are used to generate a profile based on the sequence family.
5) Steps 2 to 4 are repeated, this time using the generated profile as a reference.
6) A final list of aligned sequences is obtained; the final alignment incorporates information about all related sequences.

Comment

The combination of the MaxHom algorithm with equation 3.7 threshold curve (refinement) was denominated MaxHom2 and the resulting database denominated HSSP2.

These refinements were done with the intention of introducing more diversity and improving sensitivity in homologue selection, where an accurate and sensitive distinction between true and false positives is important for automatic database searches. The curves shown here (Eqs. 3.6 and 3.7) proved slightly more sensitive (higher coverage) and more accurate than the previously proposed curve (Sander & Schneider, 1991). The accuracy increased significantly by applying the 'more-similar-than-identical' rule. However, accuracy was gained at the expense of coverage. Which is more important? Clearly, the evolutionary information contained in multiple alignments was the single most important contribution to improving protein structure prediction in the 90's (Rost & Sander, 1996; Rost & O'Donoghue, 1997). Was the gain by increased diversity more important than the loss of accuracy when using alignments for structure prediction? The answer depends on the particular prediction goal. For example, secondary structure prediction diversity is more important than accuracy (cutoff at 25% versus that at 30%), whereas for the prediction of solvent accessibility the opposite is true (Schneider personal communication). Furthermore, as databases grow coverage may be less important than accuracy. Irrespective of individual preferences, the sharper the knife cutting between true and false positives, the better (Rost, 1999).

Figure 3.26 illustrates the balance of accuracy and coverage. Using the same data set as in Figure 3.25, the MaxHom2 method was compared with other methods, such as BLASTP and PSI-BLAST (Altschul et al., 1997). MaxHom2 outperformed PSI-BLAST for all high accuracy levels. For example, at 95% accuracy, MaxHom2 finds about 14% of all possible true positives. By comparison, PSI-BLAST finds only about 9%.

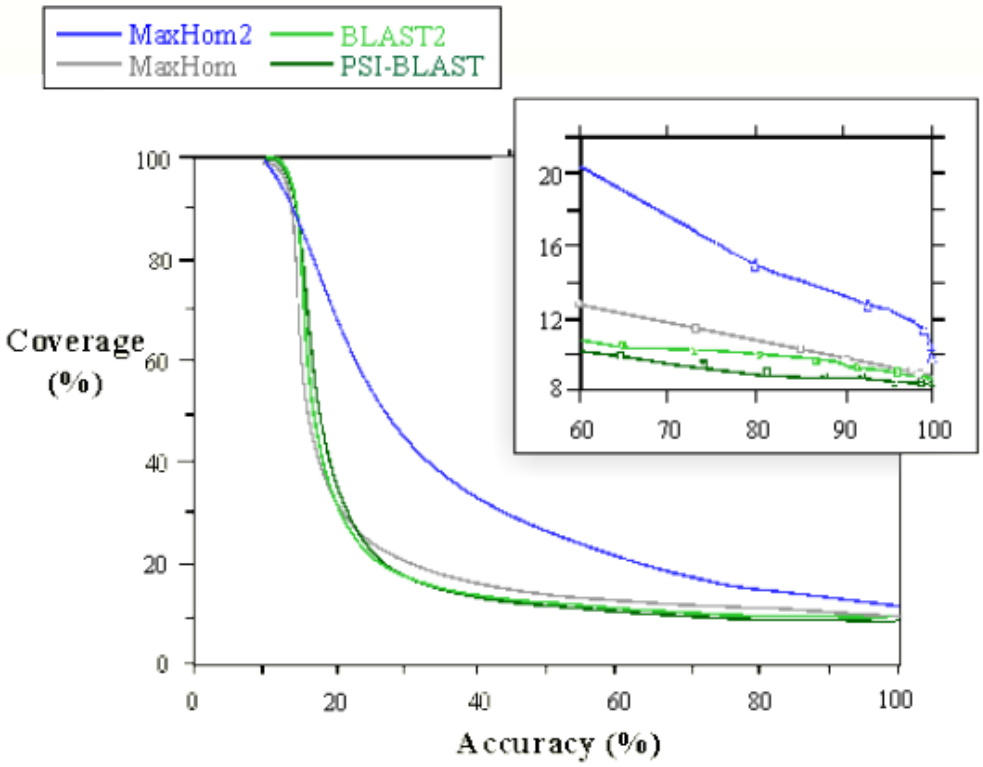


Figure 3.26: Balance of accuracy and coverage (adapted from (Rost, 1999))

PSSH

The Protein Sequence-to-Structure Homologies (PSSH) database derived from HSSP2 database (Rost, 1999), was an improved version of the HSSP database (Dodge et al., 1998). Whereas each HSSP entry lists all protein sequences related to a given 3D structure, PSSH is the 'inverse', with each entry listing all structures related to a given sequence. Two other tables were derived at that time: HSSPchain, in which each entry lists all sequences related to a given PDB chain, and HSSPalign, in which each entry gives details of one sequence aligned onto one PDB chain. That re-organization made it easier to navigate from sequence to structure, and mapping sequence features onto 3D structures (Fig. 3.27).

In September 2002, PSSH provided structural information for over 400.000 protein sequences, covering 48% of SWALL (RostLab internal sequences database) and 61% of Swiss-Prot sequences; HSSPchain provided sequence information for over 25000 PDB chains, and HSSPalign hold it over 14 million sequence-to-structure alignments. The databases were accessed via SRS 3D (O'Donoghue et al., 2004), an extension to the SRS (Etzold & Argos, 1993) system.

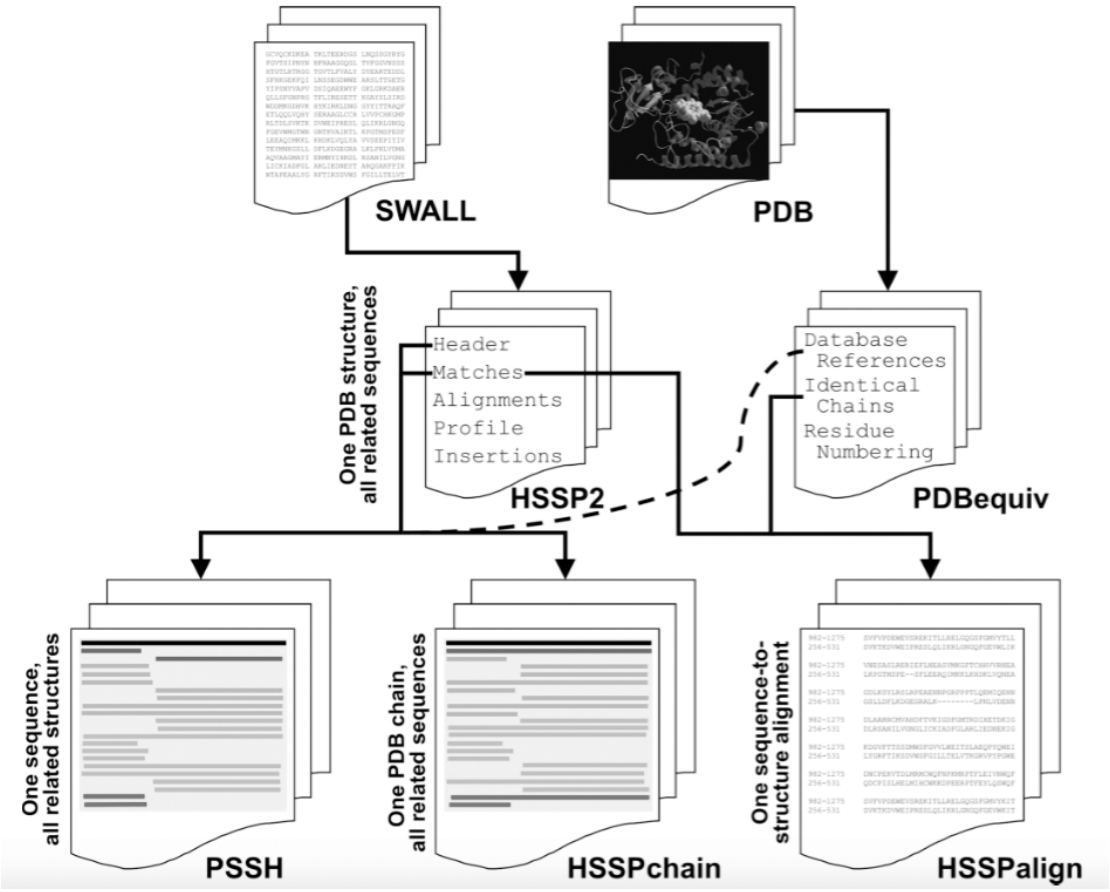


Figure 3.27: Schematic representation of the derivation of the PSSH-related databases. For each PDB entry, all related sequences in SWALL (RostLab group internal sequences database) are aligned using MaxHom2 and the alignment details are stored as one entry in HSSP2. For each PDB chain, it was stored additional information in an intermediate database, PDBequiv. Each HSSP2 entry was then processed to generate the remaining databases: all sequences that align onto one PDB chain were stored as one HSSPchain entry; each individual alignment in HSSP2 were stored as one entry in HSSPalign, with additional information extracted from PDBequiv. As each HSSP2 alignment was read, it was also appended to a separate file named by the sequence accession number, hence accumulating the PSSH database (Schafferhans et al., 2003).

Ten years later (in 2012) it was made a comparison between MaxHom2 (HSSP2/PSSH algorithm), PSI-BLAST (Altschul et al., 1997), PFAM (Punta et al., 2012) AND HHBlits (Remmert et al., 2011) to evaluate the performance of the former over the COPS dataset (Suhrer et al., 2009) and using TOPOFIT (Ilyin et al., 2004) as alignment gold-standard applying sensitivity (denotes the number of correct aligned columns compared to the structural alignment over the length of the sequence alignment) and specificity (denotes the number of correct aligned columns compared to the structural alignment over the length of the structural alignment). The conclusion was that sensitivity lied only at 5% for MaxHom2/HSSP2 (Fig. 3.28A) (Wellmann, 2012). Figure 3.28B reveal that in direct comparison MaxHom2/HSSP2 lost its higher

specicity to HHblits. These results lead to the conclusion that MaxHom2 filtered out the "hard cases" and only "easy cases" remained, which results in an overall high ratio of good alignments. Furthermore, the computational overhead of the employed MaxHom2 algorithm to compute alignments make the continuation of HSSP2/PSSH updates questionable (as a reminder: it was called HSSP, HSSP2 and PSSH databases, but in fact, they were databanks of text files).

This showed that MaxHom2 was clearly outdated and could not be relied on anymore.

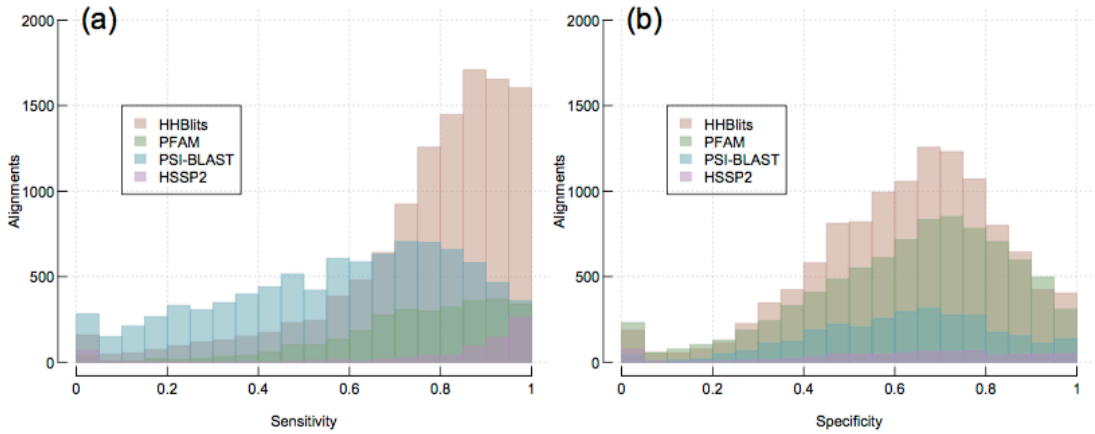


Figure 3.28: A) Sensitivity, B) Specicity of aligned PDB sequences for each alignment method. HHblits outperforms other methods, producing more alignments with higher quality based on all scores, followed next by Pfam, PSI-BLAST and HSSP2 (Wellmann, 2012).

A last and final attempt was made by me to validate the reliability of HSSP2/MaxHom2 in the twilight zone using DAPS (Database of Aligned Protein Structures) (Mallick et al., 2001). DAPS is a subset of FSSP (Holm & Sander, 1994) which contains alignments from those entries which have a low sequence identity percentage (25% or less). The DAPS was composed of 252 homologues pairwise alignments, but MaxHom only detected 47 homologues of the 252. PSI-BLAST detected 7 more homologues in a total of 54.

Figure 3.29 shows red points (PSI-BLAST pairwise alignments), green points (MaxHom alignments), red and green numbers – near to the previous points (that identify the DAPS alignment number), and finally color lines (green – MaxHom pairwise alignment better compared with PSI-BLAST alignment; red – MaxHom worst compared with PSI-BLAST alignment; yellow – Irrelevant according with Schneider,

because even MaxHom alignments were better, Schneider wanted an increase in length, which didn't occurred in that cases).

The DAPS benchmark results performed by me together with the previous ones (Wellmann, 2012) certified that MaxHom was not reliable.

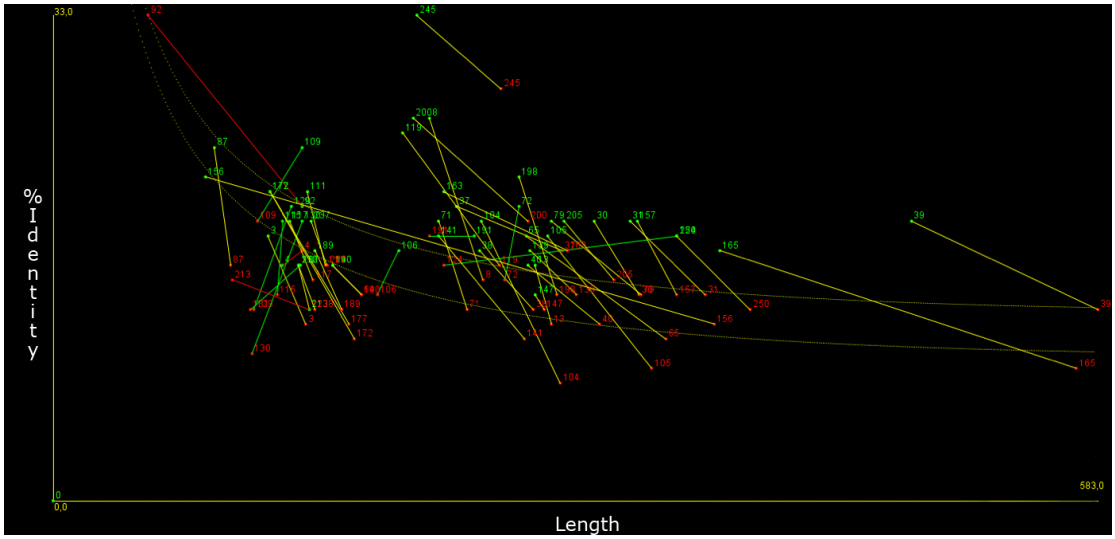


Figure 3.29: Displays 47 MaxHom alignments and 47 PSI-BLAST alignments, where: red points (PSI-BLAST pairwise alignments); green points (MaxHom alignments); red and green numbers (that identify the DAPS alignment number); and finally color lines (green – MaxHom pairwise alignment better compared with PSI-BLAST alignment; red – MaxHom worst compared with PSI-BLAST alignment; yellow – Irrelevant according with Schneider, because even that the MaxHom alignments were better, Schneider wanted a rise in Length, which did not occur in that cases).

PSSH/MaxHom2 kept during its existence sequences with significant similarity (homology) to proteins of know structure leading to a database of homology-derived of protein sequences with structural information several times larger than PDB (Fig. 3.30), reducing this way the gap between sequences (Swiss-Prot) and structural (PDB) databases.

Figure 3.30: Relation between, PDB (bottom), PSSH (middle) and UniProt (top) in 2010 (left) and 2013 (right). It is notorious the advantage of the existence of PSSH by adding structural information in sequences (1D) that had none (figure prepared by Christian Stolte & Sean O'Donoghue for poster D05 of VIZBI 2013).

PART III

METHODOLOGY

4. Aquaria

This chapter was partially published in:

"Aquaria: simplifying discovery and insight from protein structures", S. O'Donoghue, K. Sabir, M. Kalemanov, C. Stolte, B. Wellmann, V. Ho, M. Roos, **N. Perdigão**, F. Buske, J. Heinrich, B. Rost, A. Schafferhans, *Nature Methods*, Vol. 12, pp. 98–99,
doi:10.1038/nmeth.3258, 2015
http://www.nature.com/nmeth/journal/v12/n2/full/nmeth.3258.html

4.1. Summary

Viewing 3D models of a protein's molecular structure can give insight into function, especially when mapped with sequence features (e.g., domains, SNPs, or post-translational modifications). Homology-based 3D models are now available for many proteins, however it is often difficult to find the most relevant models and to map sequence features onto them. Thus we developed Aquaria, a new web resource that provides 46 million models (more than double the number previously available) derived from a systematic all-against-all comparison of Swiss-Prot (1D) and PDB (3D) sequences. Aquaria provides at least one model for 87% of Swiss-Prot proteins, with a median of 35 models per protein. Aquaria has been designed for ease of use, so that more life scientists can advance their research by taking advantage of the wealth of structural data now available. Aquaria is freely available at http://aquaria.ws.

4.2. Introduction

Many key insights into the molecular machinery of life have been derived from atomic-scale 3D structures (O'Donoghue et al., 2010) some key examples include the leap in understanding from the discovery of the DNA double-helix, as well as applications such as rational drug design and antibody engineering.

Structure determination methods have steadily improved, producing over 100,000 experimentally-derived structures in the Protein Data Bank (PDB) (Berman et al., 2000). This lags far behind the growth of protein sequence information, with less than 0.1% of UniProt (Consortium, 2014) proteins linked to a PDB structure. However, the understanding that evolution conserves structure more than sequence has led to large-scale computation of structural models (e.g., ModBase (Pieper et al., 2014) and SWISS-MODEL (Kiefer et al., 2009).

Currently, over 21 million models are consolidated in the Protein Model Portal (PMP) (Haas et al., 2013), providing structural information for over 5 million proteins, covering 80% of all manually annotated (Swiss-Prot) proteins and ~9% of all known protein sequences (UniProt) (Consortium, 2014).

Thus, structural modeling now scales with genomic sequencing, providing tremendous amounts of information that can give detailed functional insights, far beyond what is accessible from sequence alone. Currently, however, many biologists fail to take full advantage of this valuable information; two key reasons for this include:

(1) it is not always easy to find the most appropriate model amidst the increasing volume of other related information; and (2) 3D structures are intrinsically complex, and with existing tools a significant investment in time is needed to navigate through these complex data and derive insight.

Related to point (1), we believe there is one important and useful view of structural data that current resources do not provide: a view giving a concise visual summary of all related structural information for any specified protein. Such a view was previously available in SRS 3D (O'Donoghue et al., 2004), however that service is no longer available. PMP currently provides part of such a view, but shows only a small number of similar structures.

Related to point (2), we believe the key issue is that most existing resources disseminating 3D structures have been created primarily by and for the structural biology community. However, since models are now available for so many proteins, these models are of interest and relevance to a much broader group of scientists, many of whom are unfamiliar with the rather complex data (atomic structures) and required tools (molecular graphics, etc.).

To address these issues, we have developed Aquaria, a web resource intended to augment the ability of biochemists and molecular biologists to derive insight into protein function from structural models. Aquaria has been designed to provide a highly visual and intuitive user experience. In addition to providing unprecedented ease of access to all available structural information for any specified protein, Aquaria makes it easy to map sequence features – such as domains, SNPs, or post-translational modifications– onto 3D structures. Such feature mapping can be effective in providing functional insight (O'Donoghue et al., 2010).

In contrast to most molecular graphics tools (for example, Astex (Hartshorn, 2002) or Chimera (Pettersen et al., 2004)), the user interface of Aquaria is organized primarily by protein sequence, not structure (Fig. 4.1). A user starts by specifying a protein of interest by name and organism, by identifier or by Uniform Resource Locator (URL) (for example, http://aquaria.ws/P04637); Aquaria then generates a concise visual summary of all related PDB structures (Figs. 4.1i and 4.1ii), using a pre-calculated all-against-all comparison of Swiss-Prot and PDB sequences.

The related structures are grouped first by alignment to the specified sequence and second by oligomeric state (Fig. 4.1iii). Structures are then ranked - in both groupings - by sequence similarity to the specified protein. Users can quickly review all known

structural information for a protein and find the structures most relevant to them (Fig. 4.1iv).

Aquaria also allows mapping of UniProt and InterPro (Hunter et al., 2012) sequence features (for example, domains, single-nucleotide polymorphisms or posttranslational modifications) onto 3D structures: a simple yet effective way to gain insight into molecular function (Fig. 4.1v).

Initially, 3D structures are colored to highlight amino acid differences from the specified protein sequence, with bright, saturated colors indicating identical residues and with slightly dark and very dark coloring indicating conserved and nonconserved substitutions, respectively (Fig. 4.1i).

Aquaria is designed for biologists; its user interface creates clear and useful default views that show only the most relevant structural information tightly integrated with sequence, features and text that provide biological context. Aquaria uses a minimal set of mouse-based controls that are intuitive yet powerful (O'Donoghue et al., 2004). For example, its 'Autofocus' feature allows exploration of large complexes by focusing on one molecule at a time. Aquaria can also be controlled via hand gestures using the Leap Motion (Sabir et al., 2013). Currently, Aquaria contains 46 million pre-calculated sequence-to-structure alignments, resulting in at least one matching structure for 87% of Swiss-Prot proteins and a median of 35 structures per protein; this provides a depth of sequence-to-structure information currently not available from other resources.

4.3. Data

Aquaria uses relational databases to display its information in a browser, but several steps were taken prior to achieving the current state. The databanks that I worked directly with were Swiss-Prot and PDB protein knowledge bases. Their data repositories consisted and still consist in a set of flat files (i.e. text files containing records with a standardized nomenclature) where within a file record, one can organize the data using different types of fields. This organization emulates some of a relational database's behaviors (third generation database technology) (Chowdhury, 2004) but it isn't one and therefore generates many problems (inefficient access, security and administration problems, no concurrency access and no logical data model).

My work was centered in converting Swiss-Prot and PDB flat files databanks into relational databases (fourth generation database technology) (Chowdhury, 2004). This conversion was not easy and several types of exceptions occurred in the parsing process

taking months until a complete and reliable parse was obtained. Other problems such as memory problems due to the number of files and its dimension also occurred. Finally optimization techniques were applied for table size reduction, and for fast data access creation of table indexes. To ensure reliability several validation and 'smoke tests' were performed.

In sum, the conversion process from flat files to relational databases and its corresponding tables (see Database Section) was complex. However, it was a required process to obtain all the information concerning each protein subject to a query, in a fast and fully detailed manner, especially considering that the goal was for this bulk information to be consulted as a whole through a web service.

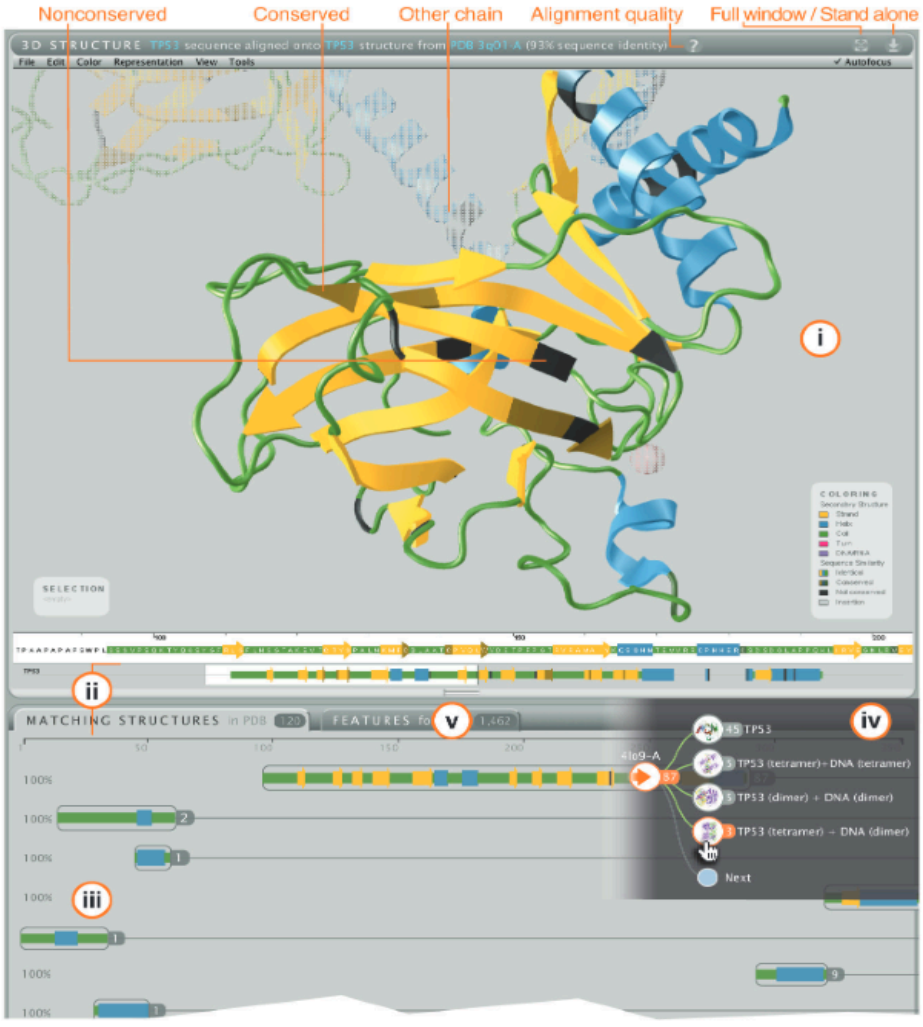


Figure 4.1: Aquaria page for human tumor suppressor protein p53. i) Initially, the PDB structure estimated to be most relevant is shown. Dark and very dark residues indicate conserved and nonconserved substitutions, respectively, between the structure and the wild-type p53 sequence. ii) Aquaria also shows all related PDB structures, grouped by region of match. iii) Clicking on a group loads the top-ranked structure; iv) clicking on a group number shows a tree view of structures organized by oligomeric state. InterPro and UniProt features v) can also be mapped onto structure (O'Donoghue et al., 2015).

4.4. Methods

At its core, Aquaria relies on aligning sequences of unknown structure (Swiss-Prot) onto sequences with known structure (PDB). The previous SRS 3D (O'Donoghue et al., 2004) used PSSH (Schafferhans et al, 2003), a database of protein sequence-to-structure homologies generated with PSI-BLAST (Altschul et al., 1997) and other alignment tools (Schafferhans, et al., 2003). PSSH new version - PSSH2 (O'Donoghue et al., 2015) – is based on HHblits (Remmert et al., 2011), an alignment method employing iterative comparisons of Hidden Markov Models (HMMs). HHblits is the key method used in HHpred (Remmert et al., 2011), a fully automated server for template-based structure prediction that was ranked best out of 79 similar servers at the CASP9 competition in 2009 (http://bit.ly/hhblits-casp9). At the 2011 CASP competition, HHpred slipped to 7th rank (http://bit.ly/hhblits-casp10), however all higher ranked servers were slower by a factor of 370 or more. Thus, it was selected HHblits as it combines both speed and reliable detection of structural templates.

4.4.1. Sequence to Structure Alignment

To ensure the highest possible final alignment quality for matches in Aquaria using HHblits (Remmert et al., 2011), it was first calculated HMM profiles for each unique PDB sequence (PDB_full) and also for each unique Swiss-Prot sequence (Fig. 4.2). For both these steps, it was used UniProt20, a database of non-redundant sequence profiles distributed with HHblits. UniProt20 is based on an all-against-all UniProt sequence comparison that was then clustered using kClust from HH-suite (Hauser et al., 2013), resulting in 4.8 million sequence clusters in which the highest pairwise sequence identity between clusters is 20%.

All sequences in each cluster were then incorporated into an HMM, thus creating one entry in the UniProt20 database. To create PDB_full, it was first ran HHblits using all unique protein sequences in PDB (derived from the PDB SEQRES records) searching against UniProt20, producing Multiple Sequence Alignments (MSA). From these MSA's, the PDB_full database files were then created using HH-suite. The PDB_full database (March 2014) contains 57,657 protein sequence profiles. It was then used the same process to create a database of HMMs for every unique Swiss-Prot sequence (540,000). Finally, it was generated PSSH2 using HHblits to find similarities between HMMs from Swiss-Prot and HMMs from PDB. This demanding calculation

required a computer cluster with sufficient RAM to hold the PDB_full HMM database. To reduce the required time, it was restricted the calculation to search only HMMs from Swiss-Prot sequences. Selecting only matches with $E \leq 10^{-10}$ resulted in a total of 46 million sequence-to-structure alignments in PSSH2; this provides at least one matching structure for 87% of Swiss-Prot entries, with a median of 35 structures per protein. Of these, 9.3 million are high confidence sequence-to-structure alignments, covering 28% of Swiss-Prot, with a median of 16 per protein. More details about PSSH2 setup (that was led by the RostLab in Munich) can be obtained in the Aquaria paper (O'Donoghue et al., 2015).

By comparison, one of the most similar existing resources, the Protein Model Portal (PMP) (Haas et al., 2013), contains a total of 22 million protein structure models for 5 million distinct UniProt sequences, an average of 4.4 models per protein, and covering 80% of Swiss-Prot. Thus, Aquaria provides more structures per sequences but is focused on a smaller set of proteins (Swiss-Prot only). However, a more fundamental difference is that Aquaria is based only on sequence-to-structure alignments, while the structural models in PMP are calculated using much more laborious comparative modeling methods such as SwissModel (Kiefer et al., 2009) and MODELLER (Eswar et al., 2007).

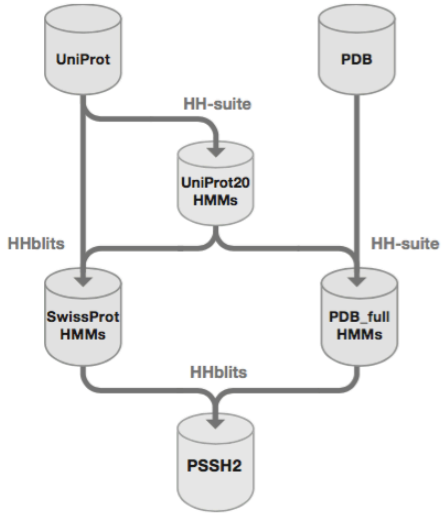


Figure 4.2: Workflow for generating PSSH2. It was used UniProt20 from HH-suite, a database of non-redundant UniProt sequence clusters in which the highest pairwise sequence identity between clusters was 20%. It was then used HH-suite to calculate Hidden Markov Model (HMM) profiles for each unique PDB sequence (PDB_full) and for each unique Uniprot (Swiss-Prot) sequence. Finally, it was generated PSSH2 using HHblits to find similarities between HMMs from PDB and HMMs from UniProt (Swiss-Prot) sequences. To save time, it was restricted this final step to only search against Uniprot (Swiss-Prot) sequences. In the future, it is planned to add to Aquaria the facility to extend PSSH2 to include any sequence on-demand (O'Donoghue et al., 2015).

4.4.2. Database

PSSH2 table: After the PDB_full generation, we searched each Uniprot (Swiss-Prot) sequence against the PDB_full database. From the search output, we created a MySQL table storing the PSSH2 (Figs. 4.2 and 4.3 left). Each PSSH2 entry contains the following: MD5 sum for the UniProt (Swiss-Prot) sequence; MD5 sum for the PDB sequence; E-value, sequence identity; and the alignment with a minimal format identifying gapless blocks.

UniProt (Swiss-Prot) related tables: I created four additional MySQL tables to store information related to UniProt (Swiss-Prot) sequences (Fig. 4.3 top). The main table ('protein_sequence') has one entry per UniProt sequence. While processing UniProt (Swiss-Prot) to build this table, we checked the MD5 sum for each sequence against a hash containing all UniProt (Swiss-Prot) MD5 sums in PSSH2. If the UniProt (Swiss-Prot) sequence had a match in PSSH2 all synonyms for the protein name (including identifiers) were added to a second table ('protein_synonyms'), all synonyms for the organism were added to a third table ('organism_synonyms'), and Latin organism names were added to a fourth table ('organism_names'). These synonym tables are then used to provide an autocomplete function for protein or organism names in the 'SPECIFY A PROTEIN' input fields (Fig. 4.4a). As a result, when a user looks up a protein by synonym, only those with matching structures in PSSH2 will be found.

PDB-related tables: I created three additional MySQL tables to store information related to PDB structures (Fig. 4.3 bottom). The main table ('PDB') has one entry per PDB entry, while another table ('PDB_chain') contains information about each PDB chain. The third table stores information about related PubMed articles. During processing of each PDB file, protein sequences were extracted from the ATOM records and aligned onto the corresponding SEQRES records – this alignment information is stored in PDB_chain, and is used on the fly when constructing the views presented in the user interface to show where the UniProt (Swiss-Prot) and PDB sequences differ (Fig. 4.4c), to map UniProt (Swiss-Prot) sequence features to PDB structures, and also to map PDB secondary structure onto UniProt (Swiss-Prot) sequences (Fig. 4.4e). Where present, we also read the first 'biounit' file (judged by the PDB to have the biologically-relevant assembly, and indicate by the file extension 'pdb1') – this information is used on the user interface to indicate oligomeric state when displaying information on molecular configuration subgroups (Fig. 4.4i). While reading each PDB

and biounit file, we calculate a transformation matrix for each chain that aligns the first three principle components of the Cα coordinates along the x-, y-, and z-axes, respectively. This matrix is then used automatically for the initial view in Aquaria, thus reducing occlusion by presenting each chain in a way that minimizes its depth along the z-axis

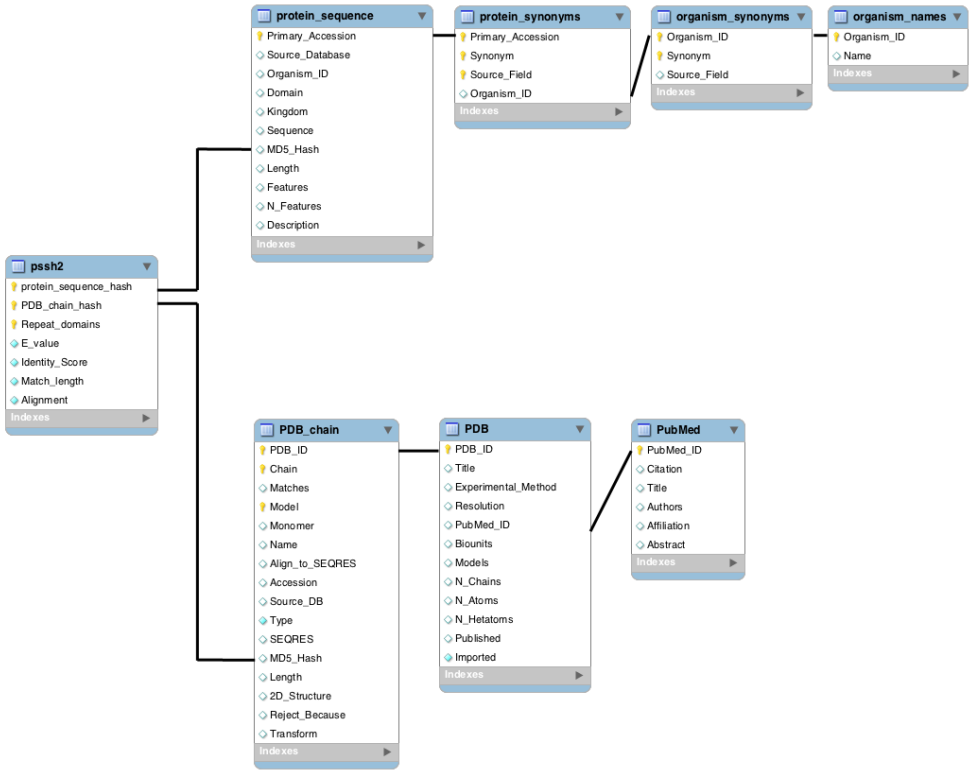


Figure 4.3: Aquaria database schema. A) PSSH2 table: *Repeat_domains* counts how often the same PDB and protein sequence hash have occurred with different alignments (indicating one PDB matching to multiple regions of a Swiss-Prot sequence). B) PDB-related tables. The *Matches* field indicates if the DBREF record of the current chain is identical to that of a previous chain from the same PDB entry. *Model* is only relevant for biounit files, where the same chain can occur in multiple model; we create a separate PDB_chain entry for each such case, and distinguish them by setting this field to be equal to the model number in which this copy of the chain occurs. The *Monomer* field contains a count of the total number of matching chains in the biounit file (or PDB entry, for NMR and cyroEM structures) – this field is used to display the oligomeric state for each structure (Fig. 4.4i). The *Transform* field holds a transformation matrix that is applied when this chain is viewed, with the result that the chain is shown centered on the screen and the first two principle components (calculated from C^{α} positions) coincide with the *x*- and *y*-axes, respectively (Fig. 4.4d). The *Type* field distinguishes protein, DNA, and RNA chains (currently, only proteins are used in Aquaria). The *Reject* field records if and why a chain was rejected (e.g., because there was no SEQRES record for that chain in the PDB file). C) UniProt-related tables: *Source_Field* distinguishes Swiss-Prot and TrEMBL sequences.

4.4.3. Web Interface

Aquaria has a novel backend created using free and open source components. The Aquaria web server was implemented using Node.js (http://nodejs.org/) to manage client-server and server-database communication, and using Express (http://expressjs.com/) for serving static files and producing dynamic files through template rendering.

For the 2D Graphics, matching structure groups (Fig. 4.4h) and features (Fig. 4.4l) are rendered by the browser as scalable vector graphics (SVG) via a customized JavaScript sequence object with single-residue resolution, created using D3.js (Bostock et al., 2011). For coloring of matching structure groups, amino acid substitutions with a McLachlan (McLachlan, 1971) score ≥ 0 were considered conserved, those with < 0 were non-conserved.

The tree view that appears upon clicking on a matching structure group (Fig. 4.4d and 4.4j) is constructed using a customized version of the standard D3.js tree layout library (d3.layout.tree).

For the 3D molecular graphics, it was used the SRS 3D (O'Donoghue et al., 2004) Viewer, a free and open source molecular graphics system that was designed to be intuitive and easy to learn. It uses hardware-accelerated rendering through OpenGL, allowing for anti-aliasing, dynamic lighting, rotation, and translation calculations without extra load on the CPU or memory. In adapting the SRS 3D Viewer for Aquaria, it was made a number of improvements. It is now based on the community maintained version of Java3D (1.6 pre 7), which uses pure Java calls to Java OpenGL (JOGL 2.0 r11) bindings.

4.5. Results

An Aquaria user first needs to specify a protein of interest, called hereafter the 'specified protein'. This is done either by typing in a protein name and species, or by composing a URL with a UniProt primary accession (e.g., http://aquaria.ws/P04637). Aquaria then automatically displays the following: synonyms for the specified protein (Fig. 4.4a); a summary of its function (Fig. 4.4b); two graphical representations of its sequence (Fig. 4.4c); a concise graphical summary of all matching structures in PDB (Fig. 4.4h); the PDB structure and chain estimated to be most relevant (Fig. 4.4d); finally, the PubMed abstract describing that PDB structure (Fig. 4.4f).

The structures shown in Aquaria (Fig. 4.4d) are commonly referred to as template models, meaning that the specified protein's sequence (Fig. 4.4c) has simply been mapped onto the unmodified 3D coordinates from the selected PDB file.

Aquaria's template models could be used as input to calculate more detailed homology models (e.g., using Modeller (Eswar et al., 2007)); however this can require a considerable time investment, thus we believe that, for most users who are not expert in structure, the template models provided by Aquaria are sufficient and easier to interpret.

Each Aquaria model (Fig. 4.4d) is initially colored to highlight where the specified protein's sequence differs to that of the original PDB structure, with slightly dark and very dark coloring indicating conserved and non-conserved amino acid substitutions, respectively. Any regions of the currently focused PDB chain that could not be aligned to the specified protein (e.g., insertions) are indicated in white coloring, while all other chains in the PDB file are initially semi-transparent. The quality of each model is primarily communicated by this color scheme, which results in high quality models having solid coloring (Fig. 4.4d), while low quality models have dull and mottled coloring. Further details about model quality can be accessed from the 3D view title-bar (Fig. 4.4d).

By default, only the specified protein's sequence is shown (Fig. 4.4c); however, the sequence of the corresponding PDB chain can also be shown by selecting 'Show PDB Sequence' from the View menu. The name and organism of the protein used to derive this PDB chain is shown in the 'ABOUT PDB' section (Fig. 4.4g). By clicking on the name of this protein, it will become the new specified protein, thus causing most views on the webpage to be updated, including the protein sequence (Fig. 4.4c), synonyms (Fig. 4.4a), function (Fig. 4.4b), matching structures (Fig. 4.4h), features (Fig. 4.4k), and chain information (Fig. 4.4g).

Often, the PDB file shown in the 3D view (Fig. 4.4d) contains additional chains, initially shown as semi-transparent; clicking on such a chain autofocuses on that chain, i.e., the molecule moves such that the chain is centered and becomes the center of rotation, while the chain is now shown with solid coloring with all other chains semi-transparent (autofocus can be disabled from the menu bar, or temporarily by double-clicking on the background). If this new chain corresponds to a different protein than the previous chain, the specified protein changes (to the UniProt protein specified for this chain in the PDB file), thus updating most views on the page.

This greatly facilitates the exploration of protein binding partners - the user can always navigate back to a previously specified protein by clicking on the previously focused chain.

The matching structures section (Fig. 4.4h) is designed to provide a graphical summary of all PDB structures with significant sequence similarity to the specified protein; these structures are organized into groups based on regions of match to the specified protein's amino acid sequence. These groups often correspond to protein domains, although not always, since many PDB structures contain multiple domains. This grouping is a key advantage of Aquaria compared to other similar resources; it rapidly communicates an overview of structural matches while providing easy access to any individual match.

Each group is colored to show the match in alignment between the specified protein and the top-ranked structure in the group (Fig. 4.4h), using the same colors initially shown on the 3D structure (Fig. 4.4d). Similar information is also communicated via alignment identity scores (Fig. 4.4h, left). As a result, a user can quickly gauge how closely the structures match to the specified protein.

Clicking on the colored portion of any group loads the top-ranked structure into the 3D view (Fig. 4.4d). To access all other structures within a group, the user can click on the gray-background number shown to the right of each group (Fig. 4.4i); member structures are then shown, further organized into a ranked list of sub-groups, based on oligomeric state, i.e., the names and number of copies of macromolecules present in each PDB entry (using the biological assembly judged to be most likely by the PDB). Finally, by clicking on the gray-background number attached to the right of each sub-group, the user can access a ranked list of individual PDB structures within that sub-group (Fig. 4.4j). In each case above, ranking of structures is based firstly on percentage sequence identity to the specified protein, then by the total number of identical residues, then by crystallographic resolution, with NMR and cryoEM structures last.

For each specified protein, clicking on the Features tab (Fig. 4.4k) reveals a set of sequence features (Fig. 4.4l) retrieved from InterPro (Hunter et al., 2012) and UniProt (Consortium, 2014).

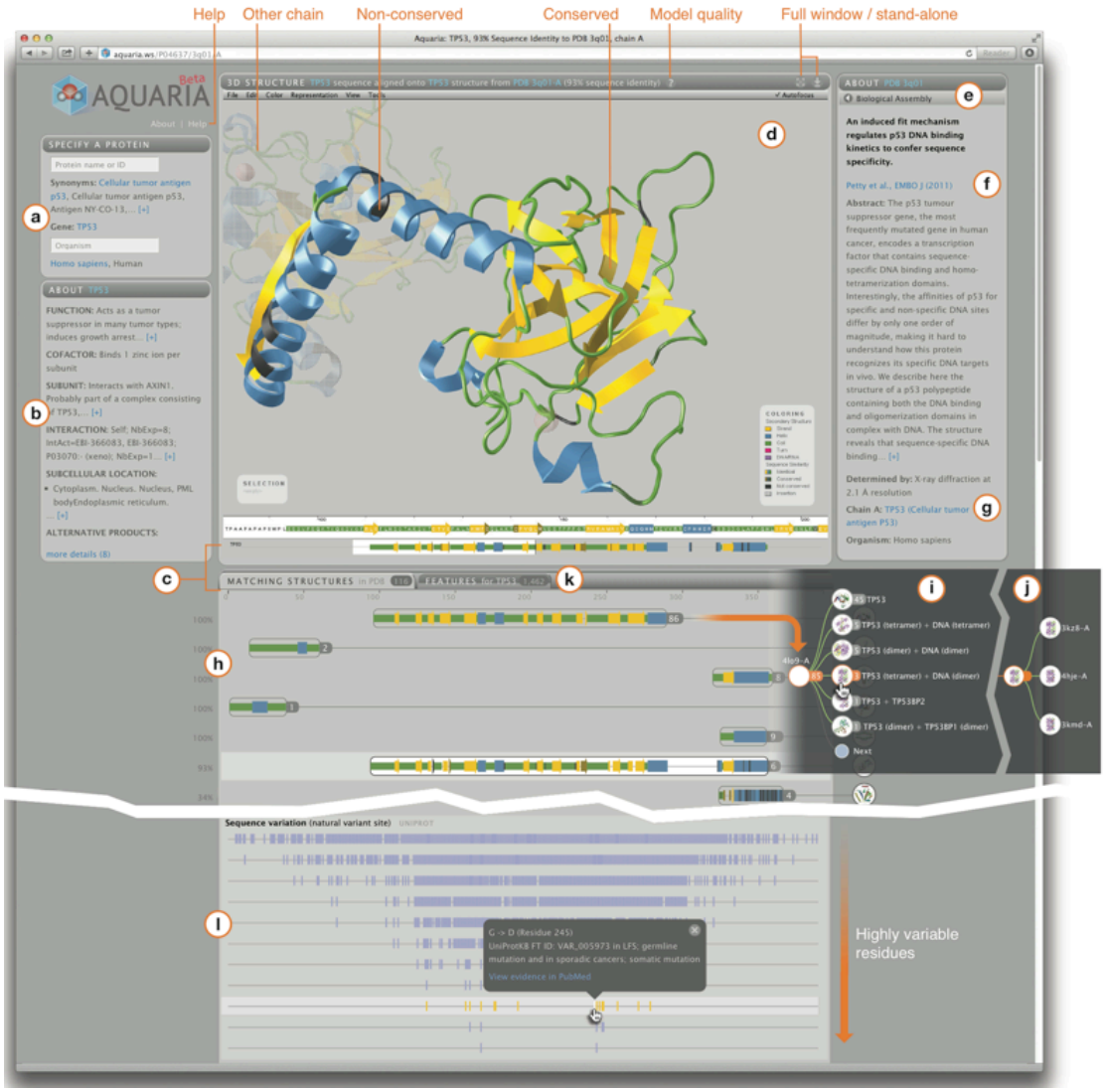


Figure 4.4: Aquaria provides access to all related 3D structures for any specified protein. a) Input fields used to specify a protein. b) UniProt summary of the specified protein's function. c) Graphical representations of the specified protein's sequence. d) Initially shows PDB structure estimated to be most relevant for the specified protein. Hue is used to indicate secondary structure: yellow = strand, blue = helix, and green = coil. Lightness is used to indicate amino acid substitutions between the specified protein and PDB structure: slightly dark = conserved, very dark = non-conserved, white = insertions, and semi-transparent = other chains. e) Initially shows the most likely biological assembly from PDB (http://bit.ly/biounit). f) Information about PDB structure. g) Information about currently focused PDB chain. h) Visual summary of all structures in PDB matching the specified protein, grouped by region of match. Clicking on a group loads the top-ranked structure into the 3D viewer. i) Clicking on a group number shows a tree view of structures in the group, organized into subgroups by oligomeric states. j) Clicking on a subgroup expands the tree to show individual PDB files. k) Provides access to sequence features. l) Shows InterPro and UniProt features for the specified protein; hovering over an individual feature reveals its details, while clicking anywhere on a feature lane uses that set of features to color the 3D structure. Features of the same type are grouped into a minimal number of lanes, avoiding overlap; for example, the specified protein shown (p53) has many sequence variants – the layout highlights residues with the largest number of distinct variations (O'Donoghue et al., 2015).

4.6. Discussion

Our design goals with Aquaria were quite different to those of comparable resources; ease of use was paramount, as was the ability to map sequence features. More fundamentally, the user interface has been organized primarily by protein sequence, not structure. Each Aquaria webpage essentially provides an interactive review of all current structural knowledge for one protein, making it clear which parts of a protein do and do not have matching structures (Fig.4.4h). Assembling this information using existing tools would take days or weeks of effort; Aquaria reduces this to seconds, freeing researchers to focus on the analysis, interpretation and understanding of structural data, rather than on the process of assembling it.

Aquaria also provides millions of new models that will potentially yield significant new insights for a wide variety of proteins. These models give an unprecedented depth of coverage, with more than double the number of structural models currently available from other comparable resources. This, in combination with Aquaria's ease of use, enables researchers to answer new kinds of scientific questions, such as whether an insight obtained by examining one model is supported by all other related models - this can be invaluable, given the uncertainties of experimental structure determination (O'Donoghue et al., 2010).

4.7. Conclusions

Since the discovery of the DNA double-helix, biologists have been aware of the enduring significance of insight gained from atomic-scale structures. Now that a wealth of such structures are available, a key challenge is to benefit from this data deluge, without being overwhelmed by it (O'Donoghue et al., 2010). We believe that Aquaria will help achieve this and, by making structures easier to access and use, will accelerate discovery in the life sciences.

4.8. Author Contributions

Nelson assisted in the co-development of two Perl scripts that parsed information from UniProt and PDB files into several tables that are part of the Aquaria database namely: Protein_sequence, protein_synonyms, organism_names, organism_sinonyms, PDB, PDBchain and PubMed. He also contributed to Aquaria user interface with use of the 'knockout.js' framework for managing communication between components

namely concerning protein identification and fetching of its corresponding info (this platform was not used in the final version). Finally, he also contributed through a script, in the generation of the circled 2D protein images show in Fig. 4.4i

5. Dark Proteome Database

This chapter was partially published in:

"The Dark Proteome Database"
N. Perdigão, A. C. Rosa, S. I. O'Donoghue, BioData Mining, Springer Nature.
(Minor Revision)
doi:
http://

"Visual analytics of gene set comparison"
N. Perdigão, T.G. Soldatos, K.S. Sabir, S.I. O'Donoghue, *IEEE Symposium on Big Data Visual Analytics*, pp. 1-2, Hobart, Australia, 2015
doi: 10.1109/BDVA.2015.7314304
http://ieeexplore.ieee.org/xpl/articleDetails.jsp?arnumber=7314304

5.1. Summary

In this chapter we will describe the Dark Proteome Database (DPD) and associated web interface that provide access to updated information about the dark proteome. The DPD provides information on the regions of proteins where 3D molecular conformation is currently unknown, either via experimental determination or via homology modelling. DPD is assembled from several external web resources (Aquaria, UniProt, Predict Protein, and the Protein Model Portal) and stored in a relational database that currently contains ~10 million entries and occupies ~2 GBytes of disk space. Availability of this database will help focus future structural and computational biology efforts to shed light on the remaining dark proteome, thus potentially revealing molecular processes of life that are currently unknown. The dark proteome database is available at http://darkproteome.ws.

5.2. Introduction

We already seen that knowledge of protein three-dimensional (3D) structure and function can be highly valuable, and has led to key discoveries in the life sciences. The PDB, or Protein Data Bank (Berman et al., 2000), that accumulates experimental structures recently past 120,000 entries – a landmark in our understanding of the molecular processes of life. This lags far behind the growth in DNA sequencing; however, since evolution conserves structure more than sequence (Chothia & Lesk, 1986; Illergård et al., 2009), high-throughput computational modeling (Haas et al., 2013; Petrey et al., 2015); can leverage the PDB to provide accurate structural predictions for a large fraction of the protein sequences inferred from genomic sequencing. Thus structural data now scales with sequencing data and can provide a wealth of detail into molecular functions. Aquaria is therefore an essential tool to determine possible structures and function to old and new protein sequences.

However, there is another side of the structure, or absence of it, which we called the dark proteome, and it's basically the core of this thesis. In this chapter I will map this dark universe (Fig 5.1 Aquaria dark regions) in the most complete and exhaustive way done till today.

5.3. Data

The starting point for the creation of DPD was the Aquaria Database (PSSH2), integrating website information from Swiss-Prot (Consortium, 2014), Predict Protein (PP) (Yachdav et al., 2014) and Protein Model Portal (PMP)) (Berman et al., 2000).

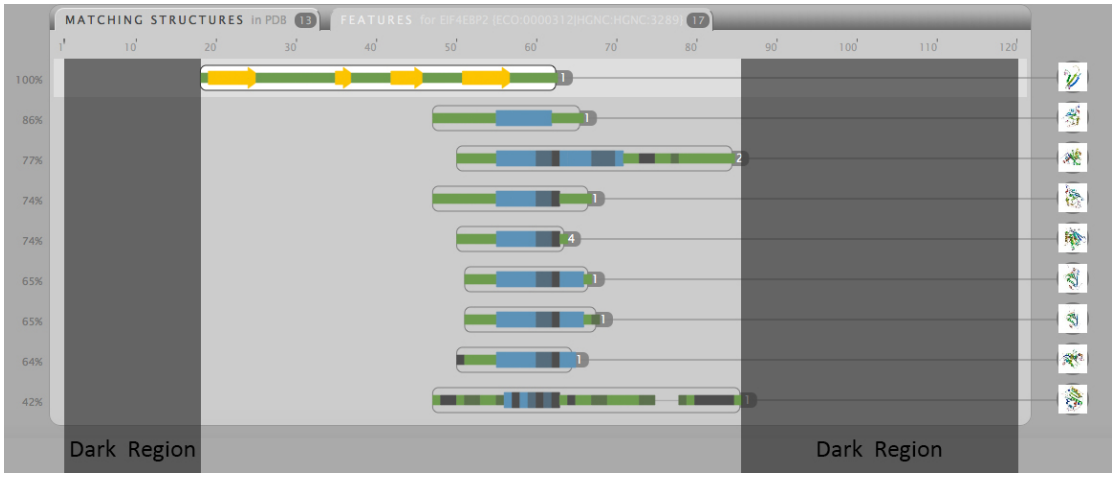


Figure 5.1: Dark Regions in Aquaria for protein Q13542.

The process of readying the data faced similar issues of those detected in Aquaria database preparation: several and many types of exceptions; information inconsistency from the sources mentioned above, when combining them together; generated information quantity and dimension originated several computational and memory problems where optimization techniques were pointed out as solutions. Several exhaustive validation and smoking tests were also made.

Like in Aquaria, the conversion into a relational database from databanks, as well as the mapping of the dark regions present in PSSH2 to its relational form, were necessary steps to perform the analyses and form conclusions.

Today, I can deduct that the Dark Proteome mapping possibly wasn't performed in the past due to the exhaustive, demanding and patient work required to prepare such a volume of data both for PSSH2 and DPD.

5.4. Methods

5.4.1. Database

DPD is created by a pipeline (Fig. 5.2A) that brings together information from Swiss-Prot (Consortium, 2014), the Protein Model Portal (PMP) (Berman et al., 2000), Predict Protein (PP) (Yachdav et al., 2014), and PSSH2 ('Protein Sequence-to-Structure Homologies'), the database underlying Aquaria (O'Donoghue et al., 2015). In the DPD pipeline, the following three initial steps are used to map the dark regions for each protein sequence present in Swiss-Prot (Fig. 5.3A):

1. The first step concerns all sequence-to-structure alignments available in PSSH2. The complete Aquaria entry for each protein is fetched (e.g., http://aquaria.ws/Q13542). This file is then analysed to determine which amino acid residues are not matched to any homologous PDB structure.
2. The second step concerns sequence-to-structure alignments recorded in the corresponding 'database cross-reference' field of each Swiss-Prot entry. These are mappings to PDB entries made using UniProt Consortium criteria (e.g., http://www.uniprot.org/uniprot/Q13542). We used this data to identify a small fraction of regions that contain sequence-to-structure alignments not detected by HHblits (Remmert et al., 2011), the PSSH2 detection algorithm.
3. Similarly, the third step fetches the corresponding PMP entry (e.g. http://www.proteinmodelportal.org/query/up/Q13542) and uses it to identify regions that contain sequence-to-structure alignments missed by both HHblits and UniProt.

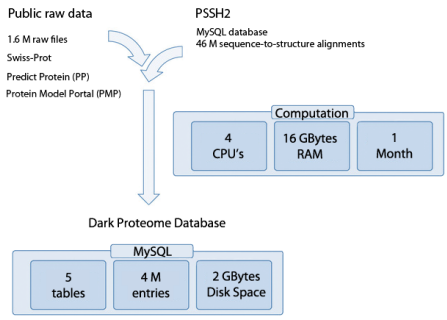
The above information is then used to assemble a MySQL table called 'dark_domains' (Fig. 5.2B). Each entry in this table corresponds to a 'white' or 'dark' region of a protein, defined as follows:

- White regions indicate a contiguous region of the amino acid sequence in which all the residues are aligned to a 3D structure in either PSSH2, UniProt, or PMP (Figs. 5.3A and 5.3B);
- Dark regions are contiguous regions of the amino acid sequence in which no residues are aligned to a structure in the previous point (Figs. 5.3A and 5.3B).

Next, I use the 'dark_domains' table to create a second table called 'dark_proteins' (Fig. 5.2B). Each entry in this table corresponds to a protein, which is assigned to be either 'White', 'Dark', or 'Grey' as follows (Fig. 5.3C):

- White, if and only if the entire amino acid sequence of the protein is a single white domain;
- Dark, if and only if the entire amino acid sequence of the protein is a single dark domain;
- Grey, if the protein contains both dark and white domains.

A



B

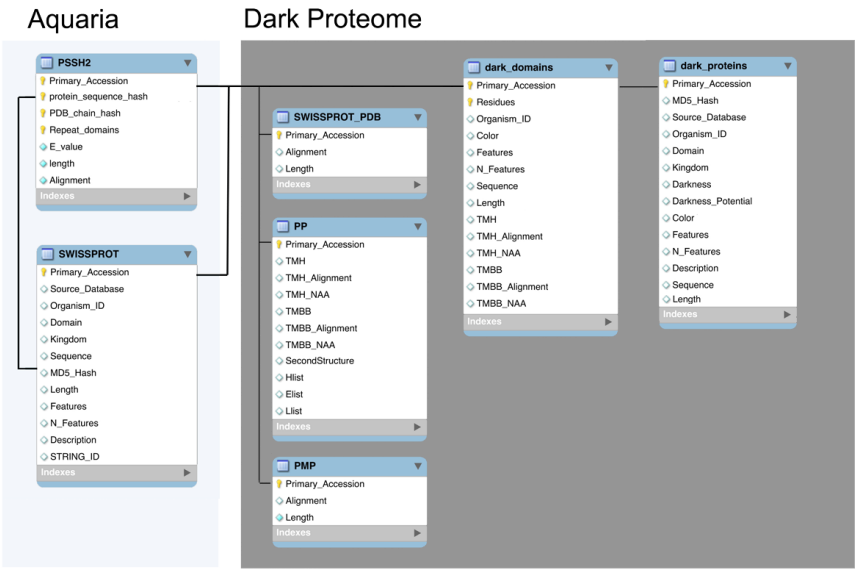


Figure 5.2: A) Flux of data into DPD. B) Overview of Aquaria and DPD schema.

Finally, I created a second version of the DPD that uses only Aquaria and UniProt data. In this version, the 'dark_domains' table is generated as follows:

- White regions indicate a contiguous region of the amino acid sequence in which all the residues are aligned to a 3D structure in either PSSH2 or UniProt;
- Dark regions are contiguous regions of the amino acid sequence in which no residues are aligned to a structure in the previous point.

Similarly, a second 'dark_proteins' table is generated based on this 'dark_domains' table (Figs. 5.4A and 5.4B).

The speed of building DPD is mostly limited by internet bandwidth, as the build process relies on fetching a very large number of files via HTTP from the source services (UniProt, Aquaria, and PMP). Overall the process of fetching and assembling data takes around 1 CPU month using a Quad-core i7; however as most of these source services have multicore servers, the process can be speed up by parallel data fetching.

5.4.2. Web Interface

The database is web-accessible allowing fast access to any Swiss-Prot protein information, revealing either the dark and non-dark regions (e.g., http://darkproteome.ws/database/domains.php?id=Q13542) (Fig. 5.3.B), or the overall percentage of dark residues (e.g., http://darkproteome.ws/database/protein.php?id=Q13542) (Fig. 5.3C). The user can also choose to see data from either version of the database, thus enabling them to use a definition of darkness that either includes (Figs. 5.3B and 5.3C) or excludes PMP (Figs. 5.4A and 5.4B).

Some functional analyses are also provided, by comparing annotations between dark and non-dark sets in a reliable manner where we applied annotation enrichment for the 'Description' field of the Swiss-Prot proteins through Fisher exact tests (Fisher, 1922; Fisher, 1925) with adjustment (Benjamini & Hochberg, 1995; Perdigão et al., 2015). The results of the analyses are presented in a Tag Cloud with pagination (Fig. 5.5) to reveal the most functional terms over- or under-represented in dark-proteins or dark-regions (Perdigão et al., 2015b). Spinning wheels are also available and can be seen as an alternative visualization method to the Tag Cloud, where the results are revealed as sorted lists (Fig. 5.6).

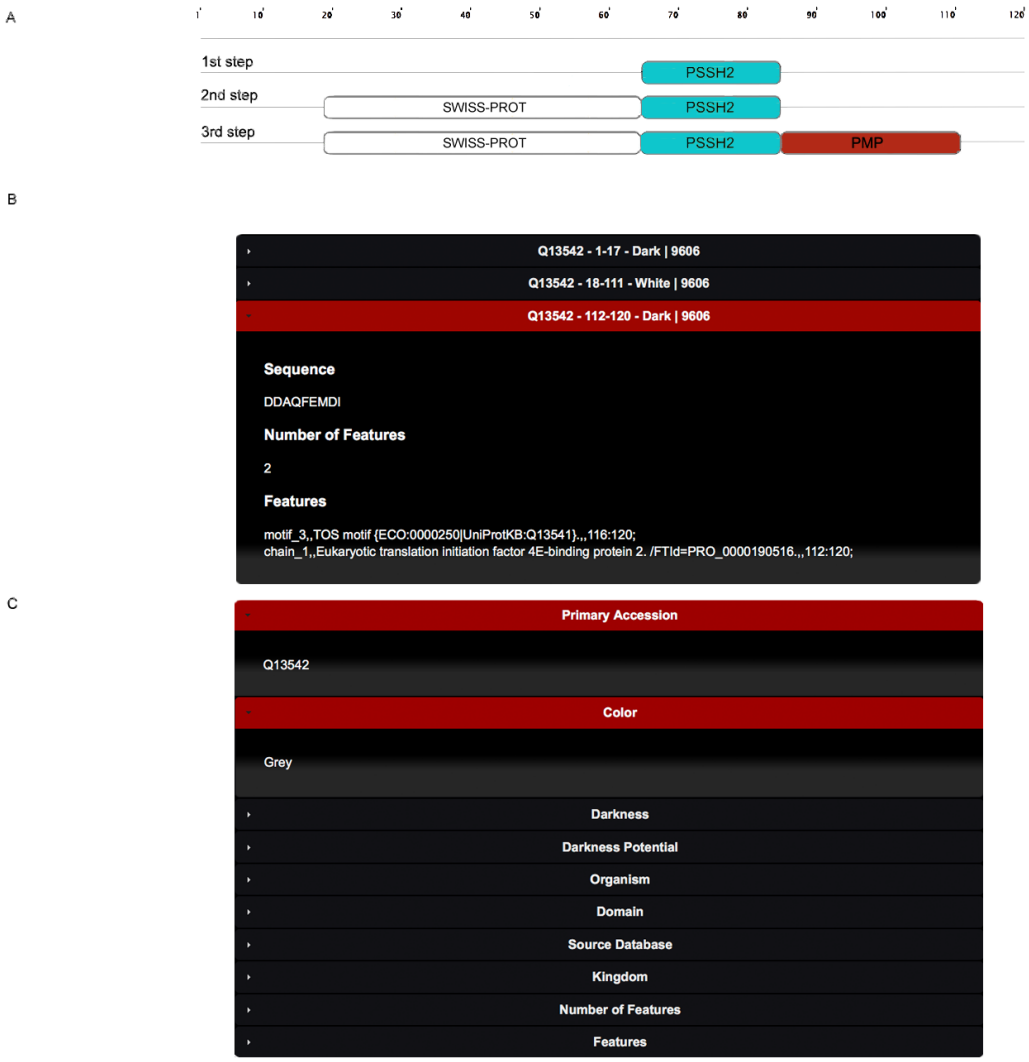
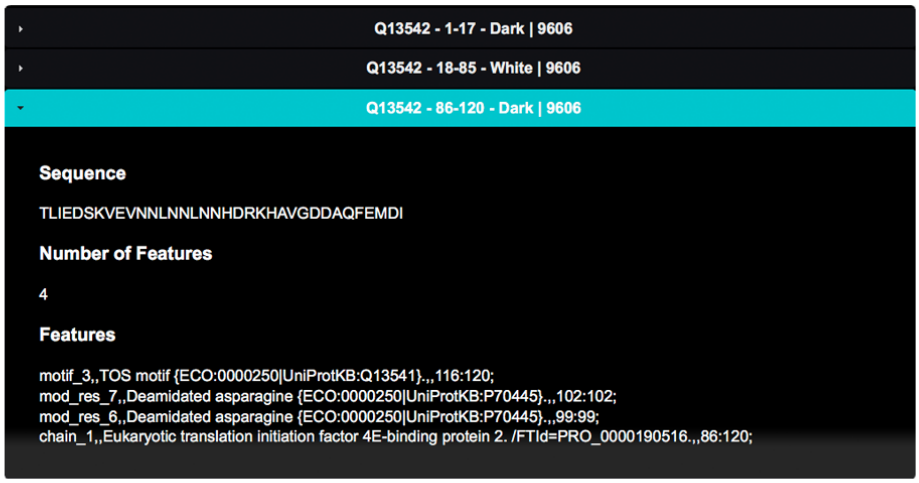


Figure 5.3: A) Three step domains fulfilment for human (organism ID number 9606) protein Q13542. B) dark_domains table holding colour domains for protein Q13542. C) dark_proteins table showing entry for protein Q13542 holding colour Grey for the full protein.

A



B

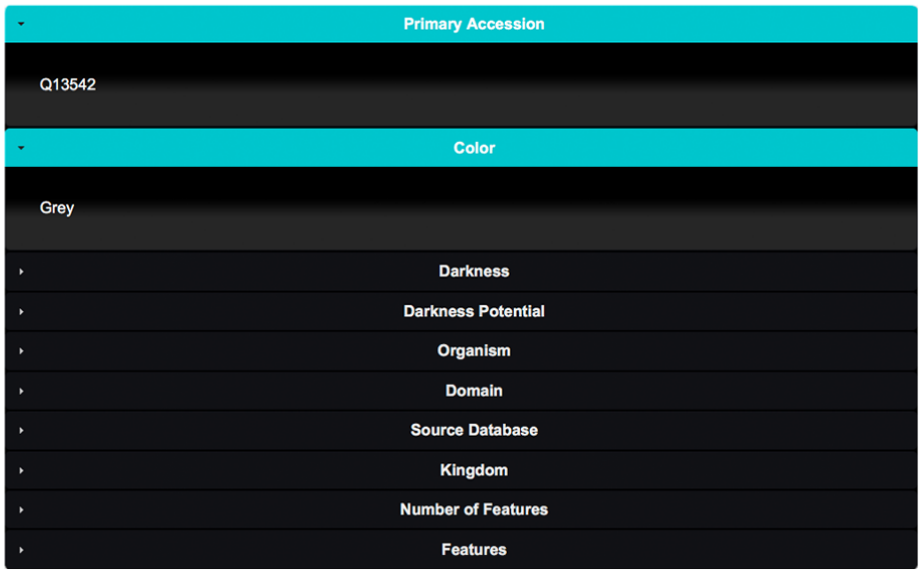


Figure 5.4: A) Dark_domains interface holding colour domains for protein Q13542, where PMP regions are ignored, i.e., they are considered dark. B) dark protein interface showing entry for protein Q13542 holding colour Grey for the full protein.

Figure 5.5: Tag-cloud visualization showing subcellular locations over- or under-represented in dark eukaryotic proteins (dark and white text, respectively). Text size terms in the tag cloud is set to the minus log of significance (score computed by adjusted Fisher's exact test). Annotations are sorted into categories and pages, helping thus making this very large set of annotations more manageable. This tool provides insight into a very wide variety of biological questions.

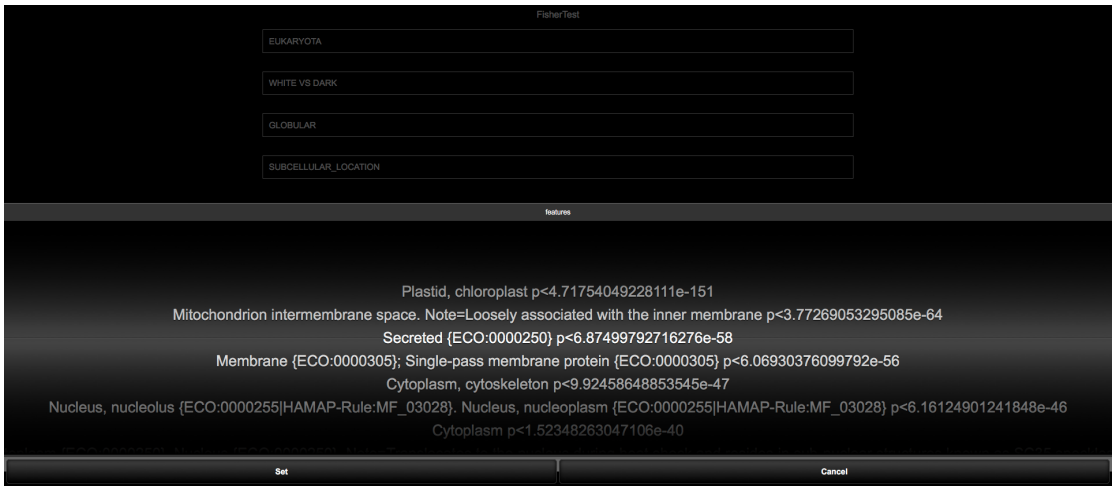


Figure 5.6: Spinning wheel visualization showing subcellular locations over- or under-represented in dark eukaryotic proteins. The order of terms in the spinning wheel is set to the minus log of significance (score computed by adjusted Fisher's exact test). This tool therefore also provides insight into a very wide variety of biological questions.

5.5. Results

The current version of DPD (October 2014) was assembled using PSSH2 and Swiss-Prot from October 2014, and PMP from October 2014. The complete database contains around 10 million entries (including entries both with and without PMP) and occupies around 2Gb in disk space.

The web interface provides users with fast access to individual entries for any protein, revealing either the dark and non-dark regions (e.g., http://darkproteome.ws/database/domains.php?id=Q13542), or the overall percentage of dark residues (e.g., http://darkproteome.ws/database/protein.php?id=Q13542).

The DPD web interface is built using Apache, PHP, MySQL, JQuery, and JQueryUI. On the DPD homepage a client-side AJAX engine initiates HTTP GET requests to the server, sending user-selected options. The AJAX engine notifies the user that a search has been initiated by displaying an animated 'throbber' icon. After the server-side PHP script receives the search options from the GET request, it constructs and executes the appropriate MySQL query on the database. Once the query has been executed, the script builds a JSON object from the result set and returns it to the AJAX engine. Upon receiving the JSON response, the AJAX engine parses it, builds the mark-up for the results and displays it in the browser window.

5.6. Discussion

We will see that specific examination of the dark proteome led to some surprising results (Chapter 6) (Perdigão et al., 2015) and has challenged some of the current beliefs and conceptions, in what concerns the proteome that remain inaccessible to structural biology. We believe there are many further discoveries waiting to be made by further studding these regions, for example, by exploring the role of the dark proteome in specific biological functions or in human health.

5.7. Conclusion

This work contribution will consolidate structural knowledge from Aquaria, UniProt, Predict Protein, and PMP into an easy-to-use interface that gives users quick access to the precise mapping of dark and non-dark regions. Thus, DPD will help focus further research shedding light into the remaining dark proteome, revealing molecular processes of life that are currently unknown.

5.8. Author Contributions

Nelson Perdigão contributions on this chapter were the coding of all algorithms for the generation of the Dark Proteome Database, built of Fisher's exact tests, Protein Model Portal validations, development of Adapted Tag Cloud tool with Pagination and corresponding Spinning Wheels, as well as, writing and revision of the corresponding paper.

PART IV

RESULTS

6. Dark Proteome

This chapter was partially published in:

"Unexpected Features of the 'Dark' Proteome"
N. Perdigão, J. Heinrich, C. Stolte , K. S. Sabir , M. Buckley , B. Tabor , B. Signal , B. S. Gloss , C. J. Hammang , B. Rost, A. Schafferhans, S. I. O'Donoghue, *Proceedings of the National Academy of Sciences*, vol. 112 no. 52, pp.15898–15903, 2015.
doi: 10.1073/pnas.1508380112
http://www.pnas.org/content/112/52/15898

6.1. Summary

We surveyed the 'dark' proteome – that is, regions of proteins never observed by experimental structure determination and inaccessible to homology modeling. For 546,000 Swiss-Prot proteins, we found that 44 – 54% of the proteome in eukaryotes and viruses was dark, compared with only ~14% in archaea and bacteria. Surprisingly, most of the dark proteome could not be accounted for by conventional explanations, such as intrinsic disorder or transmembrane regions. Nearly half of the dark proteome comprised dark proteins, in which the entire sequence lacked similarity to any known structure. Dark proteins fulfill a wide variety of functions, but a subset showed distinct and largely unexpected features, such as association with secretion, specific tissues, the endoplasmic reticulum, disulfide bonding, and proteolytic cleavage. Dark proteins also had short sequence length, low evolutionary reuse, and few known interactions with other proteins. These results suggest new research directions in structural and computational biology.

6.2. Introduction

We surveyed what we call the ‘dark’ proteome (Perdigão et al, 2015) i.e., the regions of proteins inaccessible to experimental structure determination or modelling. We found that most of the dark proteome could not be accounted for by conventional explanations (i.e., by intrinsic disorder, transmembrane proteins or compositional bias), and that nearly half of the dark proteome comprised dark proteins, in which the entire sequence lacked similarity to any known structure.

A range of previous studies have surveyed the ‘white protein universe’ of available information (Chothia, 1992; Holm & Sander, 1996; Levitt, 2009; Nepomnyachiy et al., 2014), i.e., all proteins from all organisms. From such surveys, we know much of the proteome is comprised of evolutionary conserved domains matching a relatively few 3D folds (Chothia, 1992; Holm & Sander, 1996). These surveys have focused on the ‘known’ and on extrapolating progress towards complete knowledge of all folds in the protein universe. Such studies have guided structural genomics initiatives aimed at determining at least one PDB structure for each distinct fold (Khafizov et al., 2014).

This work focuses on the structurally ‘unknown’, i.e., the fraction of the proteome with no similarity to any PDB structure, I call this fraction the ‘dark proteome’; I believe that

by studding it will clarify future research directions, as studies of dark matter have done in physics (Bertone et al., 2005).

The analogy to dark matter has inspired surveys of other 'unknown' properties of proteins; for example, Levitt examined 'orphan' protein sequences that do not match to known sequence profiles, which he termed the 'dark matter of the protein universe' (Levitt, 2009), and Taylor investigated the 'dark matter of protein fold space', i.e., theoretically plausible folds that have not been observed in native proteins (Taylor et al., 2009). The same analogy has been made to studies of so-called 'junk DNA' (Travis, 2002), which revealed a 'hidden layer' of non-coding RNAs (Mattick, 2003). Could surveying the dark proteome also reveal undiscovered biological systems?

In fact, discoveries have already resulted from studying regions of unknown structure, namely intrinsically disordered regions. Long known to confound structure determination (Oldfield et al., 2013) – thus forming part of the dark proteome – disorder was largely ignored until recently (Dunker & Obradovic, 2001), yet is now known to play key functional roles, especially in eukaryotes (Oldfield & Dunker, 2014). Another type of 'dark' regions also have specific biological functions, namely transmembrane segments (Carpenter, 2008). Thus both disorder and transmembrane regions are 'known unknowns', i.e., we know that they are often 'dark'. Could the dark proteome contain 'unknown unknowns', i.e., regions with specific functions, that confound structure determination, and that we are unaware of?

To address this question, I needed to map the dark proteome – i.e., to determine all protein regions that cannot be modeled onto any PDB structure. Most available modeling datasets – collected in the Protein Model Portal (PMP) (Haas et al., 2013) - are not well suited as they aim for breadth of coverage, typically providing only a few PDB matches per protein. Mapping the dark proteome requires depth of coverage, such as the Khafizov survey (Khafizov et al., 2014) – unfortunately however they used only a few model organisms.

Recently, I was involved in Aquaria (Chapter 4) (O'Donoghue et al., 2015) a reported resource where it was made the most detailed analysis of this kind by systematically comparing 546,000 Swiss-Prot sequences against 100,326 PDB proteins structures, which essentially covers all well-described protein sequences across a wide range of organisms. This comparison resulted in 46 million sequence-to-structure alignments (O'Donoghue et al., 2015) a depth of structural information currently not available from other resources.

In this study (Perdigão et al., 2015), I just focused on the dark regions identified by PSSH2 (Chapter 4) with the help of Dark Proteome Database (Chapter 5). By doing so (do to the huge number of entries present in the tables, the heterogenic sources of information, as well as the huge implicit information present), this became a Big Data issue, where the challenges included analysis, capture, data curation, search, storage, transfer and visualization. Accuracy in this case leaded to more confident conclusions and new breakthroughs as the ones that will be presented next.

6.3. Data

The data preparation for this Chapter can be consulted in Dark Proteome Database Chapter where a detailed description is given (Chapter 5).

6.4. Methods

6.4.1. Mapping Darkness

For each Swiss-Prot protein, each residue was categorized 'non-dark' if it met the either of following criteria (Fig. 6.1A):

(a) if the residue was aligned onto the ATOM record of any PDB entry in the corresponding Aquaria matching structures entry (e.g., http://aquaria.ws/ Q13542) or;

(b) if the residue was aligned onto a PDB entry in the corresponding UniProt entry (e.g., http://uniprot.org/uniprot/Q13542).

All other residues were considered 'dark'. I then calculated a 'darkness' score *D* for each protein using:

$$D = \frac{number\ of\ dark\ residues}{total\ number\ of\ residues} \quad \text{(Equation 6.1)}$$

For most proteins, darkness depends on criterion (a), and hence on the criteria Aquaria uses to decide when a given sequence-to-structure alignment is of sufficient quality to infer that a sequence is likely to adopt a structure similar to a given PDB entry. An advantage of using Aquaria for this task is that it is derived from a systematic, all-against-all comparison of Swiss-Prot and PDB sequences; it also uses HHblits

(Remmert et al., 2011), an iterative method that compares Hidden Markov Models (HMMs) of sequences and structures, and gave the best combination of speed and reliable detection for structural templates when compared to around 70 competing methods (http://bit.ly/hhblits-casp9 and http://bit.ly/hhblits-casp10).

Including criterion (b) above decreased the total fraction of all dark residues in Swiss-Prot by only 0.2%; mostly this is accounted for by a small fraction of very short and very long sequence-to-structure alignments missed by PSSH2 (O'Donoghue et al., 2015). In addition, the information contained in UniProt entries sometimes overestimates the region that is matched by PDB entries, including some residues that do not actually appear in the 3D structure – this has the effect of slightly underestimating darkness.

While this definition of darkness is straightforward, it has the limitation that it does not distinguish between strong and weak matches to PDB structures; in addition, we use *all* PDB structures, including those derived from low-resolution crystallography, electron microscopy, or NMR spectroscopy. Thus, we do not distinguish weak sequence matches to low resolution structures from strong matches to very reliable structures – both cases are considered equally non-dark. In Figs. 6.1B and 6.2 this issue is symbolically indicated by the white-to-black gradient in grey domains, which is suggestive of the variation in the quality of structural knowledge for these regions.

Note that Aquaria alignments are generated by first aligning to each Swiss-Prot sequence onto the PDB 'SEQRES' records – i.e., the actual peptides used in the experiments underlying each PDB entry. As a second step, we align the SEQRES records onto the PDB ATOM records; thus, in cases where a region of sequence is always missing in the ATOM records of all related PDB entries (e.g., loop regions where electron density is always missing due to large disorder), these residues will be counted as 'dark'.

Unfortunately, a different standard practice is used in NMR-derived structures; when a region lacks experimental data, coordinates for all atoms are still calculated and included in the ATOM records, resulting in highly disordered regions. Thus, these regions are considered 'non-dark' in this work which, again, slightly underestimating darkness.

Note that this definition of darkness is a stringent one, in that it underestimates darkness, or equivalently overestimates the state of structural knowledge for the proteome. We deliberately chose such a stringent definition as it gives more confidence

that the dark regions and dark proteins identified are truly dark, which suits the goals of the current work.

Most dark residues occurred within contiguous *dark regions* (Fig. 6.1A); when these are conserved across many other proteins, we call them *dark domains*. In some cases, a single dark region covers the entire sequence – we call these *dark proteins* (Fig. 6.1B). In this chapter we focus primarily on characterizing dark proteins.

6.4.2. Defining Darkness More Stringently (D_{PMP})

To test the robustness of our results, and ensure that our conclusions do not rely solely on Aquaria and HHblits, I also calculated a modified darkness score (D_{PMP}) by augmenting the above definition of non-dark residues to additionally require that:

(c) if the residue occurs in any 'twilight' or 'safe' zone model in the PMP (Haas, et al., 2013) (e.g., http://www.proteinmodelportal.org/query/up/Q13542).

The models in PMP are aggregated from a range of resources, and hence have been calculated by a variety of different methods. I excluded PMP models annotated as having very low quality ('midnight' zone, i.e., less than 10% of identity (Rost, 1997)), as many of these are expected to be inaccurate or to have the wrong fold.

However, using this more stringent criterion for defining the dark proteome, I saw very little difference in the overall distribution of dark regions and proteins across various groups of organisms (Fig. 6.1B compared to Fig. 6.2), or in the fraction of dark proteins that remain unexplained by disordered or transmembrane proteins (Fig. 6.11 compared to Fig. 6.13).

The key difference that I saw was in higher eukaryotes such as human, where dark proteins reduced from 4,382 (22%) to 2,267 (11%); similarly, dark proteins in mouse reduced from 18% to 9%. This most likely arises from the fact that several of the databases that PMP draws its models from have a bias towards modeling proteins from higher eukaryotes (Haas et al., 2013).

6.4.3. Database Biases

This work is based on Swiss-Prot (Consortium, 2014), a manually annotated database of non-redundant protein sequences from 13,110 organisms. Swiss-Prot has a bias towards well-studied proteins from model organisms; however, it is arguably the

most reliable resource available for defining a set of proteins whose existence is supported by experimental evidence. Using Swiss-Prot partly addresses one potential explanation for dark proteins – they may not be proteins, but in fact unrecognized long non-coding RNA or may arise from pseudogenes. Using Swiss-Prot reduces this likelihood.

The PDB (Berman et al., 2000) also has a similar bias towards model organisms, although this is reduced somewhat by structural genomics initiatives (Marsden et al., 2007). The effect of bias in the PDB is further reduced by the systematic modeling approach in Aquaria, which extends structure information to all detectibly related sequences in Swiss-Prot. Ultimately, these biases need to be taken into consideration in interpreting the results obtained in this work; essentially, the results document the fraction of well-described protein sequences that can be mapped onto any of the known 3D structures.

If this approach was extended to include a broader set of proteins and organisms, for example by using TrEMBL (Consortium, 2014), the distributions would be expected to change – most likely the dark proteome would increase.

The dark proteome datasets used in this chapter were compiled from Aquaria, PDB, Swiss-Prot, Predict Protein, and PMP in October 2014; thus they do not reflect structure and sequence entries deposited since then. It is planned to update the online resource annually; while many database entries change with each update, over the three years that we have studied this dataset we have observed that the key results reported in this work have not changed, as would be expected since they are supported by rather large sample sizes, with correspondingly small *p* values.

6.4.4. Density Plots

The density plots in Figs. 6.6, 6.7, 6.8 and 6.9 were created using Gaussian kernel density estimations (Silverman, 1986), as implemented in the 'stat_density' and 'stat_density2d' functions of the 'ggplot2' package in R, and using default parameters. In these plots, the total proportion of proteins within a specific range on the *x*-axis can be determined by assessing the area under the curve in that range, and divided by the area across the full range. This enables direct comparison of the relative frequency of dark and non-dark proteins. However, in some cases density plots can be misleading, as different kernel bandwidths produce different plots; for example, Fig. 6.9G shows that dark proteins have a very high but narrower peak at $x = 0$ (corresponding to 0%

transmembrane residues), while the corresponding peak for non-dark proteins is about half the height but broader. However, using other kernels and bandwidths for the same data gives very similar sized peaks at $x = 0$.

Note that for the density plots in Figs. 6.6, 6.7, 6.8 and 6.9 the strongest peak occur close to x = 0%, and occasionally a secondary peak occurs at x = 100% (Fig. 6.6A). Both these situations slightly complicate the interpretation of the area under the curve, since the kernel density method used places some of the area at x < 0% and some at x > 100% - a range of value that we could not include in Figs. 6.6, 6.7, 6.8 and 6.9. However, this minor complication does not detract from the key observation in the density plots in Figs. 6.6, 6.7, 6.8 and 6.9, namely that the majority of the density lies close to x = 0%.

For all density plots in this work, the density values (y-axis) are scaled so that the total area under the curve equals 1 - as a result, the density values depends on the range of values on the x-axis. Therefore, plots that have small range of x values, such as Fig. 6.6 (which ranges from x = 0 to 1), will have relatively large density values (in this case, up to 60).

6.4.5. Disorder

The disorder values shown in Figs. 6.6, 6.7, 6.8, and 6.9 were calculated from IUPred (Dosztányi et al., 2005), one of the most widely used methods for predicting disorder. Residues were defined as disordered if they had an IUPred score ≥ 0.5. As a control, it was also calculated a second set of disorder values using MD (Schlessinger et al., 2009), a 'META-Disorder' machine-learning method that calculates a consensus disorder from several orthogonal methods. Re-plotting the density and scatterplots from Figs. 6.6, 6.7, 6.8, and 6.9 using MD disorder gave a similar overall pattern, although some differences were apparent (Fig. 6.10). MD includes as one of its input methods DISOPRED2 (Ward et al., 2004b), which is one of several available methods that are optimized to predict residues missing from PDB structures. Methods such as this predict a mixture of both darkness and disorder, unlike methods such as IUPred, which focus on predicting disorder only. Thus, we considered IUPred to be preferable to MD (Schlessinger, 2009) or DISOPRED2 for examining the relationship between darkness and disorder, explored in Figs. 6.6, 6.7, 6.8, and 6.9.

For a small fraction of proteins there was not MD predictions; to balance the comparisons, these proteins were removed from the density and scatterplots in Figs. 6.6C, 6.7C, 6.8C, and 6.9C – thus reducing the number of proteins to 175.646, 18.999, 326.945 and 16.316, respectively.

Intrinsic disorder in proteins is a complex and poorly understood phenomenon; in addition to IUPred, many other prediction methods have been developed focusing on a range of different aspects of disorder (Ward et al., 2004) (Schlessinger et al., 2009). It would certainly be of interest to compare darkness with disorder predictions from a range of methods; however, such a detailed comparison of this single property was beyond the scope of this thesis.

6.4.6. Compositional Bias

In the proteins universe, a compositional bias is understood as a particular amino acid or a pattern of residues that are over-represented. An example of such compositional bias could be: AAAAAAAAAAAA.

A compositional bias score was calculated (shown in Figs. 6.6E, 6.7E, 6.8E, and 6.9E) for each Swiss-Prot protein by pooling all residues annotated as compositionally biased in the 'Features' section of the corresponding UniProt entry; this number was then divided by the total number of amino acids. UniProt does not annotate compositional bias occurring within known protein domains, so this method partly underestimates the total compositional bias;

6.4.7. Transmembrane

A transmembrane score was calculated (shown in Figs. 6.6, 6.7, 6.8, and 6.9) for each Swiss-Prot protein by pooling all residues annotated as either intra- and transmembrane in the 'Features' section of the corresponding UniProt entry; this number was then divided by the total number of residues. Most of these UniProt annotations derive from machine learning methods that are believed to predict transmembrane regions with >95% accuracy (Rost et al., 1995). As a control, I also calculated a second set of the transmembrane values by running systematic predictions for all Swiss-Prot sequences with PROF (Rost et al., 1995) and PROFTMB (Bigelow & Rost, 2006), which predict transmembrane helices and beta barrels, respectively. Using these values, the re-plotted density and scatterplots gave almost identical patterns to that obtained using UniProt annotations (Figs. 6.6F, 6.7F, 6.8F, and 6.9F) and also

had the same median values (i.e. zero transmembrane residues for both dark and non-dark proteins in eukaryotes, bacteria, archaea, viruses). The relatively low percentage of transmembrane proteins amongst dark proteins is somewhat surprising (Fig. 6.11); this is partly due to the success of ongoing efforts in the structural biology community to tackle this difficult class of proteins. An additional explanation is suggested by the unexpected trend seen in Figs. 6.6F, 6.7F, 6.8F, and 6.9F, where multi-pass transmembrane proteins become unexpectedly rare at ≥ 25% darkness, and where – for some groups of proteins – there appears to be a linear inverse relationship between darkness and percentage of transmembrane residues. This may be evidence that the prediction methods used here to detect transmembrane regions are progressively failing with increasing darkness, i.e., that methods such as PROF (Liu & Rost, 2001) have lower recall than is currently believed. If correct, this implies the existence of transmembrane regions not detectable using current approaches, presumably because such regions have novel features that have not been seen in existing structures of transmembrane proteins; this may occur either because they simply have not been studied or because these putative, novel transmembrane regions make proteins currently inaccessible to structure determination.

6.4.8. 2D Plots

There was a wide variety in the number of points in each 2D plots (Figs. 6.6, 6.7, 6.8, and 6.9), from ~17.000 in viruses to ~330,000 in bacteria. Thus for each plot it was, manually adjusted the point size and transparency to reveal the 2D distribution as clearly as possible. These adjustments should be taken into account when comparing different plots.

6.4.9. Linear Diagrams

To determine the fraction of dark proteins that could be accounted for by a combination of disorder, transmembrane regions, or compositional bias, it was categorized each protein as having either a ‘high’ (≥ 25%) or ‘low’ (<25%) value for each of the corresponding scores. These results were then displayed in Figs. 6.11, 6.12, and 6.13 as linear diagrams (Huntley & Golding, 2002) which can show categorical combinations (similar to Euler diagrams) for example in eukaryotes and viruses a visible fraction of proteins had both ≥ 25% disorder and ≥ 25% compositional bias. A

much smaller fraction of proteins (<<1%) had both ≥ 25% disorder and ≥ 25% transmembrane fraction however this was too small to represent in Fig. 6.11. For brevity, the fraction of proteins with < 25% for each of these properties is referred to as "ordered, globular, and as having low compositional bias".

Obviously, many important details will be obscured by the use of such a simplistic categorization based on an arbitrarily threshold (25%). Nonetheless, this approach enabled us to create a visualization that gives clear insight into the size of the 'unknown unknown' (Figs. 6.11, 6.12 and 6.13).

6.4.10. Annotation Enrichment

For each Swiss-Prot protein I extracted a set of annotations from the 'Description' field of the corresponding UniProt entry. To compare the annotations from sets of dark and non-dark proteins, I used Fisher's exact test (Fisher, 1922; Fisher, 1925) (two-tailed) to identify annotations that were either over- or under-represented in dark proteins. I applied the Benjamini-Hochberg false discovery correction (Benjamini & Hochberg, 1995) with α, the fraction of false positives considered acceptable, set to 1%, and accepting only annotations with an adjusted *p* value of ≤ 1%, calculated via:

$$p^{adjusted} = Min[p \times n / (k+1), 1] \quad \text{(Equation 6.2)}$$

where *p* is from Fisher's test, *n* is the total of number of annotations in the set, and *k* is the rank of the largest *p*-value that satisfies the false discovery criteria. This approach was then repeatedly applied to compare dark and non-dark proteins across various sets of organisms – e.g., one analysis compared all annotations from all eukaryotic proteins (Fig. 6.15C). The *p* values in Figs. 6.15 have been adjusted, as described above. The enrichment results are available in Table 6.3.

6.5. Results

Mapping the dark proteome. We based our survey on 546,000 Swiss-Prot sequences (O'Donoghue et al., 2015). Although smaller than other databases (e.g., > 50 million sequences in TrEMBL (Bairoch & Apweiler, 2000), Swiss-Prot is meticulously curated; each entry has many annotations and a high likelihood that it represents a native protein.

Fig. 6.1A shows how we mapped the dark proteome: for each Swiss-Prot sequence,

each residue was categorized as 'non-dark' if it was aligned to a PDB entry in Aquaria, and as 'dark' otherwise (see Methods).

This definition partly underestimates the dark proteome, since Aquaria includes very remote homologies using HHblits and uses all PDB entries, including low-quality structures from electron microscopy (EM) or nuclear magnetic resonance (NMR) spectroscopy. We deliberately chose this stringent definition of darkness so we can be confident that the dark proteome has *completely* unknown structure.

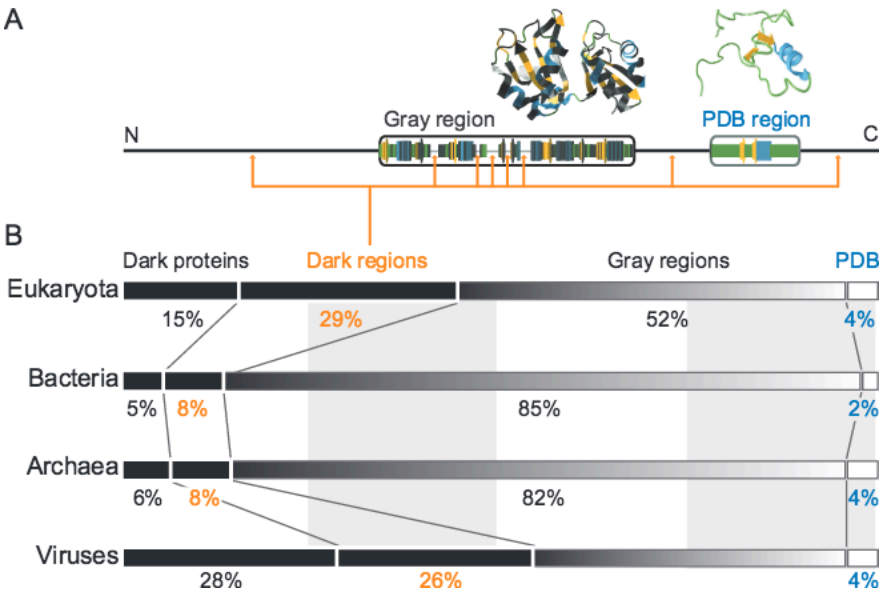


Figure 6.1: Dark proteome overview. A) For 546,000 Swiss-Prot sequences I classified each residue into four categories: 1) *PDB regions*: aligns with exact match to at least one PDB entry, 2) *Grey regions*: aligns with reliable similarity to at least one PDB entry, 3) *Dark regions*: no reliable similarity to any PDB entry, and 4) *Dark proteins*: where a single dark region spans the entire sequence. On average, eukaryotic proteins contain eight dark regions, many very short; some are *dark domains*, i.e., conserved dark regions that evolved independently. B) I pooled sequences by organism group and calculated the total fractions of amino acids in the above categories (Perdigão et al., 2015).

Most dark residues occurred in contiguous *dark regions* (Fig. 6.1); on average, eukaryotic proteins contained eight dark regions, many very short. In many cases, a single dark region covered the entire sequence; we call these *dark proteins* (Fig. 6.1B). Most non-dark residues also occurred in continuous regions: some, called *PDB regions*, exactly match to a PDB entry – these account for only 2-4% of all Swiss-Prot residues

(Fig. 6.1B). The remaining *grey regions* are detectably similar to at least one PDB entry, thus we can predict their 3D structure.

We calculated a *darkness* score for each protein, defined as the percentage of dark residues (Eq. 6.1). Thus, dark proteins have 100% darkness, while proteins with 0% darkness are those where the entire sequence is detectably similar to one or more PDB entries. The distribution of darkness scores was strongly bimodal; most proteins had either low or 100% darkness (density plots in Figs. 6.6, 6.7, 6.8, and 6.9). For brevity in this thesis, we use the term *non-dark proteins* to refer to those with < 100% darkness (noting that a small fraction has a high darkness scores).

We found that the dark proteome (i.e., the fraction of residues in dark proteins or dark regions) for archaea and bacteria was strikingly small (13-14%, Fig. 6.1B), implying that structural knowledge for these organisms approaches a level of completeness. In contrast, in eukaryotes and viruses about half (44-54%) of the proteome is dark (Fig. 6.1B). Of the total dark proteome, about half (34-52%) is comprised of dark proteins. We repeated the above analysis using an even more stringent definition for darkness – combining PMP with PSSH2 (see Methods) – but this had little effect (Figs. 6.2).

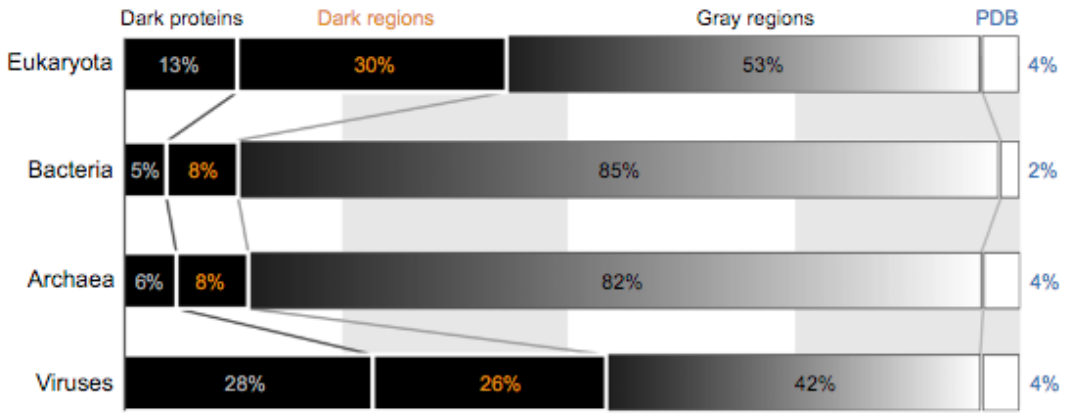


Figure 6.2: Overview of the dark proteome defined using PSSH2 and PMP. Similar distributions are plotted as for Fig. 6.1B, but now defining dark residues to be those with no matching structures in either PSSH2 or PMP. Although this definition of 'darkness' is more stringent, the overall fractions for dark regions and dark proteins are only slightly reduced. The most visible change is a very slight reduction in darkness for eukaryotes (Perdigão et al., 2015).

Figure 6.3 is what we called a darkness profile (in this case for 178.692 eukaryotes proteins) and we are asking how many of these proteins have darkness between 0% and 100%.

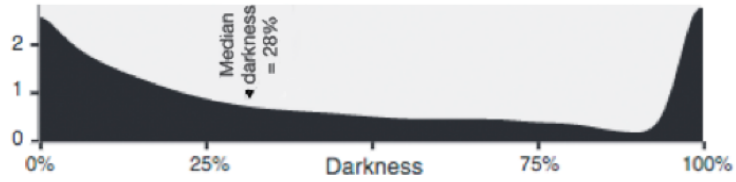


Figure 6.3: The distribution of darkness (i.e., the fraction of dark residues per protein) is bimodal; where 50% of these proteins have low darkness (≤ 28%), while 20% (36,153) have 100% darkness (in the rightest pick).

Rotating the above darkness profile and analyzing the same 178.692 eukaryotes proteins, where for each one of them we measure also a disordered value through IUPred (See Methods) we found something very interesting that is most of proteins fall above the diagonal and that means that darkness is greater than disorder and also means that most of dark regions are not disordered! This for proteins that have less than 100% darkness (Fig. 6.4).

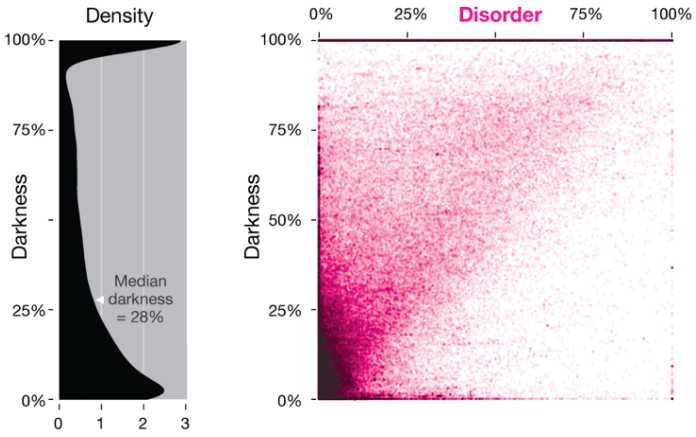


Figure 6.4: Darkness tends to increase with disorder, and the majority of highly disordered proteins are dark (Perdigão et al., 2015).

If we analyze now only the dark proteins (i.e., the proteins with 100% darkness - top right black horizontal line of Fig. 6.4) versus the non-dark proteins (including proteins like 99% dark – all the pink area square), we get the surprise that the levels of

dark between completely dark proteins and non-dark proteins are relatively similar (10% vs 6%) – Fig. 6.5.

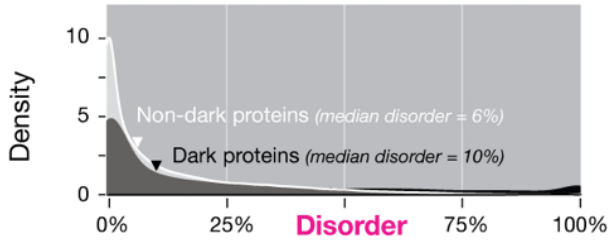


Figure 6.5: The distribution of disorder (i.e., the fraction of disordered residues per protein) shows that disorder is slightly more prevalent amongst dark proteins, but most dark proteins have low disorder (Perdigão et al., 2015).

The full relationship between the previous figures can be observed in Figs. 6.6A, 6.6B and 6.6C. The build of the following figures (Figs. 6.6, 6.7, 6.8, and 6.9) follows the same above methodology.

Dark Proteome Mostly Not Disordered. Intrinsically disordered regions are believed to account for much of the dark proteome, especially in eukaryotes (Oldfield et al., 2013). To explore this, for each protein we calculated the percentage of disordered residues using IUPred (Dosztányi et al., 2005) (see Methods). Viewing these disorder and darkness scores on a 2D scatter plot we see that darkness was greater than disorder for almost all eukaryotic proteins (most proteins above the diagonal in Figs. 6.4 and 6.6C), implying that many dark residues are not disordered or have lows levels of disorder. In this 2D plot, dark proteins are difficult to resolve as they cluster on a line at the top; thus we made density plots comparing the disorder distribution for dark vs. non-dark proteins (Figs. 6.4 and 6.6B). Surprisingly, most dark proteins had low disorder (≤ 10%), not greatly different than non-dark proteins (median 6% disorder); also, since both these medians were less than half of the median darkness 28% (Figs. 6.3 and 6.6A), this implies that – in eukaryotes – most of the dark proteome was not disordered.

In bacteria, archaea, and viruses – surprisingly – non-dark proteins had higher median disorder than dark proteins (Figs. 6.7B, 6.8B, and 6.9B). However, the median darkness was always higher still, implying that in these organisms as well, much of the dark proteome was not disordered.

For eukaryotic proteins the pattern seen in the 2D plot (Figs. 6.4 and 6.6C) also implies that – as expected – most disordered residues were dark. However, a fraction of proteins occurs below the diagonal, implying that many disordered residues were not dark. In the corresponding plots for bacteria, archaea, and viruses this fraction is even larger (Figs. 6.7C, 6.8C, and 6.9C), implying that as much as half of all disordered residues were not dark. Many of our colleagues found this last result confusing, often because they were unclear about the distinction between disorder and darkness. Thus, to clarify, disordered regions are those with evidence of structural heterogeneity (Dosztányi et al., 2005) – yet some become well-structured in particular contexts (e.g., most of the 536 Swiss-Prot proteins with 100% disorder and 0% darkness were ribosomal, and presumably well-structured within the ribosomal complex). To clarify darkness: these are regions that do not match any PDB entry – but some PDB entries are highly disordered especially those from EM or NMR (Ota et al., 2013), and any sequence aligned to them was classified as 'not dark' using our stringent definition, since *some* structural information is known.

Dark Proteome Mostly Not Compositionally Biased. Compositional bias is also known to confound structure determination (Huntley & Golding, 2002). To explore this, for each protein we calculated the percentage of compositionally biased residues (see Methods). Viewing these compositional bias and darkness scores on 2D scatter plots we see that darkness was greater than compositional bias for almost all proteins (Figs. 6.6E, 6.7E, 6.8E, and 6.9E), implying that – as expected – most compositionally biased residues were dark. Together with the density plots for compositional bias (Figs. 6.6D, 6.7D, 6.8D, and 6.9D), it is clear that darkest residues were not compositionally biased, and that most dark proteins had very low compositional bias.

Dark Proteome Mostly Not Transmembrane. Transmembrane regions are also known to confound structure determination (Oldfield et al., 2013; Carpenter et al., 2008). To explore this, for each protein we calculated the percentage of transmembrane residues (see Methods). Viewing these transmembrane and darkness scores on 2D scatter plots we see that a surprisingly large fraction of transmembrane residues was not dark (Figs. 6.6F, 6.7F, 6.8F, and 6.9F). From transmembrane density plots (Figs. 6.6G, 6.7G, 6.8G, and 6.9G) we also see that most dark proteins had no transmembrane residues; zooming these plots shows (as expected) that dark proteins

were strongly overrepresented amongst integral transmembrane proteins in bacteria and archaea – but (unexpectedly) not so in eukaryotes and viruses.

Also unexpected was that the transmembrane fraction tended to decrease with darkness in eukaryotes and – across all organisms – was unexpectedly low in proteins with 75% ≤ darkness < 100% (Fig. 6.14).

These results suggest that knowledge of eukaryotic transmembrane protein structures may be more complete than commonly believed, thanks to an ongoing focus on membrane protein structures (Punta et al., 2009). An alternative suggestion is that the methods used to predict transmembrane regions in this work progressively fail with increasing darkness – i.e., there may be transmembrane regions that are currently undetectable via PROF (Rost et al., 1995), PROFTMB (Bigelow & Rost, 2006), and other similar methods.

Shorter Sequence Length. Very short or long sequence length can confound structure determination (Slabinski et al., 2007). We found that dark proteins had 26-50% shorter median length (Figs. 6.6I, 6.7I, 6.8I, and 6.9I) and 16% had length < 50 or length > 700 amino acids, compared with 11% of non-dark proteins. So, extreme length may explain some dark proteins, but not most.

Since dark proteins are shorter, their abundance is underestimated in Fig. 6.1 which is based on the fraction of dark residues. The fractions for dark proteins were: 20% for eukaryotes, 7% for bacteria, 8% for archaea, 44% for viruses, and 13% for all Swiss-Prot proteins.

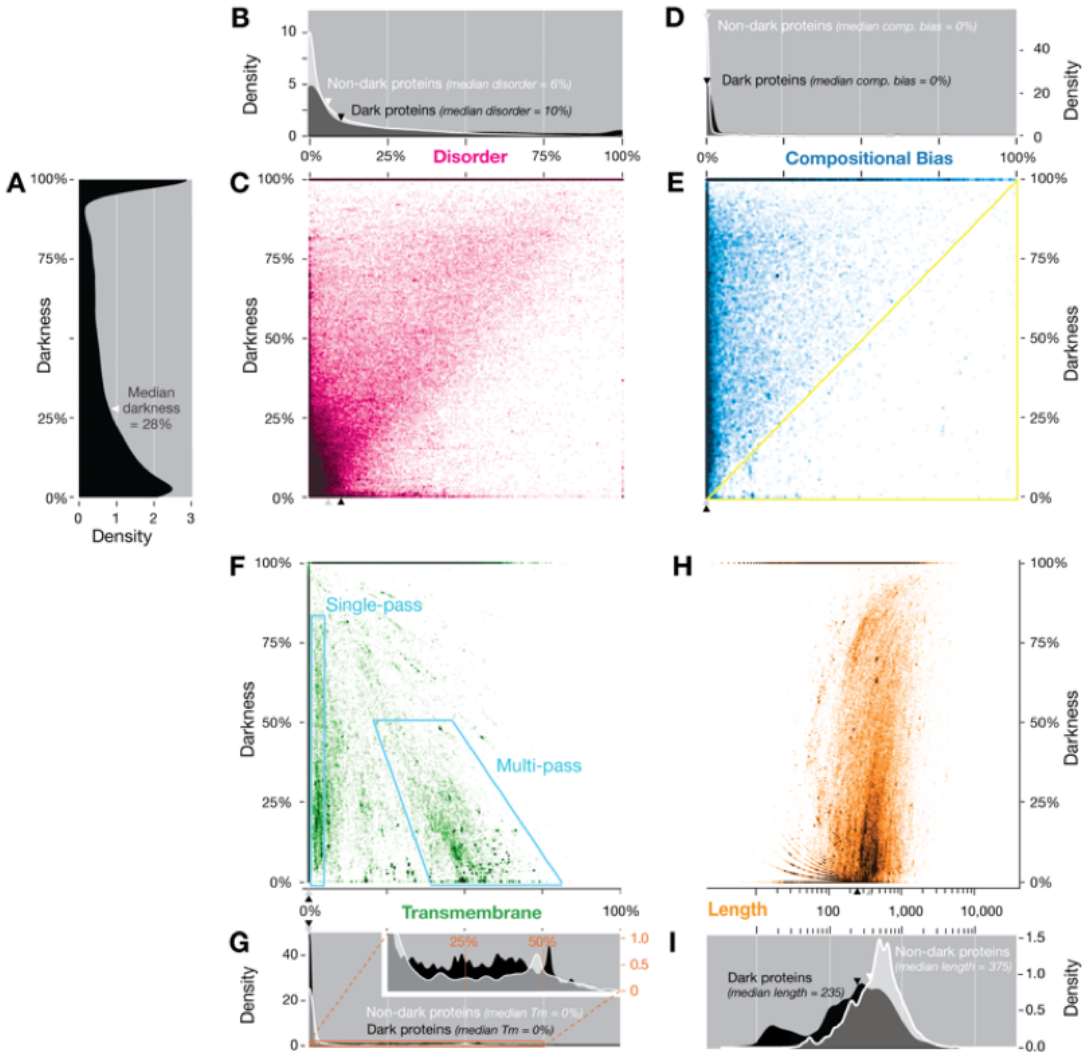


Figure 6.6: Darkness vs. other properties for 178,692 eukaryotic proteins. A) The distribution of darkness (i.e., the fraction of dark residues per protein) is bimodal; 50% of proteins have ≤ 28% darkness, while 20% (36,153) have 100% darkness. B) The distribution of disorder (i.e., the fraction of disordered residues per protein) shows that disorder is slightly more prevalent amongst dark proteins, but most dark proteins have low disorder. C) Darkness tends to increase with disorder, and the majority of highly disordered proteins are dark. D) Compositional bias is low for all proteins, but slightly more prevalent for dark. E) Very few proteins occur in the indicated triangular region, suggesting that most compositionally biased regions are dark. F) Multi-pass transmembrane proteins become unexpectedly rare at ≥25% darkness. G) Proportionally more dark proteins are multi-pass transmembrane proteins (zoomed-in insert); however, most dark proteins have no transmembrane regions. H) Darkness tends to increase with sequence length (note the log scale). I) In contrast, dark proteins tend to be shorter.

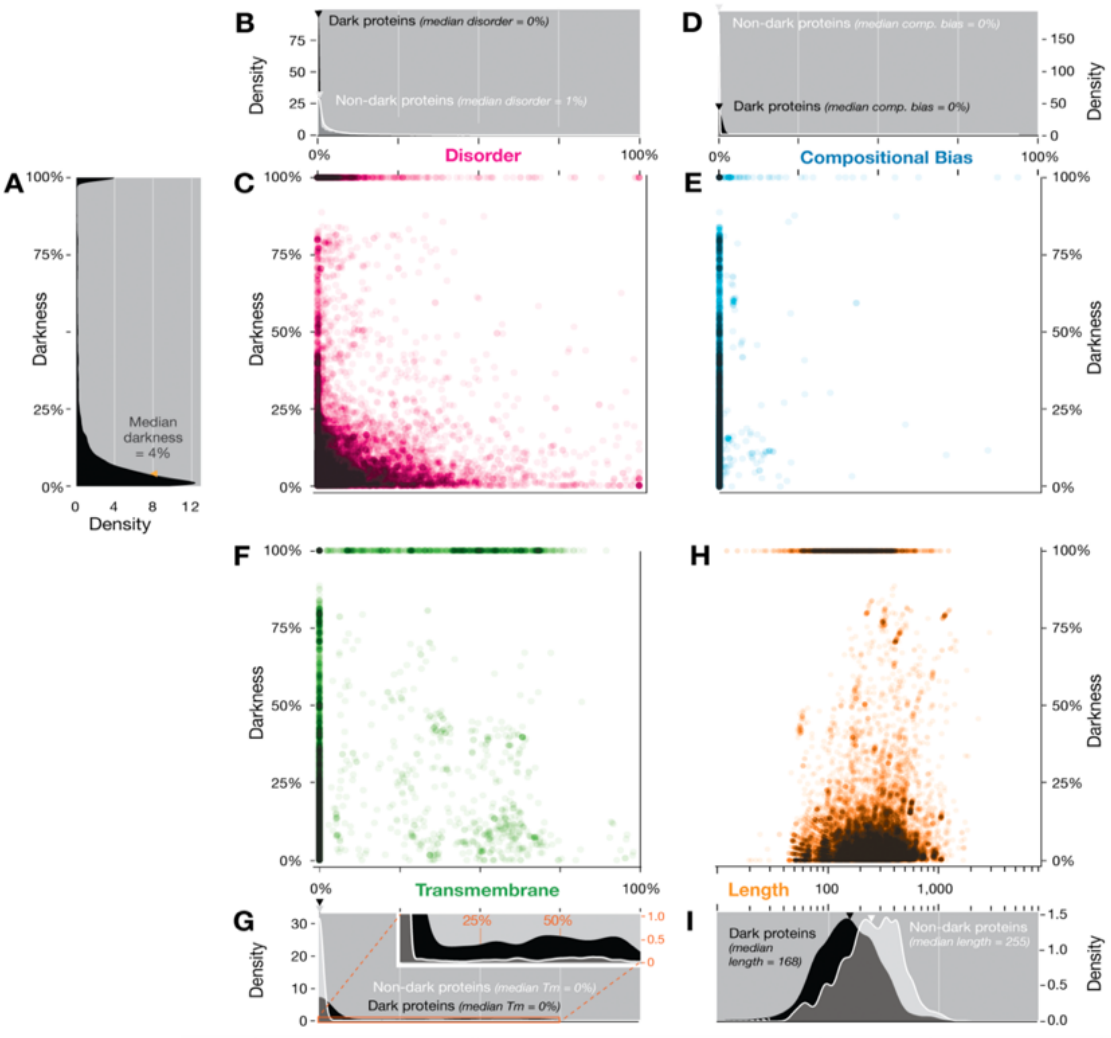


Figure 6.7: Darkness of 19,270 archaeal proteins compared to other properties. A) The distribution of darkness is strongly bimodal; at least 50% of proteins have ≤ 4% darkness, while 8% (1,612) have 100% darkness. B) The distribution of disorder shows that – surprisingly – non-dark proteins have slightly more disorder (1% median) than dark proteins (0% median). Overall, almost all dark proteins have low disorder. C) There is no clear relationship between darkness and disorder. D and E) Almost all dark proteins have very low compositional bias. F) As with eukaryotes, multi-pass transmembrane proteins become unexpectedly rare at more than 25% darkness. G) Dark proteins are much more prevalent amongst multi-pass transmembrane proteins (zoomed-in region); however, most dark proteins have no transmembrane residues. H) Darkness tends to increase with sequence length. I) Dark proteins tend to be shorter than non-dark.

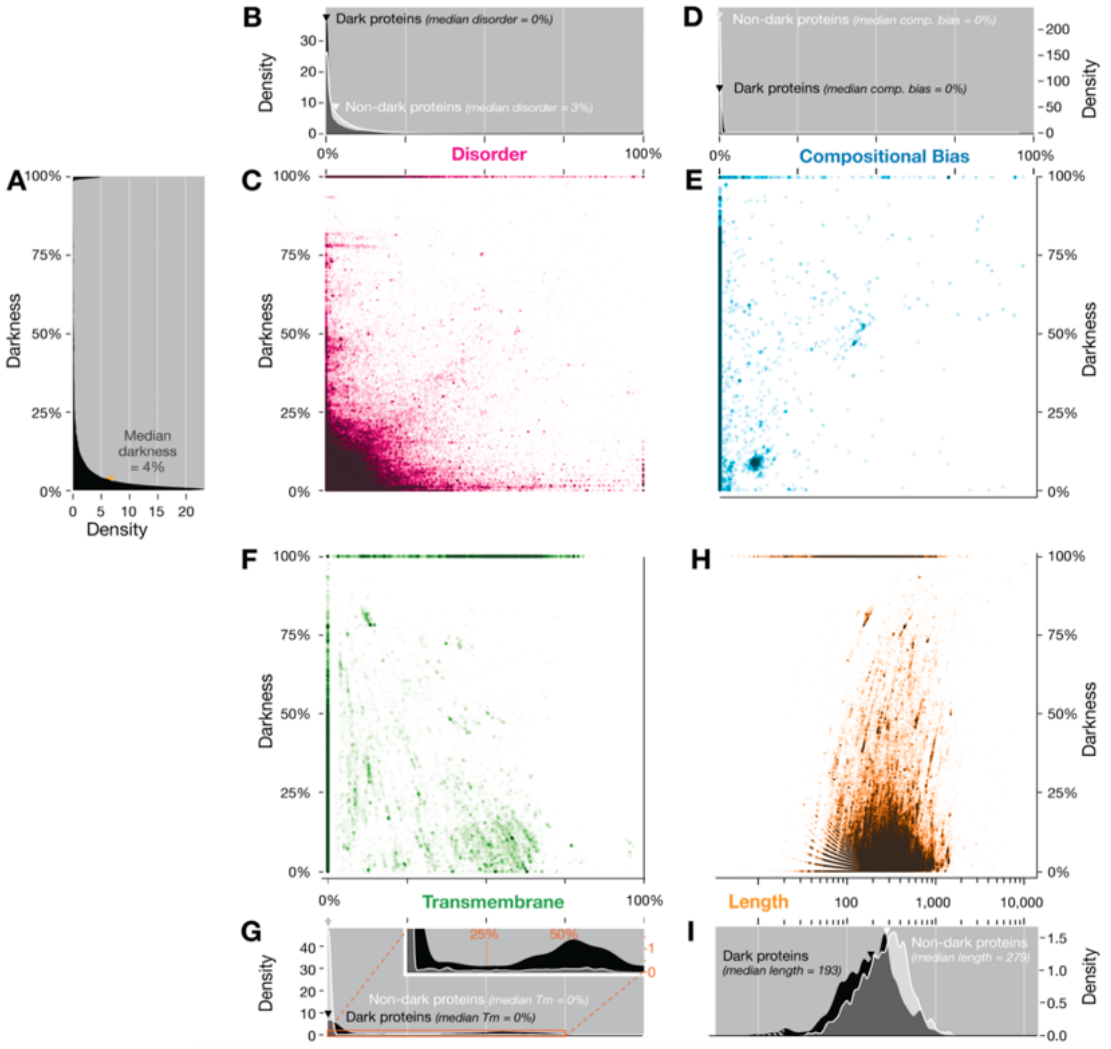


Figure 6.8: Darkness of 331,559 bacterial proteins compared to other properties. A) The distribution of darkness is strongly bimodal; at least 50% of proteins have ≤ 4% darkness, while 7% (23,540) have 100% darkness. B) The distribution of disorder shows that - surprisingly - non-dark proteins have slightly more disorder (3% median) than dark proteins (0% median). C) There is a tendency for darkness to increase with disorder, but it is very slight. D) Almost all dark proteins have very low compositional bias. E) There is a tendency for darkness to increase with compositional bias, but it is only slight. F) As with eukaryotes, multi-pass transmembrane proteins become unexpectedly rare at more than 25% darkness, and as the percentage of transmembrane residues increases, darkness tends to decrease. G) Dark proteins are much more prevalent amongst multi-pass transmembrane proteins (zoomed-in region); however, most dark proteins have no transmembrane residues. H) Darkness tends to increase with sequence length. I) Dark proteins tend to be slightly shorter than non-dark.

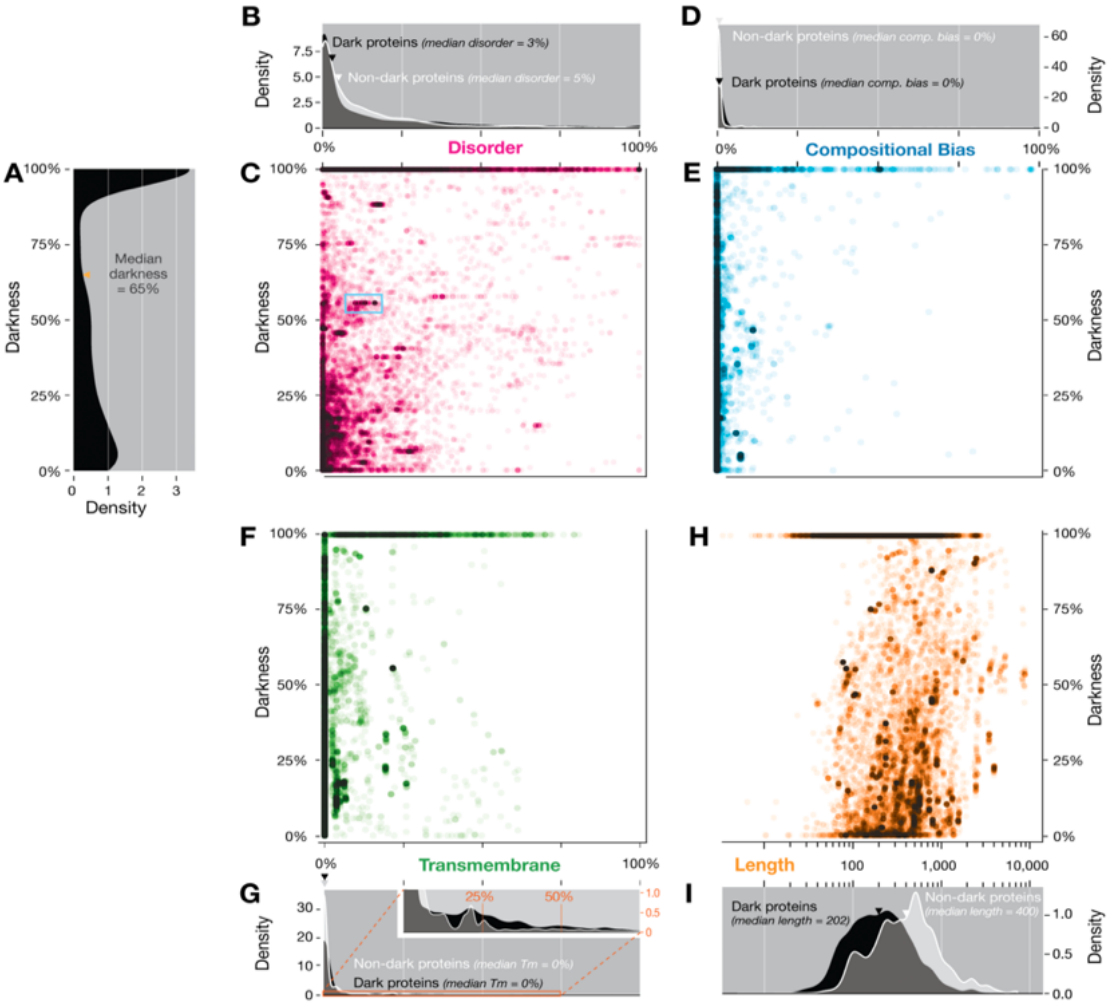


Figure 6.9: Darkness of 16,479 viral proteins compared to other properties. A) The distribution of darkness is much more evenly distributed than in archaea, bacteria, or eukaryotes, with 44% of proteins (7,316) having 100% darkness, and with an overall median of 65% darkness. B) Surprisingly, the distribution of disorder is almost identical between dark and non-dark proteins. C) There is a tendency for darkness to increase with disorder, but it is very slight. The blue rectangle indicates a striking feature that reoccurs several times on the plot, namely groups of similar viral protein groups regularly spaced in the horizontal direction. These groups mostly consist of proteins from different strains of the same virus – the key difference seen from one strain to the other is in the number of disordered residues, accounting for the regular horizontal spacing. This is consistent with the observation that the addition of disordered regions is a key aspect of viral strategies to hijack cell regulation (Davey et al., 2011). D) Almost all dark proteins have very low compositional bias. E) There is a tendency for darkness to increase with compositional bias, but it is only slight. F) Multi-pass transmembrane proteins are quite rare, and as the percentage of transmembrane residues increases, darkness tends to decrease. G) The very few multi-pass transmembrane proteins present seem fairly similarly distributed between dark and non-dark proteins. H) Darkness tends to increase with sequence length. I) Dark proteins tend to be slightly shorter than non-dark.

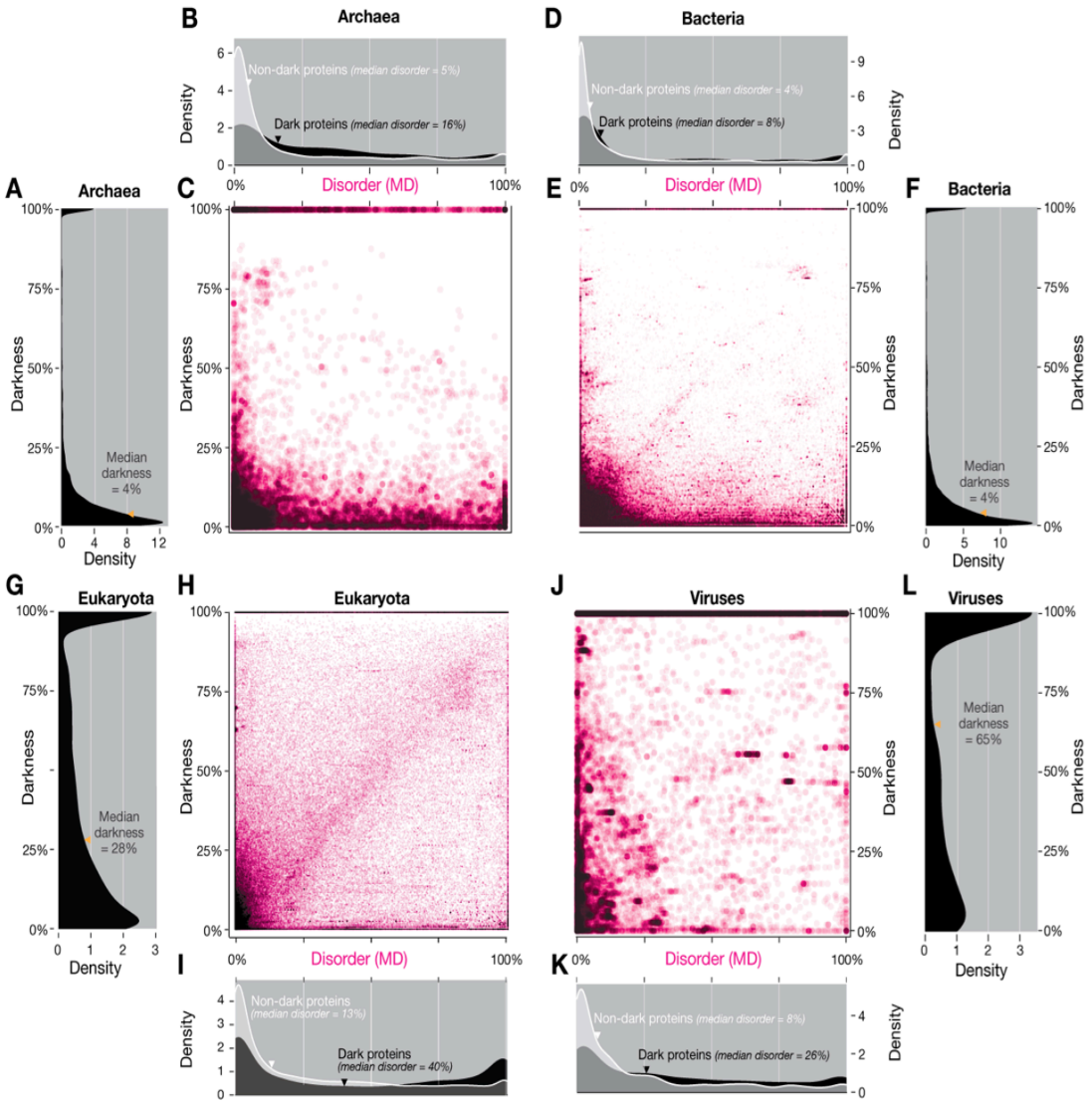


Figure 6.10: Comparing darkness with disorder defined using MD. Overall, the results are mostly similar to those obtained using only IUPred (Figs. 6.6C, 6.7C, 6.8C, and 6.9C). For eukaryotes, however, using MD results in a larger fraction of proteins occur close to the diagonal, resulting in an approximately linear relationship between disorder and darkness (H), in contrast to the upper triangular region seen with IUPred (Fig. 6.6C). However, as previously, most proteins do not show this trend. Indeed, the presence of almost as many proteins below this region as above indicates that disorder is essentially unrelated to darkness. For viruses (J), the pattern associated with disordered linear motifs is even more pronounced (Fig. 6.9C). The density plots (B, D, I, and K) show that MD disorder is more evenly distributed than IUPred disorder (Figs. 6.6D, 6.7D, 6.8D, and 6.9D, respectively).

Dark Proteins Mostly 'Unknown Unknowns'. To determine the fraction of dark proteins that could be accounted for by a combination of disorder, transmembrane regions, or compositional bias, we categorized each protein as having either a 'high' (≥ 25%) or 'low' (< 25%) value for each score (Fig. 6.11). Most of the 'known unknown' (colored fraction) is accounted for by disorder in eukaryotes and viruses, and by transmembrane regions in bacteria and archaea (consistent with Figs. 6.6G, 6.7G, 6.8G, 6.9G and Fig. 6.14). However, a surprisingly large fraction of dark proteins (45-70%) are 'unknown unknowns' (grey fraction) in that they cannot be easily accounted for by these conventional explanations (Fig. 6.11). This fraction was largest for viral dark proteins, possibly due to their rapid mutation rates (Drake et al., 1998), which would tend to increase darkness by undermining the sequence-based structure prediction used in this thesis (O'Donoghue et al., 2015; Haas et al., 2013). To further characterize 'unknown' dark proteins, we next compared them to non-dark proteins that were also ordered, globular, and had low compositional bias (i.e., grey fraction, Fig. 6.12).

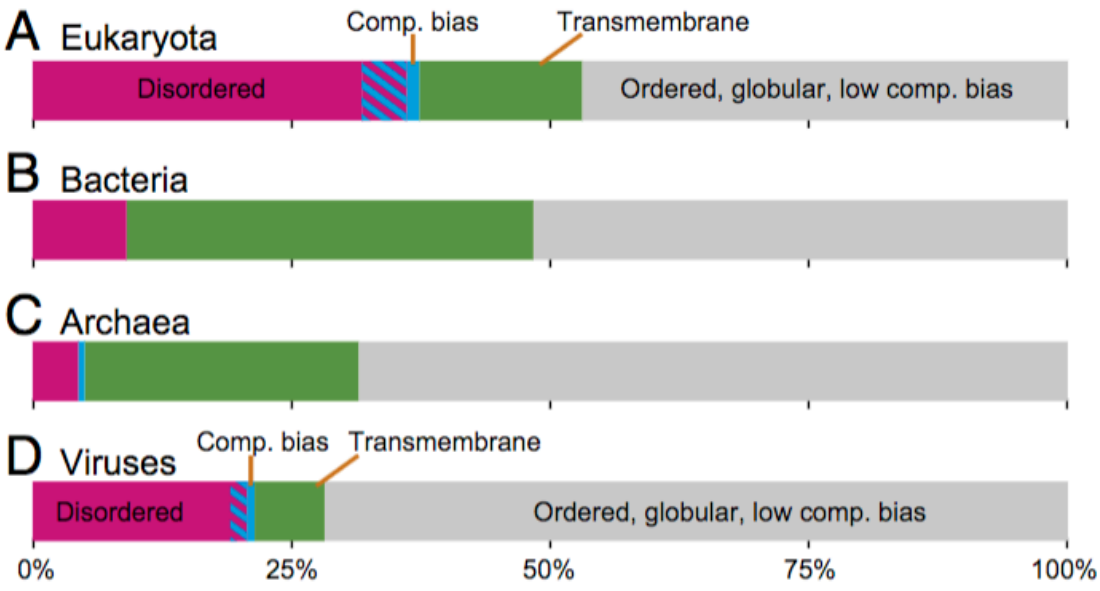


Figure 6.11: Known vs. unknown dark proteins using PSSH2. Each linear diagram (Tringe & Rubin, 2005) shows 'known' dark proteins, i.e., those with ≥ 25% of residues disordered (magenta), compositionally biased (blue), transmembrane (green), or both disordered and compositionally biased (stripes). The remaining fraction (grey) are 'unknown unknowns' – i.e., dark proteins predominately ordered, globular, and low in compositional bias. A) In eukaryotes, high disorder accounted for most of the 'known' dark proteins. Most dark proteins with high compositional bias were also highly disordered. B and C) In bacteria and archaea, highly transmembrane proteins accounted for most of the 'known' dark proteins (consistent with Figs. 6.7G, 6.8G and Fig. 6.14). D) Viruses had the largest 'unknown unknown' fraction and, like eukaryotes, had a large fraction of highly disordered dark proteins (Perdigão et al., 2015).

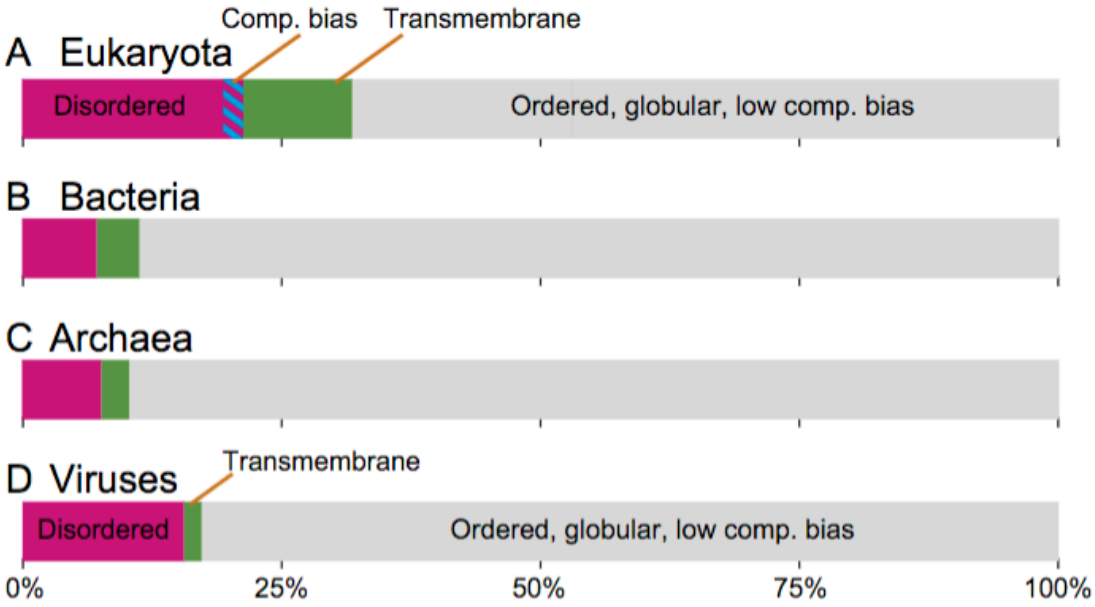


Figure 6.12: Known vs. unknown non-dark proteins using PSSH2. Disorder, compositional bias, and transmembrane fraction for non-dark proteins. Each linear diagram shows the fraction of non-dark proteins with ≥ 25% of residues disordered (magenta), compositionally biased (blue), transmembrane (green), or both disordered and compositionally biased (stripes). The remaining fractions (gray) are non-dark proteins predominately ordered, globular, and low in compositional bias. The figure shows data from eukaryotes (A), bacteria (B), archaea (C), and viruses (D). Note that in eukaryotic non-dark proteins (A), the difference in gray fraction compared with dark proteins (Fig. 6.11A) is smaller than may be expected

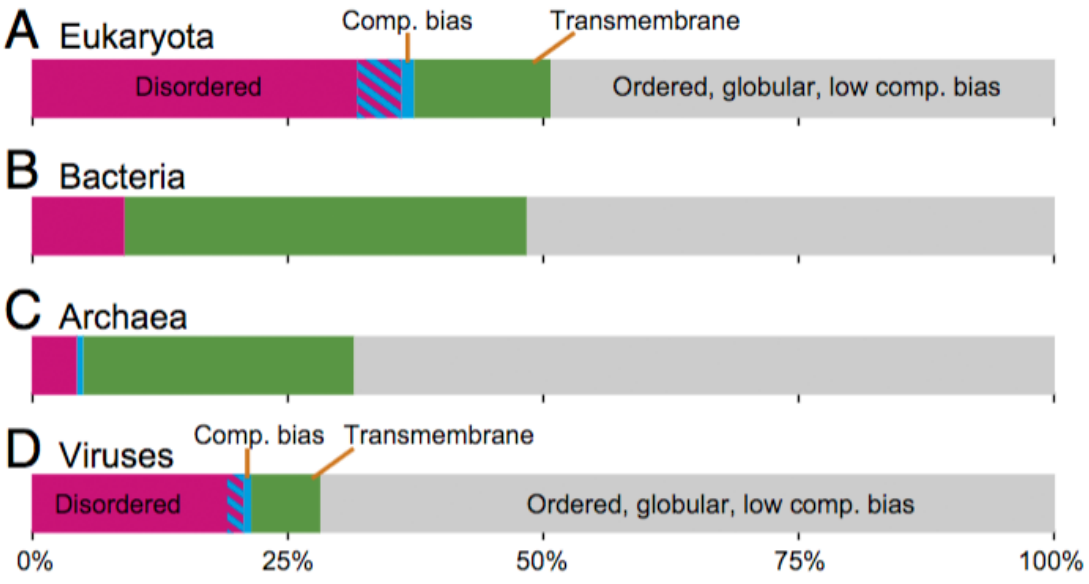


Figure 6.13: Known vs. unknown dark proteins using PSSH2 and PMP. Similar distributions are plotted as for Figure 6.11, but now defining dark residues to be those with no matching structures in either PSSH2 or PMP. Although this definition of 'darkness' is more stringent, the overall fractions for dark regions and dark proteins are only slightly reduced. (Perdigão et al., 2015).

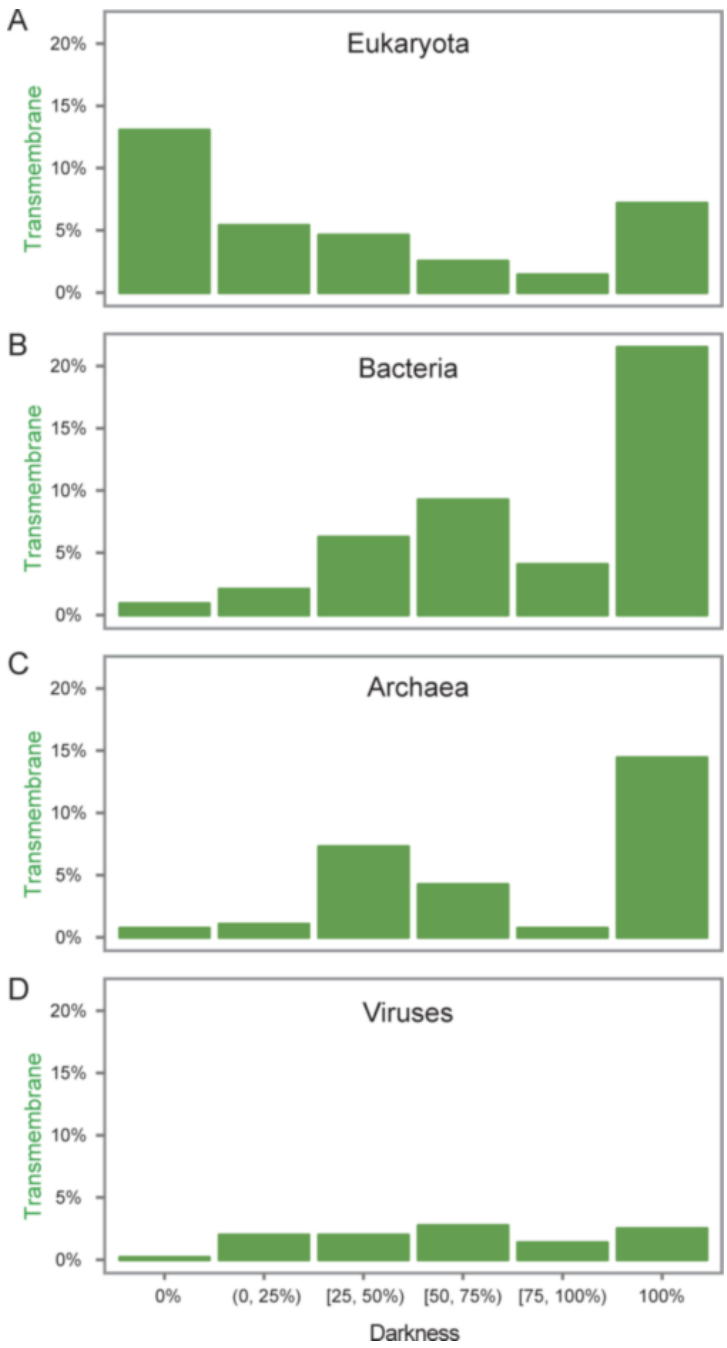


Figure 6.14: Darkness vs. transmembrane fraction. In each histogram, proteins have been binned into six groups according to their darkness score (darkness = 0%, 0% < darkness < 25%, 25% ≤ darkness < 50%, 50% ≤ darkness < 75%, 75% ≤ darkness < 100%, and darkness = 100%). We then calculated the average fraction of transmembrane residues across all proteins in each bin. A) Surprisingly, for eukaryotic proteins, the largest fraction of transmembrane residues was seen for proteins with 0% darkness, and the fraction tended to decrease with increasing darkness, although rising somewhat for dark proteins (100% darkness). B) Bacterial proteins show nearly the opposite behavior: the smallest fraction of transmembrane residues was seen for proteins with 0% darkness, and the largest for proteins with 100% darkness. Interestingly, however, there was a dip in transmembrane fraction for proteins with 75% ≤ darkness < 100%. C) Archaeal proteins show a similar overall pattern to bacteria: the transmembrane fraction tended to increase with increasing darkness, although there as a dip in 3 transmembrane fraction for proteins with 50% ≤ darkness < 100%. D) Overall, viral proteins have much lower transmembrane fraction and relatively little dependency on darkness (Perdigão et al., 2015).

Subcellular Location of Dark Proteins. For each protein I used UniProt annotations to determine its subcellular location; these data were missing for 44% for eukaryotic dark proteins compared to 29% of non-dark proteins (consistent with lower evolutionary re-use - since location is often inferred via homology) (Perdigão et al., 2015). The subset of proteins with known location was used in an enrichment analysis (see Methods) finding that, unexpectedly, eukaryotic dark proteins were most strongly over- represented in the extracellular space followed by the endoplasmic reticulum (Fig. 6.15B). This partly explains why dark proteins had few interactions (Chapter 7) since secreted proteins are often 'autonomous' compared with intracellular proteins, fulfilling their functions via fewer interactions with other proteins. Interestingly, the only subcellular location where dark proteins were under- represented was the cytoplasm (Fig. 6.15B), and the only tissue where they were under-represented was cytoplasmic-rich red blood cells (Fig. 6.15); this suggests that knowledge of cytoplasmic protein structures approaches a level of completeness – similar to that of bacterial and archaeal proteins (Fig. 6.1), most of which are also cytoplasmic.

Functions of Dark Proteins. For each protein we extracted functional descriptions from the UniProt 'CC' annotations; the median length of text in this field was 47% shorter for dark proteins, indicating that less is known about them (Fig. 6.15A) (again, consistent with lower evolutionary re-use) (Perdigão et al., 2015). The resulting set of 242,064 distinct functional annotation terms was used in an enrichment analysis (see Methods), finding that only 2,098 were under-represented in dark proteins, while 3,566 were over-represented (Tables 6.1, 6.2, 6.3, 6.4, and 6.5). This implies that, overall, dark proteins fulfill a wide variety of functions but, nevertheless, a subset has distinct biological functions.

Eukaryotic dark proteins were over-represented in specific secretory tissues and exterior environments (Fig. 6.15B), consistent with the result that many were secreted (Fig. 6.15C). They were also over-represented in disulfide-rich domains and in disulfide bonds (Fig. 6.15C; Table 6.3). Additionally, they were over-represented in cleavage and other posttranslational modifications known to prepare proteins for harsh environments and to confound experimental structure determination (Fig. 6.15C).

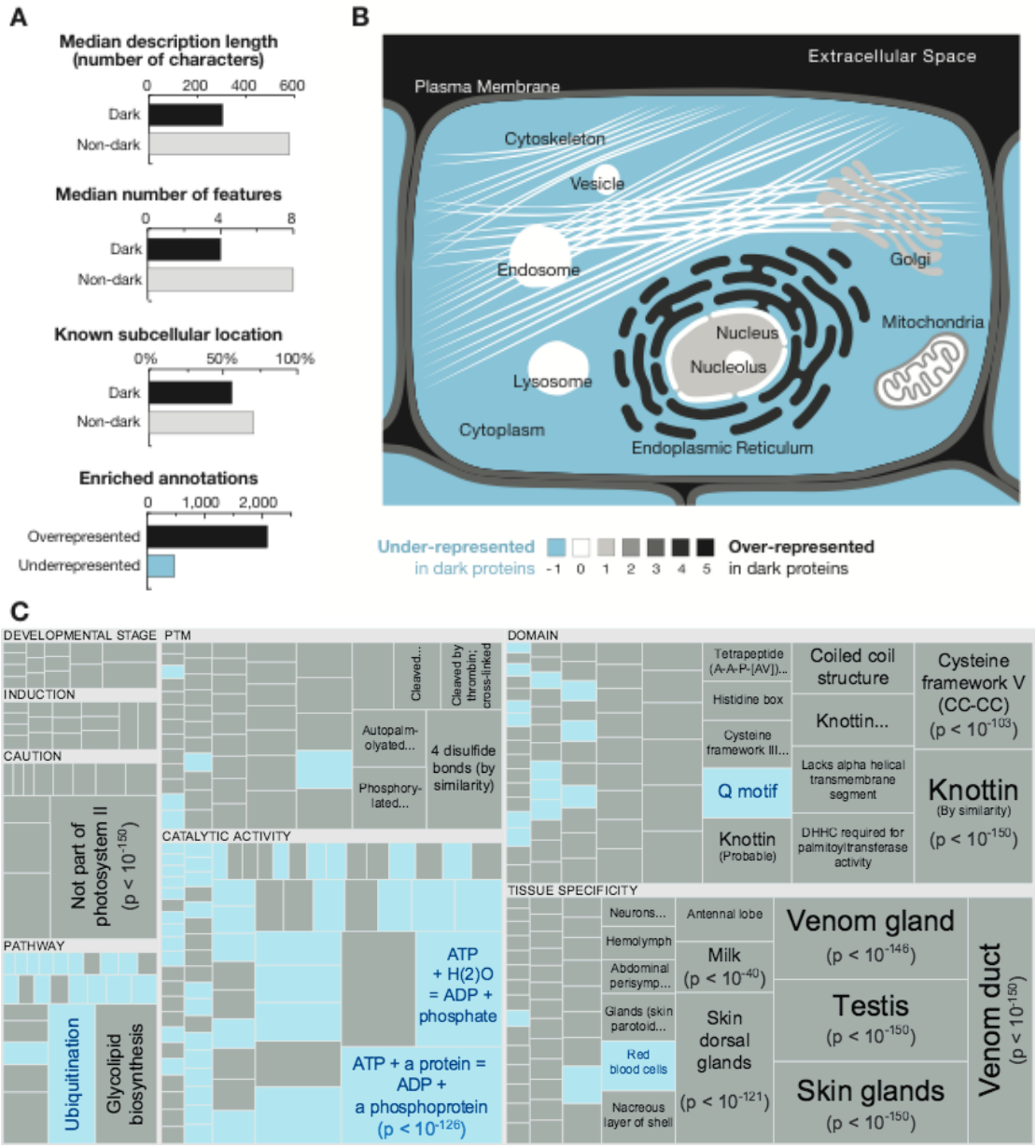


Figure 6.15: TreeMap (Perdigão et al., 2015) for Dark vs non-Dark proteins in Eukaryotes. A) Dark proteins have shorter text describing their function, fewer sequence-specific features, and less complete annotation of subcellular location. Enrichment analysis of dark proteins found four times more over-represented annotations than under-represented. B) Shows cellular regions under- or over-represented in dark proteins. C) TreeMap showing under- (blue) or over-represented annotations (black); the area of each cell is proportional to $-\log_{10}(p_j)$, where p_j is the probability associated with the annotation in the cell. Dark proteins are under-represented only in the 'Catalytic site' and 'Pathway' subcategories, where annotations generally require similarity to a PDB structure. Complete enrichment results are in Table 6.3 (Perdigão et al., 2015).

6.5.1. The Human Dark Proteome

In this thesis it was specially studied the human proteome, finding that over half of it was dark (Fig. 6.16A). Adding that less was known about the function and subcellular location of dark proteins, 56% shorter 'CC' field; missing location data for 56% compared with 22% for non-dark proteins (Fig. 6.16B). Where these data were available, we saw again that dark proteins were associated with secretion, transmembrane regions, and cleavage; in addition, we saw some association with cancer and endogenous retroviral proteins (Fig. 6.17; Table 6.5).

It was also determined which dark proteins came from sequential genes, finding seven 'dark' gene clusters. So basically you can take each protein and mapped down to the gene where the protein comes from and mapping down to chromosomes, and if we do that proteins from these clusters had many features described above as typical for dark proteins (Fig. 6.17; Table 6.6).

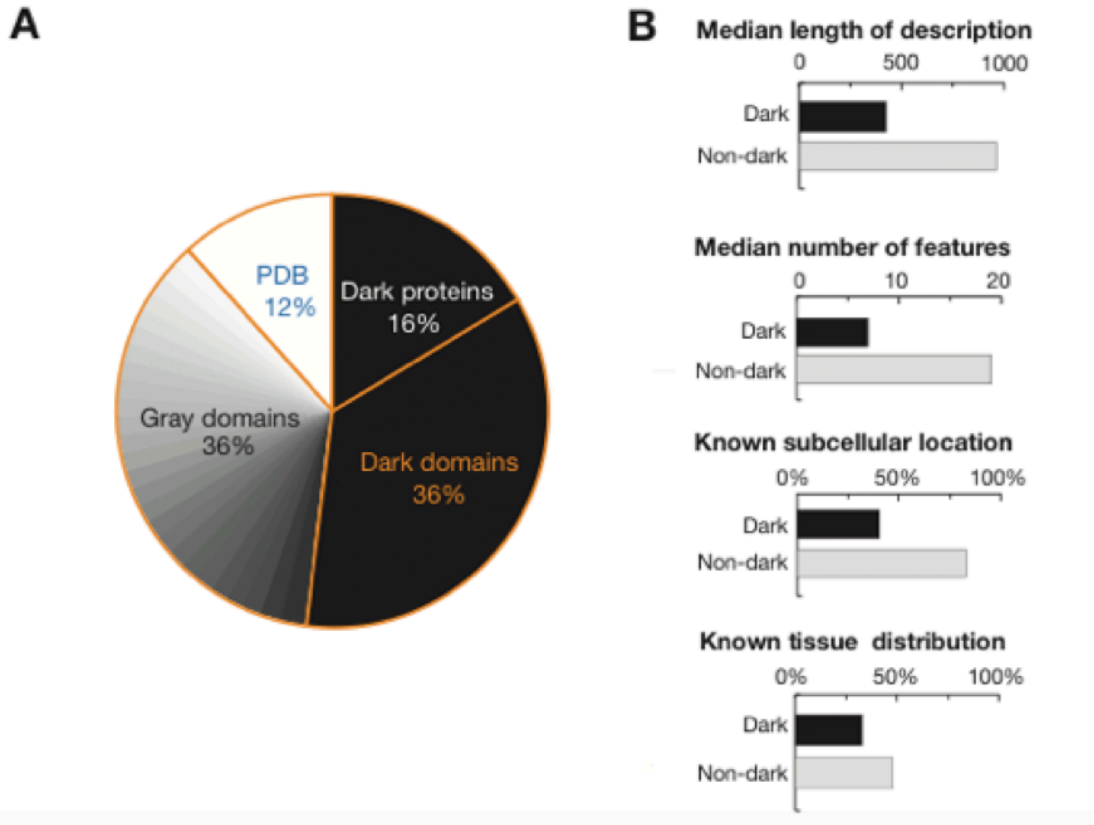


Figure 6.16: Dark vs non-dark proteins in human. A) Shows the fractions of amino acids across all 20,209 human proteins assigned to PDB domains, grey domains, dark regions, and 4,382 dark proteins. B) Dark proteins have shorter functional descriptions, fewer sequence-specific features, and less complete annotation about subcellular location and tissue distribution.

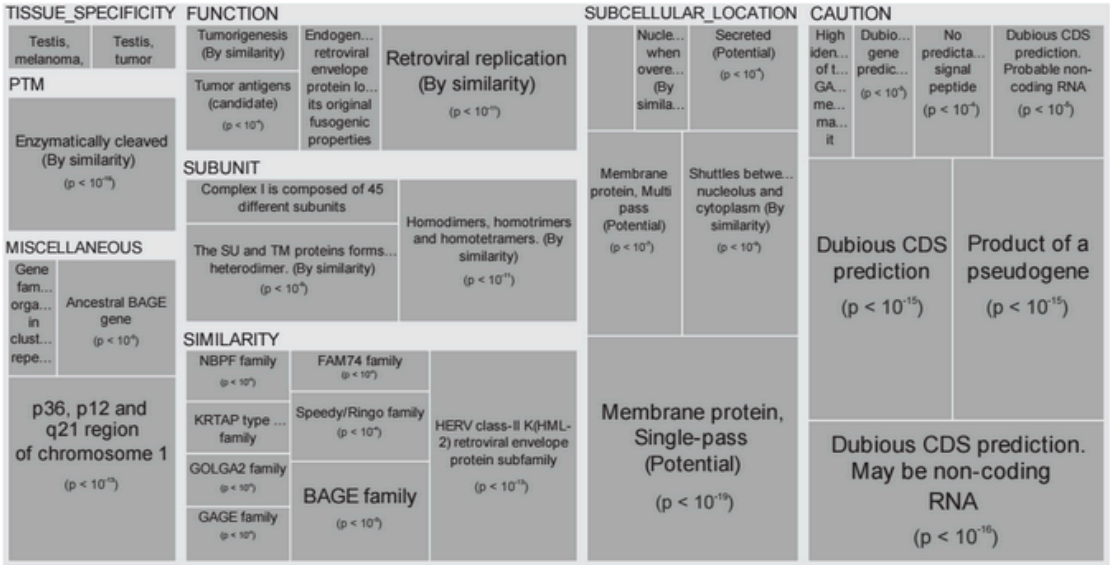


Figure 6.17: TreeMap (Perdigão et al., 2015) showing all annotations over-represented in dark proteins (details are in Table 6.5). 'Caution' annotations seen for 215 dark 'proteins' indicate they may be long non-coding RNA or arise from pseudogenes; further tests suggest only 14 are non-coding.

Finally, using Proteomics DB (Wilhelm et al., 2014) I looked at all the proteins that were expressed in 70 tissues, where every tissue has a list of expressed proteins and a level of abundance. What I have done was inspecting each tissue (like for instance, the brain), and observe which proteins where highly expressed and which fraction of those proteins were dark, associated with a darkness value, not for each protein, but for each tissue (Table 6.7). The tissue that has the highest level of darkness is the heart, which is very interesting, since it's the tissue that is associated with heart disease, one of the main cause of death in humans (Table 6.7).

6.6. Discussion

Mapping the dark proteome has revealed many unexpected features; however, more analyses remain to be done – for example, examining physiochemical properties also known to confound structure determination, e.g., isoelectric point, hydrophobicity, or irregular secondary structure (Slabinski et al., 2007). Thus, we provide our data for use by others (Tables 6.1, 6.2, 6.3, 6.4, and 6.5; full tables are online at the following URL: http://www.pnas.org/content/suppl/2015/11/17/1508380112.DCSupplemental/pnas.1508380112.sd01.xlsx).

Several insights can be gained from the dark protein features revealed in this thesis:

(a) The observation that darkest proteins had low disorder (and many highly disordered proteins are not dark) helps clarify the distinction between darkness and disorder; this in turn will help further studies into protein intrinsic disorder;

(b) The observation that transmembrane regions were rare amongst proteins with 75% ≤ darkness < 100% (especially in eukaryotes) may indicate the existence of transmembrane regions undetected by current prediction methods;

(c) The observation that many dark proteins are secreted and post-translationally modified may help focus on the development of experimental and bioinformatics methods to better manage such cases.

Mostly, however, dark proteins are a mystery; in addition to unknown structure, many have unknown location, unknown function, and no known interactions with other proteins (Chapter 7). This is partly accounted for by low evolutionary re-use, since annotation is often inferred by homology; it is also partly accounted for by expression in specific tissues and developmental stages. Ultimately, many dark proteins are simply not as well studied as non-dark proteins; this work will contribute by highlighting them for subsequent experimental and bioinformatics studies, which may reveal further 'unknown unknowns'.

The dark proteome is a moving target, changing as the PDB grows. However, as sequence databases grow at much faster rates, will the dark proteome expand or contract? The current work cannot answer this directly, but previous surveys have concluded that the number of folds is ≲ 10,000 (Koonin et al, 2002), suggesting that the dark proteome will eventually contract if improvements in detection methods (e.g., HHblits (Remmert et al., 2011)) keep pace with the rate of new sequence families. However, those surveys used databases (PDB, Swiss-Prot, etc.) with historical bias towards model organisms; newer experimental approaches are reducing this bias (e.g., structural genomics (Khafizov et al., 2014); DNA sequencing of environmental samples (Tringe & Rubin, 2005). A recent survey of 8 million protein sequences by Levitt (Levitt, 2009) concluded that eventually the number of folds may increase linearly with sequences – although uncertainty in this conclusion arose since ~22% of the proteins surveyed were 'uncharacterized' (i.e., orphans not matching any known sequence family) – many may be due to errors in predicting genes from whole genomes.

Archaea

Non-dark	Dark	Ratio	Total	Fisher's p value	Adjusted p value	Annotation Sub-Category	Annotation
222	188	24.14960798	410	4.60E-162	3.10E-158	SUBCELLULAR_LOCATION	Cell membrane; Multi-pass membrane protein (Potential).
11	76	197.0271691	87	1.08E-99	3.64E-96	CATALYTIC_ACTIVITY	5-methyl-5,6,7,8-tetrahydromethanopterin + 2-mercaptoethanesulfonate = 5,6,7,8-tetrahydromethanopterin + 2-(methylthio)ethanesulfonate.
11	74	191.8422436	85	7.66E-97	1.72E-93	PATHWAY	One-carbon metabolism; methanogenesis from CO(2); methyl-coenzyme M from 5,10-methylene-5,6,7,8-tetrahydromethanopterin: step 2/2.
0	61	0	61	9.68E-91	1.63E-87	FUNCTION	Part of a complex that catalyzes the formation of methyl-coenzyme M and tetrahydromethanopterin from coenzyme M and methyl-tetrahydromethanopterin. This is an energy-conserving, sodium-ion translocating step (By similarity).
10	68	193.9162138	78	3.70E-89	4.98E-86	SUBUNIT	The complex is composed of 8 subunits; MtrA, MtrB, MtrC, MtrD, MtrE, MtrF, MtrG and MtrH (By similarity).
0	58	0	58	2.69E-86	3.01E-83	SIMILARITY	Belongs to the UPF0218 family.
134	97	20.64296833	231	3.99E-79	3.84E-76	SUBCELLULAR_LOCATION	Cell membrane; Multi-pass membrane protein (By similarity).
0	42	0	42	1.25E-62	1.05E-59	SIMILARITY	Belongs to the UPF0179 family.
31	53	48.7550253	84	2.29E-56	1.71E-53	SUBCELLULAR_LOCATION	Membrane; Single-pass membrane protein (Potential).
3627	1	0.007862446	3628	2.95E-54	1.99E-51	SUBCELLULAR_LOCATION	Cytoplasm (By similarity).
0	34	0	34	8.20E-51	5.02E-48	SIMILARITY	Belongs to the archaeal flagellin family.
0	32	0	32	7.35E-48	4.12E-45	SIMILARITY	Belongs to the UPF0212 family.
0	29	0	29	1.96E-43	1.02E-40	SUBCELLULAR_LOCATION	Archaeal flagellum.
0	29	0	29	1.96E-43	9.44E-41	FUNCTION	Flagellin is the subunit protein which polymerizes to form the filaments of archaeal flagella.
0	28	0	28	5.86E-42	2.63E-39	FUNCTION	Important for reducing fluoride concentration in the cell, thus reducing its toxicity (By similarity).
0	28	0	28	5.86E-42	2.47E-39	SIMILARITY	Belongs to the CrcB (TC 9.B.71) family.
0	28	0	28	5.86E-42	2.32E-39	SIMILARITY	Belongs to the UPF0248 family.
0	26	0	26	5.23E-39	1.95E-36	CATALYTIC_ACTIVITY	GDP-cobinamide + alpha-ribazole = cobalamin + GMP.
0	26	0	26	5.23E-39	1.85E-36	PATHWAY	Cofactor biosynthesis; adenosylcobalamin biosynthesis; adenosylcobalamin from cob(II)yrinate a,c-diamide: step 7/7.
0	26	0	26	5.23E-39	1.76E-36	SIMILARITY	Belongs to the CobS family.

Table 6.1: Annotations enriched in dark proteins from archaea (only the first 20 entries). The table documents Swiss-Prot annotations that are over- or underrepresented in dark proteins. The annotations are derived from the Swiss-Prot 'Description' field, and are divided into 19 subcategories. The values indicated have been calculated using Fisher's exact test, then adjusted via the false discovery rate method. Each row of the table gives information on all proteins where the Swiss-Prot entry contains a match to the annotation specified in the Annotation column. Non-dark indicates the number of matching proteins that are non-dark, while Dark indicates the number that are dark. Ratio indicates the ratio of dark to non-dark, adjusted to account for the total numbers of proteins in both categories – i.e., a ratios above 1 indicate an overrepresentation of dark proteins, values below 1 indicate underrepresentation. Total indicates the sum of dark and non-dark proteins. Fisher's p value indicates the raw p- value calculated using Fisher's exact test. Adjusted p value indicates the p-value after applying a correction for false discovery. Annotation Sub-Category indicates the class of annotation. (http://www.pnas.org/content/suppl/2015/11/17/1508380112.DCSupplemental/pnas.1508380112.sd02.xlsx)

Bacteria

Non-dark	Dark	Ratio	Total	Fisher's p value	Adjusted p value	Annotation Sub-Category	Annotation
0	449	0	449	0	0	CATALYTIC_ACTIVITY	Acyl-phosphate + sn-glycerol 3-phosphate = 1-acyl-sn-glycerol 3-phosphate + phosphate.
0	561	0	561	0	0	FUNCTION	Catalyzes the dephosphorylation of undecaprenyl diphosphate (UPP). Confers resistance to bacitracin (By similarity).
0	483	0	483	0	0	FUNCTION	Transfers the N-acyl diglyceride group on what will become the N-terminal cysteine of membrane lipoproteins (By similarity).
2	563	6717.118666	565	0	0	CATALYTIC_ACTIVITY	Ditrans,octacis-undecaprenyl diphosphate + H(2)O = ditrans,octacis-undecaprenyl phosphate + phosphate.
3	453	3603.143583	456	0	0	SIMILARITY	Belongs to the CrcB (TC 9.B.71) family.
0	485	0	485	0	0	PATHWAY	Protein modification; lipoprotein biosynthesis (diacylglyceryl transfer).
0	328	0	328	0	0	FUNCTION	Assembles around the rod to form the L-ring and probably protects the motor/basal body from shearing forces during rotation (By similarity).
0	466	0	466	0	0	SUBUNIT	Probably interacts with PlsX (By similarity).
33680	62	0.04392626	33742	0	0	SUBUNIT	Homodimer (By similarity).
0	411	0	411	0	0	SIMILARITY	Belongs to the peptidase A8 family.
0	309	0	309	0	0	SIMILARITY	Belongs to the UPF0246 family.
7440	3556	11.40495138	10996	0	0	SUBCELLULAR_LOCATION	Cell inner membrane; Multi-pass membrane protein (By similarity).
1975	1970	23.80146821	3945	0	0	SUBCELLULAR_LOCATION	Cell membrane; Multi-pass membrane protein (Potential).
0	331	0	331	0	0	SUBUNIT	The basal body constitutes a major portion of the flagellar organelle and consists of four rings (L,P,S, and M) mounted on a central rod (By similarity).
0	263	0	263	0	0	SIMILARITY	Belongs to the UPF0178 family.
0	247	0	247	0	0	SIMILARITY	Belongs to the UPF0061 (SELO) family.
0	405	0	405	0	0	PATHWAY	Protein modification; lipoprotein biosynthesis (signal peptide cleavage).
3	453	3603.143583	456	0	0	FUNCTION	Important for reducing fluoride concentration in the cell, thus reducing its toxicity (By similarity).
0	448	0	448	0	0	FUNCTION	Catalyzes the transfer of an acyl group from acyl-phosphate (acyl-PO(4)) to glycerol-3-phosphate (G3P) to form lysophosphatidic acid (LPA). This enzyme utilizes acyl-phosphate as fatty acyl donor, but not acyl-CoA or acyl-ACP (By similarity).
0	466	0	466	0	0	SIMILARITY	Belongs to the PlsY family.

Table 6.2: Annotations enriched in dark proteins from bacteria (only the first 20 entries). The table documents Swiss-Prot annotations that are over- or under-represented in dark proteins. (http://www.pnas.org/content/suppl/2015/11/17/1508380112.DCSupplemental/pnas.1508380112.sd02.xlsx)

Eukaryota

Non-dark	Dark	Ratio	Total	Fisher's p value	Adjusted p value	Annotation Sub-Category	Annotation
0	363	0	363	0	0	SIMILARITY	Belongs to the Casparian strip membrane proteins (CASP) family.
6886	3518	4.220125639	10404	0	0	SUBCELLULAR_LOCATION	Secreted.
166	498	24.78094242	664	0	0	TISSUE_SPECIFICITY	Expressed by the skin glands.
1588	1154	6.002772367	2742	0	0	SUBCELLULAR_LOCATION	Membrane; Multi-pass membrane protein (Potential).
222	802	29.84131505	1024	0	0	TISSUE_SPECIFICITY	Expressed by the venom duct.
0	361	0	361	0	0	SUBUNIT	Homodimer and heterodimers (By similarity).
516	591	9.460941197	1107	1.07E-267	3.58E-263	SUBCELLULAR_LOCATION	Cell membrane; Multi-pass membrane protein (By similarity).
0	257	0	257	2.65E-249	7.77E-245	SIMILARITY	Belongs to the periviscerokinin family.
24	267	91.89599482	291	3.02E-225	7.86E-221	DOMAIN	The presence of a 'disulfide through disulfide knot' structurally defines this protein as a knottin (By similarity).
4120	1443	2.89311488	5563	8.77E-221	2.05E-216	SUBCELLULAR_LOCATION	Secreted (By similarity).
0	202	0	202	4.44E-196	9.44E-192	FUNCTION	Mediates visceral muscle contractile activity (myotropic activity).
0	196	0	196	2.83E-190	5.53E-186	SIMILARITY	Belongs to the PsbN family.
0	195	0	195	2.63E-189	4.74E-185	CAUTION	Based on experiments in Thermosynechococcus vulcanus this is probably not a component of photosystem II.
53	250	38.96374595	303	2.36E-185	3.94E-181	TISSUE_SPECIFICITY	Testis.
482	455	7.797599449	937	1.86E-184	2.90E-180	SUBCELLULAR_LOCATION	Endoplasmic reticulum membrane; Multi-pass membrane protein (By similarity).
0	174	0	174	5.46E-169	7.99E-165	SIMILARITY	Belongs to the pyrokinin family.
36	203	46.57899363	239	5.65E-156	7.78E-152	SIMILARITY	Belongs to the conotoxin O1 superfamily.
3539	1133	2.644514248	4672	2.67E-150	3.47E-146	TISSUE_SPECIFICITY	Expressed by the venom gland.
0	153	0	153	1.13E-148	1.39E-144	SIMILARITY	Belongs to the FARP (FMRFamide related peptide) family.
1	154	1272.088378	155	1.68E-147	1.97E-143	SIMILARITY	Belongs to the protamine P1 family.

Table 6.3: Annotations enriched in dark proteins from eukaryotes (only the first 20 entries). (http://www.pnas.org/content/suppl/2015/11/17/1508380112.DCSupplemental/pnas.1508380112.sd02.xlsx)

Viruses

Non-dark	Dark	Ratio	Total	Fisher's p value	Adjusted p value	Annotation Sub-Category	Annotation
40	221	15.49995482	261	2.45E-87	2.90E-83	SUBCELLULAR_LOCATION	Host membrane; Single-pass membrane protein (Potential).
45	154	9.600776439	199	3.40E-51	2.01E-47	FUNCTION	Plays a role in virus cell tropism, and may be required for efficient virus replication in macrophages (By similarity).
1	68	190.7686747	69	1.56E-38	6.14E-35	PTM	Phosphorylated (By similarity).
47	122	7.282158421	169	1.37E-35	4.04E-32	SUBCELLULAR_LOCATION	Host membrane; Multi-pass membrane protein (Potential).
6	72	33.66506024	78	6.13E-35	1.45E-31	SUBUNIT	Interacts with major capsid protein L1 (By similarity). Interacts with host importins (By similarity).
7	72	28.85576592	79	5.13E-34	7.57E-31	FUNCTION	Minor protein of the capsid that localizes along the inner surface of the virion, within the central cavities beneath the L1 pentamers.. …. (By similarity).
7	72	28.85576592	79	5.13E-34	8.66E-31	SIMILARITY	Belongs to the papillomaviridae L2 protein family.
0	57	0	57	7.63E-34	1.00E-30	SIMILARITY	Belongs to the asfivirus MGF 360 family.
7	72	28.85576592	79	5.13E-34	1.01E-30	PTM	Highly phosphorylated (Potential).
0	53	0	53	1.62E-31	1.91E-28	SUBUNIT	Homomultimer. Interacts with envelope E protein in the budding compartment of the host cell, which is located between endoplasmic reticulum and the Golgi complex. ... This interaction probably participates in RNA packaging into the virus (By similarity).
1	52	145.8819277	53	2.43E-29	2.61E-26	SIMILARITY	Belongs to the coronaviruses M protein family.
0	48	0	48	1.31E-28	1.29E-25	SIMILARITY	Belongs to the asfivirus MGF 110 family.
209	0	0	209	4.08E-28	3.71E-25	SIMILARITY	Belongs to the influenza viruses hemagglutinin family.
2	52	72.94096386	54	4.88E-28	4.11E-25	SIMILARITY	Belongs to the orthohepadnavirus protein X family.
207	0	0	207	6.13E-28	4.83E-25	PTM	In natural infection, inactive HA is matured into HA1 and HA2 outside the cell by one or more trypsin-like, arginine-specific endoprotease secreted by the bronchial epithelial cells. ….(By similarity).
0	46	0	46	1.90E-27	1.40E-24	SUBCELLULAR_LOCATION	Host cytoplasm. Note=Found in spherical cytoplasmic structures, called virus factories, that appear early after infection and are the site of viral replication and packaging (By similarity).
1	47	131.8548193	48	1.78E-26	1.24E-23	FUNCTION	Component of the viral envelope that plays a central role in virus morphogenesis and assembly via its interactions with other viral proteins (By similarity).
1	45	126.2439759	46	2.48E-25	1.63E-22	SUBCELLULAR_LOCATION	Virion membrane; Multi-pass membrane protein (Potential). Host Golgi apparatus membrane; Multi-pass membrane protein (Potential). Note=Largely embedded in the lipid bilayer (By similarity).
183	0	0	183	8.36E-25	5.20E-22	CATALYTIC_ACTIVITY	Hydrolysis of alpha-(2->3)-, alpha-(2->6)-, alpha-(2->8)-glycosidic linkages of terminal sialic acid residues in oligosaccharides, glycoproteins, glycolipids, colominic acid and synthetic substrates.
27	78	8.104551539	105	1.27E-24	7.49E-22	SUBCELLULAR_LOCATION	Membrane; Single-pass membrane protein (Potential).

Table 6.4: Annotations enriched in dark proteins from viruses (only the first 20 entries). (http://www.pnas.org/content/suppl/2015/11/17/1508380112.DCSupplemental/pnas.1508380112.sd02.xlsx)

Human

Non-dark	Dark	Ratio	Total	Fisher's p value	Adjusted p value	Annotation Sub-Category	Annotation
124	28	18,485372	152	9,71E-25	6,65E-20	SUBCELLULAR_LOCATION	Membrane; Single-pass membrane protein (Potential).
0	11	0	11	7,55E-22	2,59E-17	CAUTION	Product of a dubious CDS prediction. May be a non-coding RNA.
162	27	13,643965	189	7,11E-21	1,62E-16	CAUTION	Could be the product of a pseudogene.
5	12	196,473096	17	5,29E-20	9,05E-16	CAUTION	Product of a dubious CDS prediction.
0	9	0	9	5,27E-18	6,01E-14	SIMILARITY	Belongs to the beta type-B retroviral envelope protein family. HERV class-II K(HML-2) env subfamily.
0	9	0	9	5,27E-18	7,22E-14	MISCELLANEOUS	Encoded by one of the numerous copies of NBPF genes clustered in the p36, p12 and q21 region of the chromosome 1.
0	8	0	8	4,39E-16	3,76E-12	SUBUNIT	Forms homodimers, homotrimers, and homotetramers via a C-terminal domain. Associates with XPO1 and with ZNF145 (By similarity).
0	8	0	8	4,39E-16	4,30E-12	FUNCTION	Retroviral replication requires the nuclear export and translation of unspliced, singly-spliced and multiply-spliced derivatives of the initial genomic transcript. Rec interacts with a highly structured RNA element (RcRE) present in the viral 3'LTR and recruits the cellular nuclear export machinery. This permits export to the cytoplasm of unspliced genomic or incompletely spliced subgenomic viral transcripts (By similarity).
1	8	654,910321	9	3,91E-15	2,98E-11	PTM	Specific enzymatic cleavages in vivo yield the mature SU and TM proteins (By similarity).
0	7	0	7	3,66E-14	2,51E-10	SUBCELLULAR_LOCATION	Cytoplasm (By similarity). Nucleus, nucleolus (By similarity). Note=Shuttles between the nucleus and the cytoplasm. When in the nucleus, resides in the nucleolus (By similarity).
3	8	218,30344	11	7,02E-14	4,37E-10	SUBUNIT	The surface (SU) and transmembrane (TM) proteins form a heterodimer. SU and TM are attached by noncovalent interactions or by a labile interchain disulfide bond (By similarity).
218	21	7,885961 4	239	2,55E-12	1,46E-08	SUBCELLULAR_LOCATION	Membrane; Multi-pass membrane protein (Potential).
0	5	0	5	2,54E-10	1,16E-06	CAUTION	Product of a dubious CDS prediction. Probable non-coding RNA.
0	5	0	5	2,54E-10	1,24E-06	SIMILARITY	Belongs to the BAGE family.
0	5	0	5	2,54E-10	1,34E-06	MISCELLANEOUS	The ancestral BAGE gene was generated by juxtacentromeric reshuffling of the KMT2C/MLL3 gene. The BAGE family was expanded by juxtacentromeric movement and/or acrocentric exchanges. BAGE family is composed of expressed genes that map to the juxtacentromeric regions of chromosomes 13 and 21 and of unexpressed gene fragments that scattered in the juxtacentromeric regions of several chromosomes, including chromosomes 9, 13, 18 and 21.
215	17	6,472950 85	232	4,87E-09	2,09E-05	SUBCELLULAR_LOCATION	Secreted (Potential).
2	5	204,659475	7	5,22E-09	2,10E-05	FUNCTION	Retroviral envelope proteins mediate receptor recognition and membrane fusion during early infection. Endogenous envelope proteins may have kept, lost or modified their original function during evolution. This endogenous envelope protein has lost its original fusogenic properties.
3	5	136,43965	8	1,38E-08	5,25E-05	SUBUNIT	Complex I is composed of 45 different subunits.
0	4	0	4	2,11E-08	6,88E-05	FUNCTION	Unknown. Candidate gene encoding tumor antigens.
0	4	0	4	2,11E-08	7,23E-05	SIMILARITY	Belongs to the Speedy/Ringo family.

Table 6.5: Annotations enriched in dark proteins from human (only the first 20 entries). (http://www.pnas.org/content/suppl/2015/11/17/1508380112.DCSupplemental/pnas.1508380112.sd02.xlsx)

Gene	Protein	Length	Binds	Bias
Chromosome 1 (q21.3): PQCK-rich, keratinocyte proteins				
LCE5A	Late cornified envelope 5A	118	3	21
CRCT1	Cysteine-rich C-terminal 1	99	2	25
LCE3E	Late cornified envelope 3E	92	2	16
LCE3E	Late cornified envelope 3E	92	2	17
LCE3D	Late cornified envelope 3D	92	2	19
LCE3C	Late cornified envelope 3C	94	4	19
LCE3B	Late cornified envelope 3B	95	2	24
LCE3A	Late cornified envelope 3A	89	1	21
LCE2D	Late cornified envelope 2D	110	1	20
LCE2C	Late cornified envelope 2C	110		21
LCE2B	Late cornified envelope 2B	110		22
LCE2A	Late cornified envelope 2A	106	1	20
LCE4A	Late cornified envelope 4A	99	3	18
KPRP	Keratinocyte proline-rich protein	579	1	20
LCE1F	Late cornified envelope 1F	118		22
LCE1E	Late cornified envelope 1E	118	1	22
LCE1D	Late cornified envelope 1D	114	1	22
LCE1C	Late cornified envelope 1C	118	1	21
LCE1B	Late cornified envelope 1B	118		20
LCE1A	Late cornified envelope 1A	110	1	14
LCE6A	Late cornified envelope 6A	80		17
SMCP	Sperm mitochondria-associated	116		29
SPRR4	Small proline-rich protein 4	79		22
SPRR3	Small proline-rich protein 3	169		29
SPRR1B	Small proline-rich protein 1B	89		38
SPRR2D	Small proline-rich protein 2D	72		38
SPRR2A	Small proline-rich protein 2A	72	1	39
SPRR2B	Small proline-rich protein 2B	72	1	39
SPRR2E	Small proline-rich protein 2E	72		36
SPRR2F	Small proline-rich protein 2F	72		40
SPRR2G	Small proline-rich protein 2G	73		26
LELP1	Late cornified envelope-like	98		21
Chromosome 4 (q13.3): P-rich, mouth and digestive secreted proteins				
CSN1S1	Casein alpha s1	185	2	11
CSN2	Casein beta	226	2	17
STATH	Statherin	62	2	11
HTN3	Histatin 3	51	1	14
HTN1	Histatin 1	57	2	12
C4orf40	Proline-rich protein 27	219		21
ODAM	Odontogenic, ameloblast asssociated	279		15
C4orf7	Follicular dendritic cell secreted	85		19
CSN3	Casein kappa	182	2	16
SMR3B	Salivary gland androgen regulated	79	1	39
MUC7	Mucin 7, secreted	377	4	20
AMTN	Amelotin	209	1	15
AMBN	Enamel matrix protein	447	1	15
IGJ	Immunoglobulin J chain	159	1	9
UTP3	Processome component	479	1	13
Chromosome 11 (q12.1-q12.2): LS-rich, transmembrane complex members				
MS4A3	Member 3	214		13
MS4A2	Member 2, receptor for	244		12
MS4A6A	Member 6A	248	2	14
MS4A4E	Putative member 4E	132	2	11
MS4A4A	Member 4	239	1	11
MS4A6E	Member 6E	147	2	16
MS4A7	Member 7	240	1	15
MS4A5	Member 5	200		13
Chromosome 17 (q21.2): CS-rich, keratin associated proteins				
KRTAP3-3	Keratin associated protein 3-3	98		19
KRTAP3-2	Keratin associated protein 3-2	98		19
KRTAP3-1	Keratin associated protein 3-1	98		18
KRTAP3-1	Keratin associated protein 3-1	98		24
KRTAP1-5	Keratin associated protein 1-5	174		25
KRTAP1-1	Keratin associated protein 1-1	177		27
KRTAP2-1	Keratin associated protein 2-1	128		27
KRTAP2-1	Keratin associated protein 2-1	128		35
KRTAP4-11	Keratin associated protein 4-11	195		37
KRTAP4-12	Keratin associated protein 4-12	201		37
KRTAP4-4	Keratin associated protein 4-4	166		36
KRTAP4-3	Keratin associated protein 4-3	195		35
KRTAP4-2	Keratin associated protein 4-2	136		34
KRTAP4-1	Keratin associated protein 4-1	146		36
KRTAP17-1	Keratin associated protein 17-1	105		19
Chromosome 21 (q22.11): GYSC-rich, keratin-associated proteins				
CLDN17	Claudin 17	224	1	14
CLDN8	Claudin 8	225	1	10
KRTAP24-1	Keratin associated protein 24-1	254	2	19
KRTAP25-1	Keratin associated protein 25-1	102		21
KRTAP26-1	Keratin associated protein 26-1	210	2	18
KRTAP27-1	Keratin associated protein 27-1	207	2	16
KRTAP23-1	Keratin associated protein 23-1	65	2	20
KRTAP13-2	Keratin associated protein 13-6, pseudogene	175		23
KRTAP13-1	Keratin associated protein 13-1	172		23
KRTAP13-3	Keratin associated protein 13-3	172		22
KRTAP13-4	Keratin associated protein 13-4	160		21
KRTAP19-1	Keratin associated protein 19-1	90		42
KRTAP19-2	Keratin associated protein 19-2	52		27
KRTAP19-3	Keratin associated protein 19-3	81		43
KRTAP19-4	Keratin associated protein 19-4	84		27
KRTAP19-5	Keratin associated protein 19-5	72		39
KRTAP19-7	Keratin associated protein 19-7	63	2	33
KRTAP6-2	Keratin associated protein 6-2	62		32
KRTAP6-1	Keratin associated protein 6-1	71		38
KRTAP20-1	Keratin associated protein 20-1	56		36
KRTAP20-2	Keratin associated protein 20-2	65		37
KRTAP20-3	Keratin associated protein 20-3	44	4	25
KRTAP21-1	Keratin associated protein 21-1	79	2	35
KRTAP8-1	Keratin associated protein 8-1	63		24
KRTAP11-1	Keratin associated protein 11-1	163		15
KRTAP19-8	Keratin associated protein 19-8	63	4	35
Chromosome X (p11.23): EPG-rich, GAGE and PAGE family proteins				
GAGE10	G antigen 10	116		17
GAGE12J	G antigen 12J	117		16
GAGE12F	G antigen 6	117		17
GAGE13	G antigen 13	117		17
GAGE2E	G antigen 8	116		17
GAGE2D	G antigen 8	116		16
GAGE2C	G antigen 2C	116		18
GAGE12B	G antigen 12B	117		17
GAGE2A	G antigen 2A	116		17
GAGE1	G antigen 6	139		14
GAGE4	Cancer/testis antigen 4.4	117		17
PAGE1	P antigen family, member 1	146		18
PAGE4	P antigen family, member 4	102		15
Chromosome X (p11.22): EP-rich; contains XAGE family proteins				
XAGE2B	X antigen family, member 2B	111		13
XAGE1B	G antigen member; Cancer/testis antigen 12.1	81		15
SSX7	Synovial sarcoma, X breakpoint 7	188		12
SSX2B	Synovial sarcoma, X breakpoint 2B	188		12
SPANXN5	SPANX family, member N5	72		14
XAGE5	X antigen family, member 5	108		12
XAGE3	X antigen family, member 3	111		15
FAM156A	Family with sequence similarity 156, member B	213		12

Table 6.6: Human gene clusters containing dark proteins. *Length* indicates the number of amino acids; *'Binds'* indicates the number of known binding partners in the same cluster from STRING (Franceschini et al., 2013)(Chapter 7); *'Bias'* indicates the largest single amino acid composition (e.g., a value of '42%' indicates that one amino acid accounts for 42% of the entire sequence) – the most frequently occurring amino acids are given for each cluster (e.g., 'CS-rich' indicates Cys is the most common, followed by Ser). The proteins arising from these gene clusters exhibit typical characteristics of dark proteins: they tend to be short, have few known interactions, have atypical amino acid composition, and are often secreted, transmembrane, or skin-associated. The 1q21.3 cluster arises from gene duplication (Rost et al., 1995); it contains many skin proteins with significant compositional bias. The 4q13.3 cluster does not appear to have been previously characterized; it contains proteins related to the mouth, salivary glands, and secretion, implying that these genes share related functions. The 11q12 cluster arises from gene duplication during vertebrate evolution (Bigelow & Rost, 2006); it contains proteins that all have a 4-pass membrane-spanning region and are components of a multimeric receptor complexes. The 17q21.2 and 21q22.11 clusters have also been previously identified (Cedano et al., 1997; Drake et al., 1998); they contain hair-associated proteins. The Xp11.23 and Xp11.22 clusters are both very recent evolutionary developments (Andrade et al., 1998); they contain proteins that are expressed only in testis and in cancer - some are also unique to human

Rank	Tissue	Ratio Dark Residues
1	Heart	50%
2	Cervical Mucosa	50%
3	Natural Killer Cell	50%
4	Lung	49%
5	Testis	49%
6	Rectum	49%
7	Proximal Fluid Coronary Sinus	49%
8	Pancreas	49%
9	B. Lymphocyte	49%
10	Colon Muscle	49%
11	Bone Marrow Stromal Cell	48%
12	Hair Follicle	48%
13	Cytotoxic T Lymphocyte	48%
14	Helper T Lymphocyte	48%
15	Colon	48%
16	Ovary	48%
17	Stomach	48%
18	Spinal Cord	47%
19	Placenta	47%
20	Vitreous Humor	47%
21	Blood Platelet	47%
22	Prostate Gland	47%
23	Retina	47%
24	Salivary Gland	47%
25	Uterus	46%

Table 6.7: Tissues with the highest levels of darkness (only the first 25 entries).

In the current survey of half a million carefully curated Swiss-Prot sequences we found that ~13% are dark proteins; while many were not orphans (just hard to determine folds), most were, as evidenced by low evolutionary re-use scores. Although we used a very different approach to Levitt (a focus on structure versus sequence, and very different methods, thresholds, and cut-off values), both our studies are in broad agreement. Thus, our results suggest that many of the uncharacterized orphan sequences reported by Levitt (or the 'dark matter of the protein universe') are indeed proteins; this strengthens the suggestion that folds will eventually increase linearly with sequences (Levitt, 2009), and implies that dark proteins may remain a sizeable and irreducible feature of the protein universe.

6.7. Conclusion

The dark proteome is a key remaining frontier in the understanding of biological systems. We believe the current work will help focus future structural genomics and computational biology efforts to shed light on the remaining dark proteome, thus revealing currently unknown molecular processes of life.

6.8. Author Contributions

Nelson Perdigão contributions on this chapter were the writing of all algorithms for generation of Dark Proteome Database and the database itself from where the results were obtained, built of Fisher's exact tests, Protein Model Portal validations, writing and revision of the corresponding paper.

7. Dark Autonomy

This chapter was partially published in:

"Unexpected Features of the 'Dark' Proteome"
N. Perdigão, J. Heinrich , C. Stolte , K. S. Sabir , M. Buckley , B. Tabor , B. Signal , B. S. Gloss , C. J. Hammang , B. Rost, A. Schafferhans, S. I. O'Donoghue, *Proceedings of the National Academy of Sciences,* vol. 112 no. 52, pp.15898–15903, 2015.
doi: 10.1073/pnas.1508380112
http://www.pnas.org/content/112/52/15898

7.1. Summary

Recently we developed the Dark Proteome Database (DPD). This chapter exposes an add on to DPD, namely the so-called 'autonomy' tables. These tables were added to analyze Protein-Protein Interactions (PPI's) for dark and non-dark proteins, where we are especially interested in the capability of dark proteins to perform their biological tasks or functions alone i.e., having no interactions with surrounding proteins. I applied the data resources known as STRING (Franceschini et al., 2013) and HIPPIE (Schaefer et al., 2012), the later only for the human organism; these provide data on the protein-protein interaction networks obtained from experimentally-based quality scores. The results show clear evidence that - independently of the organism evaluated - dark proteins have significantly fewer interactions with other proteins, in comparison with non-dark proteins.

7.2. Introduction

Till today does not exist a comprehensive map of all relevant functionally for PPI's in simple or complex organisms. The existence of this map is of crucial importance to understand cellular behavior.

Several databases started to flourish helping in construction of this global protein interactions map. Some databases are dedicated to register interaction experiments such physical binding detection among proteins (Bader et al., 2008; Christensen et al., 2007; Devos & Russell, 2007; Hu et al., 2007); others are centered on specific model organisms (Kerrien et al., 2012; Smedley et al., 2014; Szklarczyk et al., 2011).

However, there are two difficulties: the first is the "tsunami" of genome and proteome sequencing information that must be processed putting the above map in standby; The second difficulty is in the way how proteins interact i.e., they also interact through indirect associations such as shared pathways which are not registered in interaction databases, but instead are registered in pathway databases (Chatr-Aryamontri et al., 2013; Prasad et al., 2009). This is my contribution to the above map specially to its dark side.

7.3. Data

To perform this study, the usage of the Dark Proteome Database (Chapter 5) was essential to identify dark and non-dark proteins. It was also used STRING (Search Tool

for the Retrieval of Interacting Genes/Proteins), and HIPPIE (integrating protein interaction networks with experiment based quality scores) databases/sites. The most reliable is the former (as explained below); the later was used for comparison purposes only in the human organism.

7.3.1. STRING

STRING is a database of known and predicted PPI's and contains information from several sources like experimental data, computational prediction methods (Salwinski et al., 2004; Alfarano et al., 2005; Licata et al., 2011). STRING imports knowledge not only from databases of physical interactions, but also from databases of curated biological pathway knowledge, like DIP (Salwinski et al., 2004), BIND (Alfarano et al., 2005), Reactome (Vastrik et al., 2007), KEGG (Kanehisa et al., 2008), HPRD (Prasad et al., 2009), EcoCyc (Keseler et al., 2011), MINT (Licata et al., 2011), IntAct (Kerrien et al., 2012), BioGRID (Chatr-Aryamontri et al., 2013), NCI-Nature Pathway Interaction Database and Gene Ontology (GO) protein complexes. This set is also complemented by predicted computational interactions, using algorithms like (Harrington et al., 2008; Skrabanek et al., 2008).

The STRING scheme classifies its functional link confidence into three different scores: low (<400), medium (400< score <700) and high (>700) confidence (Skrabanek et al., 2008) scores measuring the confidence in pair-wise functional interactions of the networks produced. Even assuming that sequence data is accurate computational tools can introduce noise when generation sequence similarity data occurs. Taking this noise into account, it is suggested to set a cut-off score above which an interaction is highly probable. In terms of functional classification accuracy what matters is high confidence score (=> 700) (Mazandu & Mulder, 2011), however low and medium confidence is also shown for comparison purposes.

7.3.2. HIPPIE

HIPPIE is a dataset of experimentally measured human protein-protein interactions (PPI) derived from several publicly available PPI datasets. For reference HIPPIE consists of 72,916 interactions (Harrington et al., 2008; Schaefer et al., 2012), which was used in this manuscript for several descriptive analyses. The live version of HIPPIE is monthly updated, which allows the automatically retrieve of the newest interaction data from most of the manually curated source databases like DIP (Salwinski et al.,

2004), BIND (Alfarano et al., 2005), MINT (Licata et al., 2011), IntAct (Kerrien et al., 2012), BioGRID (Chatr-Aryamontri et al., 2013) and integrate the new interactions and updated evidence records into HIPPIE.

7.4. Methods

7.4.1. Mapping Autonomy

For each Swiss-Prot protein, I categorized its autonomy as:

$$Autonomy\ Score = \begin{cases} 1 - 0.N\ if\ mN = 0\ and\ 0 \leq N \leq 900 \\ 0\ if\ mN \neq 0\ and\ N > 900 \end{cases} \quad \text{(Equation 7.1)}$$

Where *mN* indicate the number of matches that occur for link score of *N*. This means if the protein that have m0 equal to zero matches, then the protein is fully autonomous because at the lowest quality cut-off score no interactions occur between it and other proteins. On the other side, if at the highest cut-off score still exists interactions with other proteins (i.e., m900 is not zero) then it can be concluded that the protein is completely non-autonomous.

7.4.2. Density Plots

The density plots in Figs. 7.2, 7.3, 7.4, 7.5 and 7.6 (like in Chapter 6) were created using Gaussian kernel density estimations (Silverman, 1986), as implemented in the 'stat_density' and 'stat_density2d' functions of the 'ggplot2' package in R, and using default parameters like performed in Chapter 6 (Density Plots).

7.5. Results

In this section, it will be presented results obtained using the STRING database to study protein-proteins interactions for dark and non-dark proteins; each interaction is classified as either low, medium or high confidence (which indicate the most important interactions). The organisms benchmarked were archaea, bacteria, eukaryotes and human. For the last case (human) results from HIPPIE will also be presented (Fig. 7.1). For each protein, the number of interaction partners was determined using STRING interactions (Franceschini et al., 2013) that have high quality (700) or greater (http://bit.ly/1x0D8k6); this retrieves only interactions that are considered to be of high confidence. For comparison, low quality (score 300) and medium quality (score 500) interaction counts were also computed.

The lower number of interactions seen for dark proteins is quite striking (Figs. 7.2, 7.3, 7.4 and 7.5) – at first it may seem that this arises simply because dark proteins have not been as well studied; however, STRING's annotation process aggregates multiple types of evidence for interactions, primarily high-throughput experimental studies as well as text mining, and in some cases interaction is inferred via homology. This would reduce potential study bias. Each of the interaction profiles (Figs. 7.2, 7.3, 7.4 and 7.5) also shows a prominent peak at around 100-120 interactions; most likely, this arises from the ribosome complex, a common and well-studied feature for which STRING provides interaction information across many organisms. Note that, using this high confidence threshold, lack of known interactions does not necessarily imply that a protein has no interactions - rather it implies that all known evidence for interaction with other proteins is rather weak.

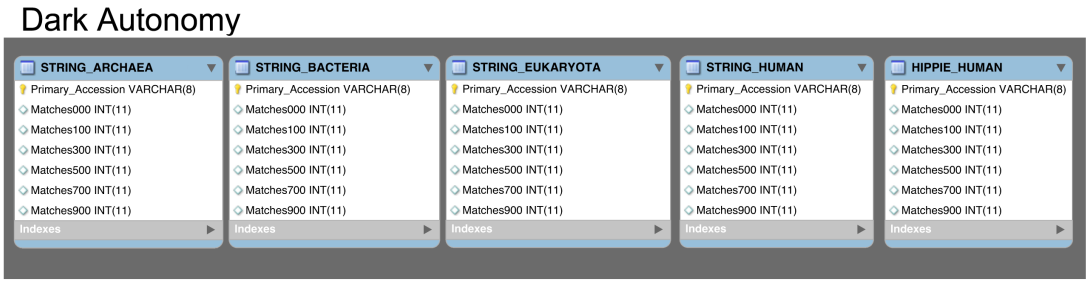


Figure 7.1: Dark Autonomy database with Archaea, Bacteria, Eukaryota and Human tables using STRING with 0, 100, 300, 500, 700 and 900 score thresholds. For Human, the HIPPIE table with the same thresholds is also included.

Figure 7.2 shows the results of protein-protein interactions for dark and non-dark proteins of Archaea for the three levels of quality. For high quality confidence, it is observed that dark proteins have much less interactions compared with non-dark proteins.

Concerning Bacteria Figure 7.3 shows the results of protein-protein interactions for dark and non-dark proteins for the three levels of quality. Again, for high quality confidence it is observed that dark proteins have much less interactions compared with non-dark proteins.

In the Eukaryotes case Figure 7.4 shows the results of protein-protein interactions for dark and non-dark proteins for the three levels of quality. For high quality confidence it is observed that dark proteins have much less interactions compared with non-dark proteins. The pattern repeats itself. Note also that I did not calculate the profile of interaction partners for Viral proteins since STRING provides no information for them.

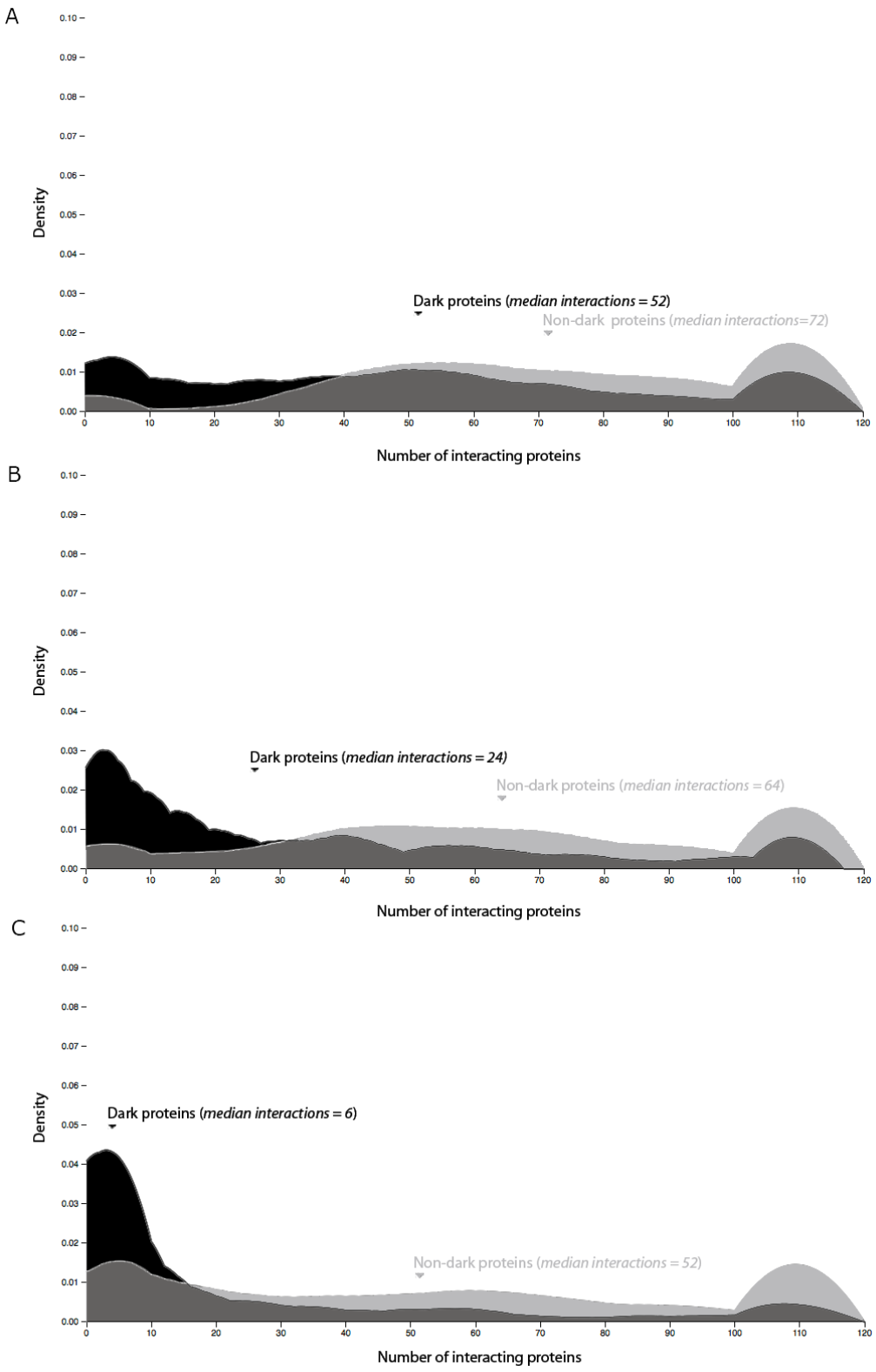


Figure 7.2: Protein-protein interactions for Archaea using STRING. A) For low quality it was observed that dark proteins have fewer interactions (median = 52) compared to non-dark proteins (median = 72); B) For medium quality, dark proteins have even fewer interactions with other proteins (median = 24), compared to non-dark proteins (median = 64); C) For high quality dark proteins have even less interactions with other proteins (median = 6), compared to non-dark proteins (median = 52).

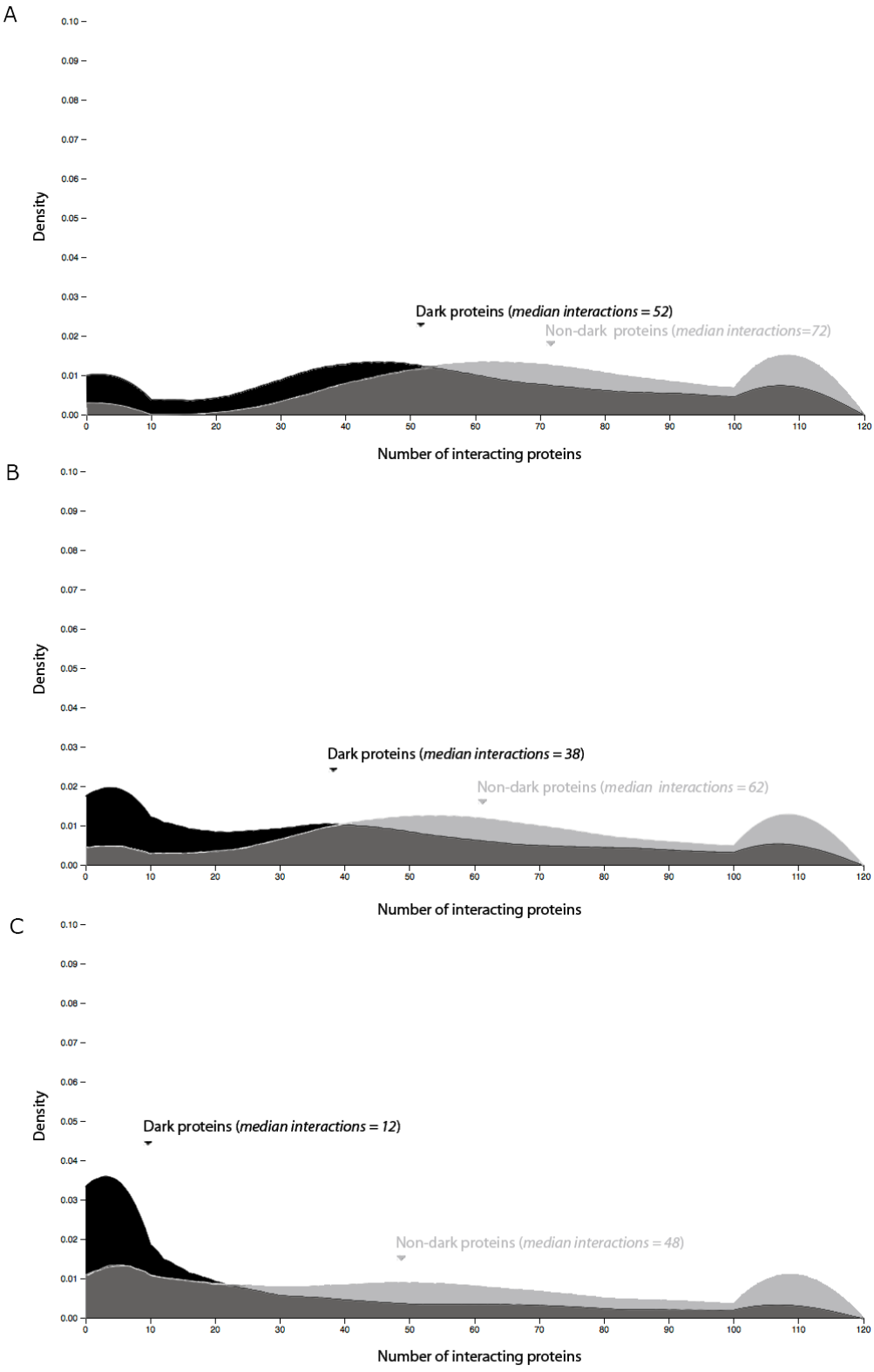


Figure 7.3: Protein-protein interactions for Bacteria using STRING. A) For low quality it was observed that dark proteins have fewer interactions (median = 52) compared to non-dark proteins (median = 72); B) For medium quality dark proteins have even fewer interactions with other proteins (median = 38), compared to non-dark proteins (median = 62); C) For high quality dark proteins have even less interactions with other proteins (median = 12), compared to non-dark proteins (median = 48).

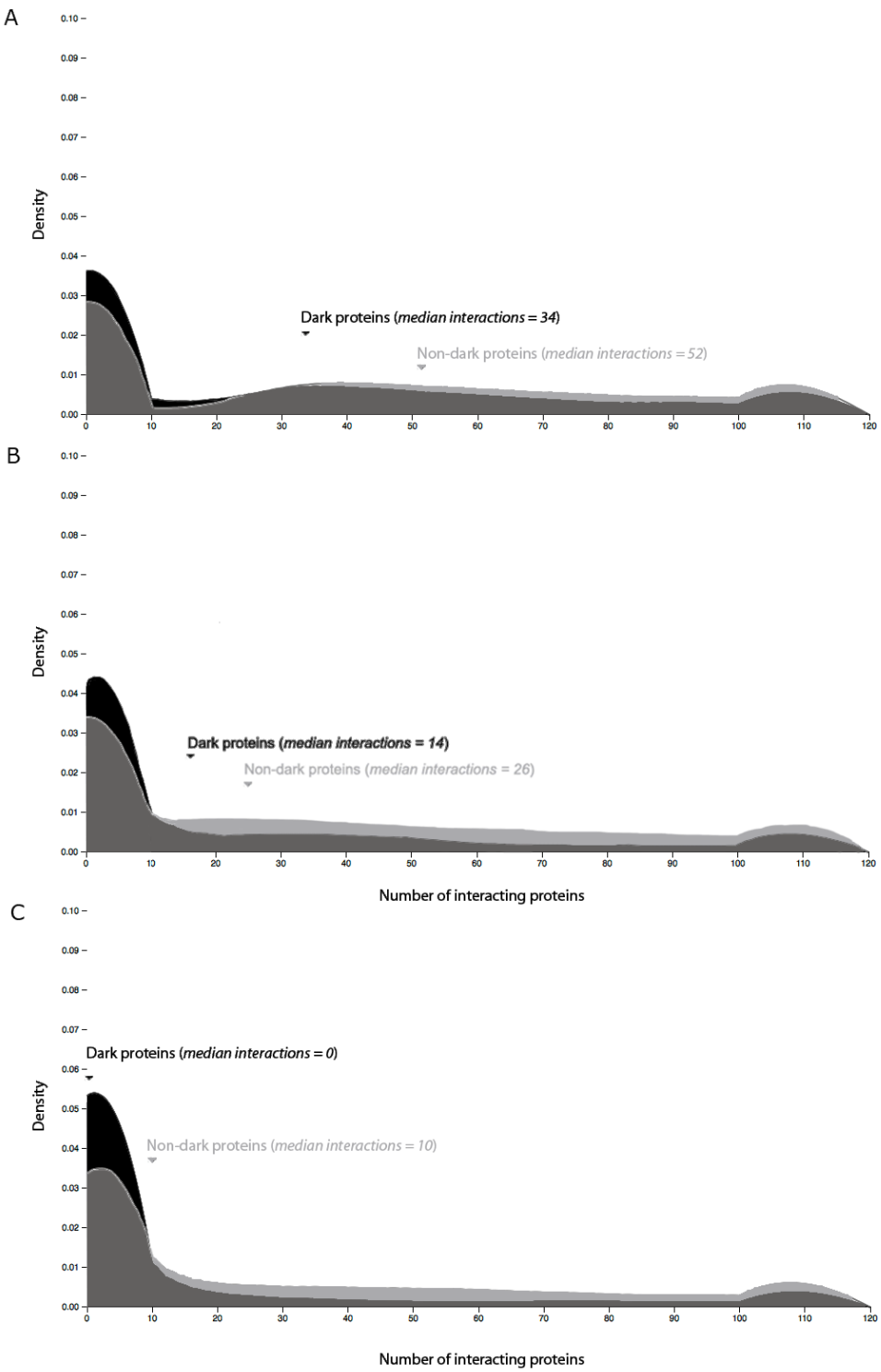


Figure 7.4: Protein-protein interactions for Eukaryotes using STRING. A) For low quality its observed that dark proteins have fewer interactions (median = 34) compared to non-dark proteins (median = 52); B) For medium quality its observed that dark proteins have fewer interactions (median = 14) compared to non-dark proteins (median = 26); C) For high quality dark proteins have no interactions with other proteins (median = 0), compared to non-dark proteins (median = 10).

7.5.1. The Human Dark Autonomy

As stated previously, I also made a special case study for the Human organism where Figure 7.5 shows the results of protein-protein interactions for dark and non-dark proteins for the three levels of quality. Similarly, human dark proteins have fewer interactions with others proteins and are associated with secretion, transmembrane regions, and cleavage (Table 5.5);

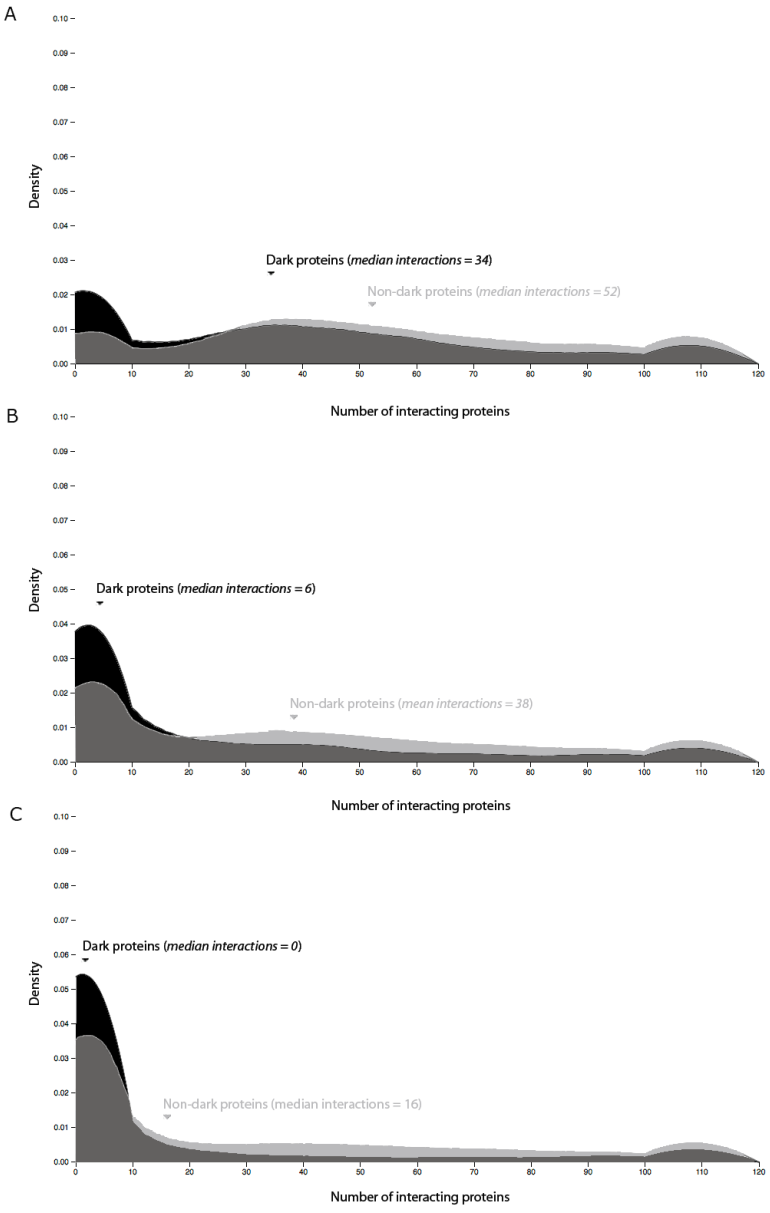


Figure 7.5: Protein-proteins interactions for Human using STRING. A) For low quality it was observed that dark proteins have fewer interactions (median = 34) compared to non-dark proteins (median = 52); B) For medium quality, dark proteins have even fewer interactions with other proteins (median = 6), compared to non-dark proteins (median = 38); C) For high quality, dark proteins have no interactions with other proteins (median = 0), compared to non-dark proteins (median = 16).

Next will be presented the results for HIPPIE database concerning Human organism. Figure 7.6 shows the results of protein-protein interactions for dark and non-dark proteins of human for the three levels of quality.

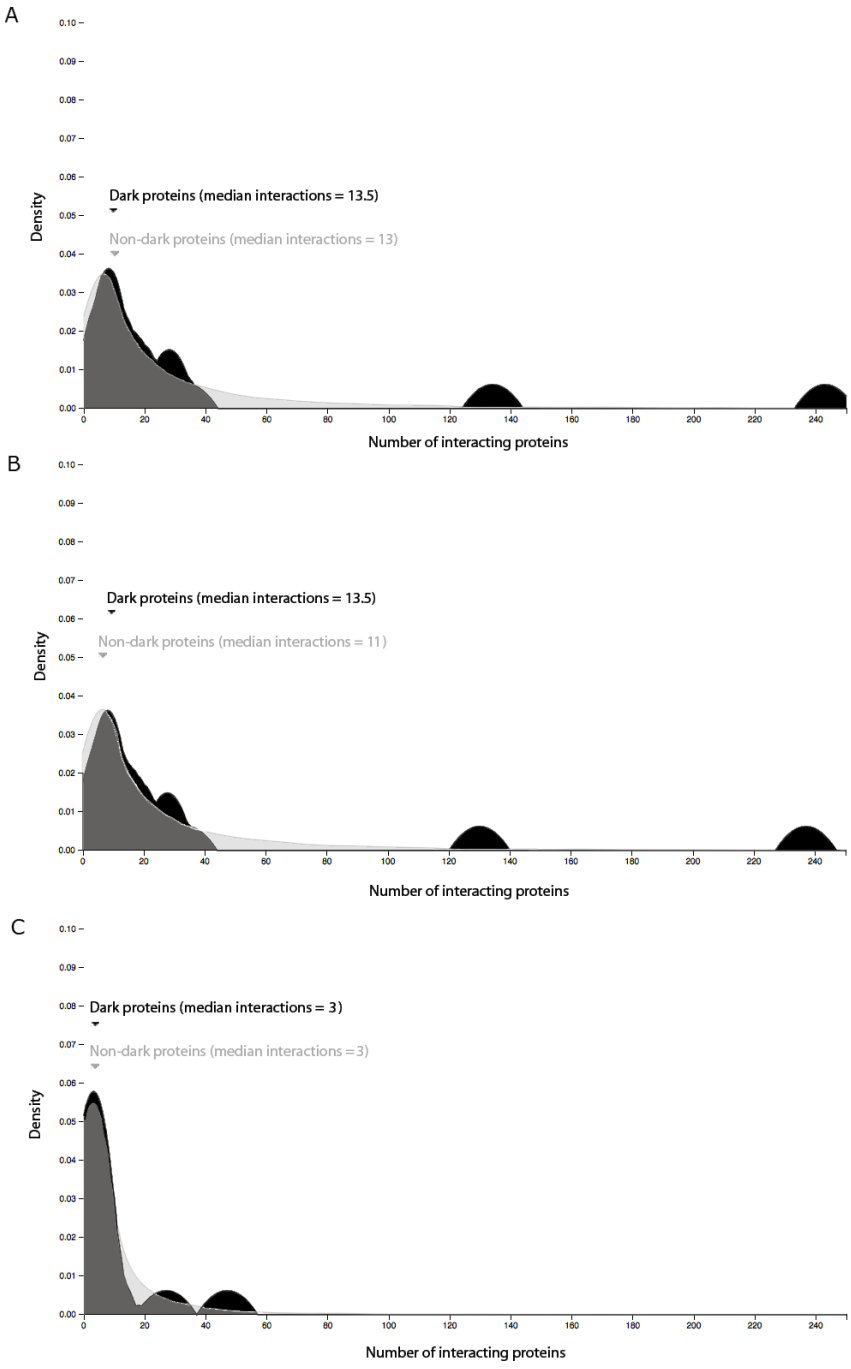


Figure 7.6: Protein-proteins interactions for Human using HIPPIE. A) For low quality, it was observed that dark proteins have slight more interactions (median = 13,5) compared to non-dark proteins (median = 13); B) For medium quality, dark proteins have also slight more interactions with other proteins (median = 13,5), compared to non-dark proteins (median = 11); C) For high quality, dark proteins have the same number of interactions with other proteins (median = 3), compared to non-dark proteins (median = 3).

For high quality confidence it was observed that dark proteins have substantially the same interactions compared with non-dark proteins. Here the pattern observed, namely the prominent peak at around 100-120 interactions, has a shift to the right of about 125; that is still consistent from the ribosome complex (let us consider in this case lack of precision) but another unusual pattern comes around the 240 interactions.

7.6. Discussion

Concerning HIPPIE results (and in comparison with STRING results), it was observed that was detected a peak for proteins with 240 interactions (something not seen in STRING) in low and medium link quality. According with HIPPIE non-dark proteins has more interactions in comparison with the dark ones (a marginal difference - Figure 7.6). When we move to high quality, and again concerning HIPPIE, there is no observable difference, i.e., the interaction counts for dark and non-dark proteins are the same (3 median interactions).

As I said in the beginning of this chapter, STRING is regarded as a reference database, since it is much more widely used than HIPPIE. In addition, STRING imports much more protein association knowledge not only from databases of physical interactions, but also from databases of curated biological pathway knowledge, in comparison with HIPPIE. Thus, the results observed for STRING are more likely to correctly reflect reality.

7.7. Conclusions

The conclusion is clear from the results shown above, and are sustained across all the organisms evaluated. In short, there is clear and consistent evidence that dark proteins tend to have less interactions with other proteins in comparison with the non-dark ones. As the interaction quality increases from low to high, this observation becomes even more evident. Therefore, it could be said that dark proteins are more autonomous than non-dark proteins, since they appear to fulfill their biological function with substantially fewer interactions with proteins.

7.8. Author Contributions

Nelson Perdigão contributions on this chapter were all algorithms for generation of Dark Autonomy database, the database itself and its validation.

PART V

CONCLUSIONS

8. General Discussion

Currently, the Protein Data Bank (PDB) archives the world's knowledge about protein structures contains just over 120,000 experimentally determined structures of large biological molecules. These data can be used as input for high-throughput computational modeling studies that can accurately predict structures for many protein sequences not in the PDB. In this thesis, the success of these modeling efforts has been mapped – we see that many proteins contain regions of unknown structure, and so they have been considered part of the dark proteome. This systematic mapping and exploration of the dark proteome could help clarify future research directions, in an analogous way to which studies of dark matter have done in physics.

In this study, the features of the dark proteome's proteins were analyzed and it was shown that much of these unknown regions of proteins cannot readily be explained. The fact that so much is still unknown shows there is still much potential for a broad spectrum of research into the complexity of biology. By pushing forth the set of unexpected features of the dark proteins, it has raised much discussion throughout the structural biology research community.

In this research, the majority of the dark proteome was proven to be determined by different factors than those which were initially expected, such as disorder, transmembrane regions, or compositional bias (i.e., the known unknowns). Instead, most of the dark proteome was found to be ordered, globular, and with low compositional bias (and therefore, 'unknown unknowns'). It was also unveiled that dark proteins have a diversity of functions, but many have unknown function, and it is unclear to what extent they interact with other proteins. Curiously, an overrepresented set of dark proteins are extracellular; they also tend to be shorter than non-dark proteins.

The dark proteome in bacteria represent a smaller portion, implying that knowledge of their structural biology is more complete, which may prove useful in the research and development of specific types of antibacterial drugs. Similar results were seen for the archaea proteome. But the proteome of the Eukaryotes and Viruses is the opposite, in that most of it is dark, or of unknown structure.

Structural biology experiments with non-dark proteins have been, until now, the focus of research. This is mostly due to the fact that those proteins were more tangible

in terms of isolation and crystallization. It may be that the dark proteins requires expertise that has yet to be developed. This research raises many questions that are left without answer. However, it does clarify the fact that there is still much to learn about the protein structure and function, and that in order to facilitate future research, interdisciplinary collaboration is required to reach the necessary tools that would allow this exploration.

I am certain, that many proteins that are part of the dark proteome are involved in a many different functions in the cell, like cellular signaling or cellular organization; undoubtedly, many of these proteins will be associated with diseases, such as cancer, diabetes, cardiovascular disease, neurodegenerative diseases such as Parkinson or Alzheimer. Therefore, mapping the dark proteome is likely to have an important impact in human biology, and in medicine. After this mapping the next step will be to focus on decoding the dark proteome, so that new drugs or therapies could occur, and this poses many challenges.

I am certain once again, like computer science played a central and important role in mapping the dark proteome, computer science will play again an important role in the dark proteome decoding, even more that the traditional structural biology methods like crystallography, as these are physically limited. Computation predictions will also be a key player for more closely examining the role of intrinsic disorder as well as for transmembrane proteins.

As we gain more knowledge about the structure of proteins that are currently ‘dark’, we will gain insight into their function, which in turn will give new insight into a range of diseases, as well as other important advances in the life science.

Like Donald Knuth said “can’t be as confident about computer science as I can about biology. Biology easily has 500 years of exciting problems to work on. It’s at that level.” The work presented in this thesis provides yet another demonstration that – going forward - Biology cannot solve its problems without computer science.

References

Alfarano, C., Andrade, C. E., Anthony, K., Bahroos, N., Bajec, M., Bantoft, K., … Hogue, C. W. V. (2005). The Biomolecular Interaction Network Database and related tools 2005 update. *Nucleic Acids Research*, *33*(DATABASE ISS.), 418–424. http://doi.org/10.1093/nar/gki051

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., & Lipman, D. J. (1990). Altschul et al.. 1990. Basic Local Alignment Search Tool.pdf. *Journal of Molecular Biology*. http://doi.org/10.1016/S0022-2836(05)80360-2

Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Research*. http://doi.org/10.1093/nar/25.17.3389

Andrade, M. A., O'Donoghue, S. I., & Rost, B. (1998). Adaptation of protein surfaces to subcellular location. *Journal of Molecular Biology*, *276*(2), 517–525. http://doi.org/10.1006/jmbi.1997.1498

Anfinsen, C. B. (1973). Principles that Govern the Folding of Protein Chains. *Science*, *181*(4096), 223–230. http://doi.org/10.1126/science.181.4096.223

Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., … Yeh, L.-S. L. (2004). UniProt: the Universal Protein knowledgebase. *Nucleic Acids Research*, *32*(Database issue), D115-9. http://doi.org/10.1093/nar/gkh131

Bader, S., Kühner, S., & Gavin, A.-C. (2008). Interaction networks for systems biology. *FEBS Letters*, *582*(8), 1220–1224. http://doi.org/10.1016/j.febslet.2008.02.015

Bairoch, A., & Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. *Nucleic Acids Research*, *28*(1), 45–48. http://doi.org/10.1093/nar/28.1.45

Benjamini, Y., & Hochberg, Y. (1995). Controlling the False Discovery Rate: A Practical and Powerful Approach to Multiple Testing. *Journal of the Royal Statistical Society. Series B (Methodological)*, *57*(1), 289–300. http://doi.org/10.2307/2346101

Berman, H., Henrick, K., Nakamura, H., & Markley, J. L. (2007). The worldwide Protein Data Bank (wwPDB): Ensuring a single, uniform archive of PDB data. *Nucleic Acids Research*, *35*(SUPPL. 1). http://doi.org/10.1093/nar/gkl971

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., … Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research*, *28*(1), 235–242. http://doi.org/10.1093/nar/28.1.235

Bernstein, F. C. (1977). The Protein D a t a Bank: A Computer-based Archival File for Macromolecular Structures. *Alcohol*, *112*, 535–542.

Bertone, G., Hooper, D., & Silk, J. (2005). Particle dark matter: Evidence, candidates and constraints. *Physics Reports*. http://doi.org/10.1016/j.physrep.2004.08.031

Bigelow, H., & Rost, B. (2006). PROFtmb: A web server for predicting bacterial transmembrane beta barrel proteins. *Nucleic Acids Research*, *34*(WEB. SERV. ISS.). http://doi.org/10.1093/nar/gkl262

Bostock, M., Ogievetsky, V., & Heer, J. (2011). D³ Data-Driven Documents. *IEEE Transactions on Visualization and Computer Graphics*, *17*(12), 2301–2309. http://doi.org/10.1109/TVCG.2011.185

Bright, J. N., Woolf, T. B., & Hoh, J. H. (2001). Predicting properties of intrinsically unstructured proteins. *Progress in Biophysics and Molecular Biology*. http://doi.org/10.1016/S0079-6107(01)00012-8

Carpenter, E. P., Beis, K., Cameron, A. D., & Iwata, S. (2008). Overcoming the challenges of membrane protein crystallography. *Current Opinion in Structural Biology*, *18*(5), 581–586. http://doi.org/10.1016/j.sbi.2008.07.001

Cedano, J., Aloy, P., Pérez-Pons, J. a, & Querol, E. (1997). Relation between amino acid composition and cellular location of proteins. *Journal of Molecular Biology*, *266*(3), 594–600. http://doi.org/10.1006/jmbi.1996.0804

Chatr-Aryamontri, A., Breitkreutz, B. J., Heinicke, S., Boucher, L., Winter, A., Stark, C., … Tyers, M. (2013). The BioGRID interaction database: 2013 Update. *Nucleic Acids Research*, *41*(D1), 470–478. http://doi.org/10.1093/nar/gks1158

Chothia, C. (1992). Proteins. One thousand families for the molecular biologist. *Nature*, *357*(6379), 543. article.

Chothia, C., & Lesk, a M. (1986). The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, *5*(4), 823–826. http://doi.org/060 fehlt

Christensen, C., Thakar, J., & Albert, R. (2007). Systems-level insights into cellular regulation: inferring, analysing, and modelling intracellular networks. *IET Systems Biology*, *1*(2), 61–77. http://doi.org/10.1049/iet-syb:20060071

Clark, K., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., & Sayers, E. W. (2016). GenBank. *Nucleic Acids Research*, *44*(D1), D67–D72. http://doi.org/10.1093/nar/gkv1276

Cochrane, G., Akhtar, R., Aldebert, P., Althorpe, N., Baldwin, A., Bates, K., … Birney, E. (2008). Priorities for nucleotide trace, sequence and annotation data capture at the Ensembl Trace Archive and the EMBL Nucleotide Sequence Database. *Nucleic Acids Research*, *36*(SUPPL. 1). http://doi.org/10.1093/nar/gkm1018

Crick, F. (1956). Ideas on Protein Synthesis. In *Symp. Soc. Exp. Biol. XII* (pp. 139–163). Retrieved from http://profiles.nlm.nih.gov/ps/retrieve/ResourceMetadata/SCBBFT

Crick, F. (1958). On Protein Synthesis. In *The Symposia of the Society for Experimental Biology* (pp. 138–166).

Crick, F. (1970). Central dogma of molecular biology. *Nature*, *227*(5258), 561–563. http://doi.org/10.1038/227561a0

Davey, N. E., Travé, G., & Gibson, T. J. (2011). How viruses hijack cell regulation. *Trends in Biochemical Sciences*. http://doi.org/10.1016/j.tibs.2010.10.002

Dayhoff, M., Schwartz, R. M., & Orcutt, B. C. (1978). Atlas of protein sequence and structure (Vol. 5, pp. 345–352). inbook, Silver Spring, Maryland: National Biomedical Research Foundation.

Devos, D., & Russell, R. B. (2007). A more complete, complexed and structured interactome. *Current Opinion in Structural Biology*. http://doi.org/10.1016/j.sbi.2007.05.011

Dodge, C., Schneider, R., & Sander, C. (1998). The HSSP database of protein structure sequence alignments and family profiles. *Nucleic Acids Research*, *26*(1), 313–315.

Dosztányi, Z., Csizmok, V., Tompa, P., & Simon, I. (2005). IUPred: Web server for the prediction of intrinsically unstructured regions of proteins based on estimated energy content. *Bioinformatics*, *21*(16), 3433–3434. http://doi.org/10.1093/bioinformatics/bti541

Drake, J. W., Charlesworth, B., Charlesworth, D., & Crow, J. F. (1998). Rates of spontaneous mutation. *Genetics*, *148*(4), 1667–1686. http://doi.org/citeulike-article-id:610966

Dunker, A. K., & Obradovic, Z. (2001). The protein trinity — linking function and disorder, *99124*.

Dunker, a. K., Lawson, J. D., Brown, C. J., Williams, R. M., Romero, P., Oh, J. S., … Obradovic, Z. (2001). Intrinsically disordered protein. *Journal of Molecular Graphics and Modelling*, *19*(1), 26–59. http://doi.org/10.1016/S1093-3263(00)00138-8

Dunker, a K., Oldfield, C. J., Meng, J., Romero, P., Yang, J. Y., Chen, J. W., … Uversky, V. N. (2008). The unfoldomics decade: an update on intrinsically disordered proteins. *BMC Genomics*, *9 Suppl 2*, S1. http://doi.org/10.1186/1471-2164-9-S2-S1

Dyson, H. J., & Wright, P. E. (2005). Intrinsically unstructured proteins and their functions. *Nature Reviews. Molecular Cell Biology*, *6*(3), 197–208. http://doi.org/10.1038/nrm1589

Eswar, N., Webb, B., Marti-Renom, M. A., Madhusudhan, M. S., Eramian, D., Shen, M.-Y., … Sali, A. (2007). Comparative protein structure modeling using MODELLER. *Current Protocols in Protein Science / Editorial Board, John E. Coligan ... [et Al.]*, *Chapter 2*, Unit 2.9. http://doi.org/10.1002/0471140864.ps0209s50

Etzold, T., & Argos, P. (1993). SRS--an indexing and retrieval tool for flat file data libraries. *CABIOS*, *9*(1), 49–57. http://doi.org/10.1093/bioinformatics/9.1.49

Fisher, R. (1922). On the interpretation of χ2 from contingency tables, and the calculation of P. *Journal of the Royal Statistical Society*, *85*(1), 87–94. http://doi.org/10.2307/2340521

Fisher, R. (1925). *Statistical methods for research workers*. *Biological monographs and manuals*. http://doi.org/10.1056/NEJMc061160

Fisher, R. A. (1922). On the Interpretation of χ^2 from Contingency Tables, and the Calculation of P. *Journal of the Royal Statistical Society*, *85*(1), 87–94. http://doi.org/10.2307/2340521

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., … Jensen, L. J. (2013). STRING v9.1: Protein-protein interaction networks, with increased coverage and integration. *Nucleic Acids Research*, *41*(D1), 808–815. http://doi.org/10.1093/nar/gks1094

Gribskov, M., McLachlan, a D., & Eisenberg, D. (1987). Profile analysis: detection of distantly related proteins. *Proceedings of the National Academy of Sciences of the United States of America*, *84*(13), 4355–4358. http://doi.org/10.1073/pnas.84.13.4355

Haas, J., Roth, S., Arnold, K., Kiefer, F., Schmidt, T., Bordoli, L., & Schwede, T. (2013). The Protein Model Portal--a comprehensive resource for protein structure and model information. *Database : The Journal of Biological Databases and Curation*, *2013*(8), bat031. http://doi.org/10.1093/database/bat031

Harrington, E. D., Jensen, L. J., & Bork, P. (2008). Predicting biological networks from genomic data. *FEBS Letters*. http://doi.org/10.1016/j.febslet.2008.02.033

Hartshorn, M. J. (2002). AstexViewer: a visualisation aid for structure-based drug design. *Journal of Computer-Aided Molecular Design*, *16*(12), 871–881. http://doi.org/10.1023/A:1023813504011

Hauser, M., Mayer, C. E., Soding, J., & Söding, J. (2013). kClust: fast and sensitive clustering of large protein sequence databases. *BMC Bioinformatics*, *14*(1), 248. http://doi.org/10.1186/1471-2105-14-248

Hilbert, M., & López, P. (2011). The World's Technological Capacity to Store, Communicate, and Compute Information. *Science*, *332*, 60–65. http://doi.org/10.1126/science.1200970

Hobohm, U., Scharf, M., Schneider, R., & Sander, C. (1992). Selection of representative protein data sets. *Protein Science : A Publication of the Protein Society*, *1*(3), 409–17. http://doi.org/10.1002/pro.5560010313

Holm, L., & Sander, C. (1994). The FSSP database of structurally aligned protein fold families. *Nucleic Acids Research*, *22*(17), 3600–3609. http://doi.org/061/2

Holm, L., & Sander, C. (1996). Mapping the protein universe. *Science (New York, N.Y.)*, *273*(5275), 595–603. http://doi.org/10.1126/science.273.5275.595

Holm, L., & Sander, C. (1997). Dali/FSSP classification of three-dimensional protein folds. *Nucleic Acids Research*, *25*(1), 231–234. http://doi.org/10.1093/nar/25.1.231

Hu, Z., Mellor, J., Wu, J., Kanehisa, M., Stuart, J. M., & DeLisi, C. (2007). Towards zoomable multidimensional maps of the cell. *Nature Biotechnology*, *25*(5), 547–554. http://doi.org/10.1038/nbt1304

Hunter, L. (2009). *The processes of life: an introduction to molecular biology*. book, Mit Press Cambridge, MA.

Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T. K., Bateman, A., … Yong, S. Y. (2012). InterPro in 2011: New developments in the family and domain prediction database. *Nucleic Acids Research*, *40*(Database Issue), 306–312. http://doi.org/10.1093/nar/gkr948

Huntley, M. A., & Golding, G. B. (2002). Simple sequences are rare in the Protein Data Bank. *Proteins: Structure, Function and Genetics*, *48*(1), 134–140. http://doi.org/10.1002/prot.10150

Illergård, K., Ardell, D. H., & Elofsson, A. (2009). Structure is three to ten times more conserved than sequence--a study of structural response in protein cores. *Proteins*, *77*(3), 499–508. http://doi.org/10.1002/prot.22458

Ilyin, V. A., Abyzov, A., & Leslin, C. M. (2004). Structural alignment of proteins by a novel TOPOFIT method, as a superimposition of common volumes at a topomax point. *Protein Science : A Publication of the Protein Society*, *13*(7), 1865–74. http://doi.org/10.1110/ps.04672604

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., … Yamanishi, Y. (2008). KEGG for linking genomes to life and the environment. *Nucleic Acids Research*, *36*(SUPPL. 1). http://doi.org/10.1093/nar/gkm882

Kerrien, S., Aranda, B., Breuza, L., Bridge, A., Broackes-Carter, F., Chen, C., … Hermjakob, H. (2012). The IntAct molecular interaction database in 2012. *Nucleic Acids Research*, *40*(D1), 1–6. http://doi.org/10.1093/nar/gkr1088

Keseler, I. M., Collado-Vides, J., Santos-Zavaleta, A., Peralta-Gil, M., Gama-Castro, S., Muniz-Rascado, L., … Karp, P. D. (2011). EcoCyc: A comprehensive database of Escherichia coli biology. *Nucleic Acids Research*, *39*(SUPPL. 1), 583–590. http://doi.org/10.1093/nar/gkq1143

Prasad, T. S., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., … Pandey, A. (2009). Human Protein Reference Database--2009 update. *Nucleic Acids Research*, *37*(Database issue), D767–D772. http://doi.org/10.1093/nar/gkn892

Khafizov, K., Madrid-Aliste, C., Almo, S. C., & Fiser, A. (2014). Trends in structural coverage of the protein universe and the impact of the Protein Structure Initiative. *Proceedings of the National Academy of Sciences of the United States of America*, *111*(10), 3733–8. http://doi.org/10.1073/pnas.1321614111

Kiefer, F., Arnold, K., Künzli, M., Bordoli, L., & Schwede, T. (2009). The SWISS-MODEL Repository and associated resources. *Nucleic Acids Research*, *37*(SUPPL. 1). http://doi.org/10.1093/nar/gkn750

Komdeur, J., Eikenaar, C., Brouwer, L., & Richardson, D. S. (2009). Encyclopedia of life sciences. *Life Sciences*, *21*(1), 1–8. http://doi.org/10.1002/047001590X

Koonin, E. V, Wolf, Y. I., & Karev, G. P. (2002). The structure of the protein universe and genome evolution. *Nature*, *420*(6912), 218–223. http://doi.org/10.1038/nature01256

Koshland, D. E. (2002). Special essay. The seven pillars of life. *Science (New York, N.Y.)*, *295*(5563), 2215–2216. http://doi.org/10.1126/science.1068489

Krogh, a, Larsson, B., von Heijne, G., & Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology*, *305*, 567–580. http://doi.org/10.1006/jmbi.2000.4315

Lagerkvist, U. (2005). *Enigma of Ferment: From the Philosopher's Stone to the First Biochemical Nobel Prize*. *Enigma of Ferment: From the Philosopher's Stone to the First Biochemical Nobel Prize*. http://doi.org/10.1142/5900

Lesk, A. M. (2002). *Introduction to Bioinformatics*. book, Oxford University Press.

Levitt, M. (2009). Nature of the protein universe. *Proceedings of the National Academy of Sciences of the United States of America*, *106*(27), 11079–11084. http://doi.org/10.1073/pnas.0905029106

Licata, L., Briganti, L., Peluso, D., Perfetto, L., Iannuccelli, M., Galeota, E., … Cesareni, G. (2011). MINT, the molecular interaction database: 2012 update. *Nucleic Acids Research*, *40*(D1), gkr930-. http://doi.org/10.1093/nar/gkr930

Liu, J., & Rost, B. (2001). Comparing function and structure between entire proteomes. *Protein Science : A Publication of the Protein Society*, *10*(10), 1970–1979. http://doi.org/10.1110/ps.10101

Mallick P., R. D. & E. D. (2001). DAPS: Database of Distant Aligned Protein Structures. site, http://www.doe-mbi.ucla.edu/DAPS/.

Mardis, E. R. (2011). A decade's perspective on DNA sequencing technology. *Nature*, *470*(7333), 198–203. http://doi.org/10.1038/nature09796

Marsden, R. L., Lewis, T. A., & Orengo, C. A. (2007). Towards a comprehensive structural coverage of completed genomes: a structural genomics viewpoint. *BMC Bioinformatics*, *8*, 86. http://doi.org/10.1186/1471-2105-8-86

Mashima, J., Kodama, Y., Kosuge, T., Fujisawa, T., Katayama, T., Nagasaki, H., … Takagi, T. (2016). DNA data bank of Japan (DDBJ) progress report. *Nucleic Acids Research*, *44*(D1), D51–D57. http://doi.org/10.1093/nar/gkv1105

Mattick, J. S. (2003). Challenging the dogma: The hidden layer of non-protein-coding RNAs in complex organisms. *BioEssays*, *25*, 930–939. http://doi.org/10.1002/bies.10332

Mazandu, G. K., & Mulder, N. J. (2011). Scoring Protein Relationships in Functional Interaction Networks Predicted from Sequence Data. *PLoS ONE*, *6*(4), e18607. http://doi.org/10.1371/journal.pone.0018607

McLachlan, A. D. (1971). Tests for comparing related amino-acid sequences. Cytochrome c and cytochrome c 551 . *Journal of Molecular Biology*, *61*(2), 409–424. http://doi.org/10.1016/0022-2836(71)90390-1

Meyers, R. a. (2005). *Encyclopedia of Molecular Cell Biology and Molecular Medicine*. *Molecular Cell* (Vol. 16). http://doi.org/3-527-30543-2

Mirsky, a. E., & Paulin, L. (1936). On the structure of native, denatured, and coagulated proteins. *Proceedings of the National Academy of Sciences of the United States of America*, *22*(1892), 439–447. http://doi.org/10.1017/CBO9781107415324.004

Nepomnyachiy, S., Ben-Tal, N., & Kolodny, R. (2014). Global view of the protein universe. *Proceedings of the National Academy of Sciences*, 201403395. http://doi.org/10.1073/pnas.1403395111

Nurse, P. (2003). The great ideas of biology. In *Clinical Medicine* (Vol. 3, pp. 560–568). http://doi.org/10.7861/clinmedicine.3-6-560

O'Donoghue, S. I., Gavin, A.-C., Gehlenborg, N., Goodsell, D. S., Hériché, J.-K., Nielsen, C. B., … Wong, B. (2010). Visualizing biological data-now and in the future. *Nature Methods*, *7*(3 Suppl), S2–S4. http://doi.org/10.1038/nmeth.f.301

O'Donoghue, S. I., Meyer, J. E. W., Schafferhans, A., & Fries, K. (2004). The SRS 3D module: Integrating structures, sequences and features. *Bioinformatics*, *20*(15), 2476–2478. http://doi.org/10.1093/bioinformatics/bth260

O'Donoghue, S. I. O., Sabir, K. S., Kalemanov, M., Stolte, C., Wellmann, B., Ho, V., … others. (2015). Aquaria: simplifying discovery and insight from protein structures. *Nature Methods*, *12*(2), 98–99. article. http://doi.org/10.1038/nmeth.3258

Oldfield, C. J., & Dunker, a K. (2014). Intrinsically disordered proteins and intrinsically disordered protein regions. *Annual Review of Biochemistry*, *83*, 553–84. http://doi.org/10.1146/annurev-biochem-072711-164947

Oldfield, C. J., Xue, B., Van, Y. Y., Ulrich, E. L., Markley, J. L., Dunker, A. K., & Uversky, V. N. (2013). Utilization of protein intrinsic disorder knowledge in structural proteomics. *Biochimica et Biophysica Acta - Proteins and Proteomics*, *1834*(2), 487–498. http://doi.org/10.1016/j.bbapap.2012.12.003

Ota, M., Koike, R., Amemiya, T., Tenno, T., Romero, P. R., Hiroaki, H., … Fukuchi, S. (2013). An assignment of intrinsically disordered regions of proteins based on NMR structures. *Journal of Structural Biology*, *181*(1), 29–36. http://doi.org/10.1016/j.jsb.2012.10.017

Overington, J. P., Al-Lazikani, B., & Hopkins, A. L. (2006). Opinion - How many drug targets are there? *Nature Reviews Drug Discovery*, *5*(12), 993–996. http://doi.org/10.1038/nrd2199

Parry, D. A. D., & Squire, J. M. (1998). Fibrous Proteins. *Journal of Structural Biology*, *122*(1–2), 1–2. http://doi.org/10.1006/jsbi.1998.3996

Pauling, L., & Coryell, C. D. (1936). The Magnetic Properties and Structure of Hemoglobin, Oxyhemoglobin and Carbonmonoxyhemoglobin. *Proceedings of the National Academy of Sciences of the United States of America*, *22*, 210–216. http://doi.org/10.1073/pnas.22.4.210

Pearson, W. R., & Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proceedings of the National Academy of Sciences of the United States of America*, *85*(8), 2444–8. http://doi.org/10.1073/pnas.85.8.2444

Perdigão, N., Heinrich, J., Stolte, C., Sabir, K. S., Buckley, M. J., Tabor, B., … O'Donoghue, S. I. (2015). Unexpected features of the dark proteome. *Proceedings of the National Academy of Sciences*. http://doi.org/10.1073/pnas.1508380112

Perdigão, N., Soldatos, T. G., Sabir, K. S., & O Donoghue, S. I. (2015). Visual Analytics of Gene Sets Comparison. In *2015 Big Data Visual Analytics (BDVA)* (pp. 1–2). IEEE. http://doi.org/10.1109/BDVA.2015.7314304

Petrey, D., Chen, T. S., Deng, L., Garzon, J. I., Hwang, H., Lasso, G., … Honig, B. (2015). Template-based prediction of protein function. *Current Opinion in Structural Biology*, *32*, 33–38. article. http://doi.org/10.1016/j.sbi.2015.01.007

Pettersen, E. F., Goddard, T. D., Huang, C. C., Couch, G. S., Greenblatt, D. M., Meng, E. C., & Ferrin, T. E. (2004). UCSF Chimera--a visualization system for exploratory research and analysis. *Journal of Computational Chemistry*, *25*(13), 1605–1612. http://doi.org/10.1002/jcc.20084

Pieper, U., Webb, B. M., Barkan, D. T., Schneidman-Duhovny, D., Schlessinger, A., Braberg, H., … Sali, A. (2014). ModBase, a database of annotated comparative protein structure models, and associated resources. *Nucleic Acids Research*, *39*(suppl 1), D465–D474. http://doi.org/10.1093/nar/gkq1091

Punta, M., Coggill, P., Eberhardt, R., Mistry, J., Tate, J., Boursnell, C., … Finn, R. (2012). The Pfam protein families databases. *Nucleic Acids Res 40: D290-D301.*, *30*(1), 1–12. http://doi.org/10.1093/nar/gkp985

Punta, M., Love, J., Handelman, S., Hunt, J. F., Shapiro, L., Hendrickson, W. A., & Rost, B. (2009). Structural genomics target selection for the New York consortium on membrane protein structure. *J Struct Funct Genomics*, *10*(4), 255–268. http://doi.org/10.1007/s10969-009-9071-1

Remmert, M., Biegert, A., Hauser, A., & Söding, J. (2011). HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nature Methods*. http://doi.org/10.1038/nmeth.1818

Rost, B. (1997). Protein structures sustain evolutionary drift. *Folding and Design*, *2*, S19–S24. http://doi.org/10.1016/S1359-0278(97)00059-X

Rost, B. (1999). Twilight zone of protein sequence alignments. *Protein Engineering Design and Selection*, *12*(2), 85–94. http://doi.org/10.1093/protein/12.2.85

Rost, B., Casadio, R., Fariselli, P., & Sander, C. (1995). Transmembrane helices predicted at 95% accuracy. *Protein Science : A Publication of the Protein Society*, *4*(3), 521–533. http://doi.org/10.1002/pro.5560040318

Rost, B., & O'Donoghue, S. (1997). Sisyphus and prediction of protein structure. *Computer Applications in the Biosciences : CABIOS*, *13*(4), 345–56. Retrieved from http://www.ncbi.nlm.nih.gov/pubmed/9283749

Rost, B., & Sander, C. (1996). Bridging the protein sequence-structure gap by structure predictions. *Annual Review of Biophysics and Biomolecular Structure*, *25*, 113–136. http://doi.org/10.1146/annurev.biophys.25.1.113

Sabir, K., Stolte, C., Tabor, B., & O'Donoghue, S. I. (2013). The molecular control toolkit: Controlling 3D molecular graphics via gesture and voice. In *BioVis 2013 - IEEE Symposium on Biological Data Visualization 2013, Proceedings* (pp. 49–56). http://doi.org/10.1109/BioVis.2013.6664346

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., & Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. *Nucleic Acids Research*, *32*(Database issue), D449–D451. http://doi.org/10.1093/nar/gkh086

Sander, C., & Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. *Proteins*, *9*(1), 56–68. http://doi.org/10.1002/prot.340090107

Schaefer, M. H., Fontaine, J. F., Vinayagam, A., Porras, P., Wanker, E. E., & Andrade-Navarro, M. A. (2012). Hippie: Integrating protein interaction networks with experiment based quality scores. *PLoS ONE*, *7*(2). http://doi.org/10.1371/journal.pone.0031826

Schafferhans, A., Meyer, J. E. W., & O'Donoghue, S. I. (2003). The PSSH database of alignments between protein sequences and tertiary structures. *Nucleic Acids Research*. http://doi.org/10.1093/nar/gkg110

Schlessinger, A., Punta, M., Yachdav, G., Kajan, L., & Rost, B. (2009). Improved disorder prediction by combination of orthogonal approaches. *PLoS ONE*, *4*(2), e4433. http://doi.org/10.1371/journal.pone.0004433

Schneider, R. (1994). Sequenz und Sequenz-Struktur Vergleiche und deren Anwendung für die Struktur- und Funktionsvorhersage von Proteinen. thesis, Heidelberg University.

Shneiderman, B. (1992). Tree visualization with tree-maps: 2-d space-filling approach. *ACM Transactions on Graphics*. http://doi.org/10.1145/102377.115768

Silverman, B. (1986). *Density estimation for statistics and data analysis*. *Chapman and Hall* (Vol. 37). http://doi.org/10.2307/2347507

Skrabanek, L., Saini, H. K., Bader, G. D., & Enright, A. J. (2008). Computational prediction of protein-protein interactions. *Molecular Biotechnology*. http://doi.org/10.1007/s12033-007-0069-2

Slabinski, L., Jaroszewski, L., Rodrigues, A. P. C., Rychlewski, L., Wilson, I. A., Lesley, S. A., & Godzik, A. (2007). The challenge of protein structure determination--lessons from structural genomics. *Protein Science : A Publication of the Protein Society*, *16*(11), 2472–2482. http://doi.org/10.1110/ps.073037907

Smedley, D., Köhler, S., Czeschik, J. C., Amberger, J., Bocchini, C., Hamosh, A., … Robinson, P. N. (2014). Walking the interactome for candidate prioritization in exome sequencing studies of Mendelian diseases. *Bioinformatics (Oxford, England)*, *30*(22), 1–8. http://doi.org/10.1093/bioinformatics/btu508

Smith, T. F., & Waterman, M. S. (1981). Identification of Common Molecular Subsequences. *J. Mol. Biol.*, *147*, 195–197. http://doi.org/10.1016/0022-2836(81)90087-5

Stroud, R. M., Choe, S., Holton, J., Kaback, H. R., Kwiatkowski, W., Minor, D. L., … Harries, W. (2009). 2007 Annual progress report synopsis of the Center for Structures of Membrane Proteins. *Journal of Structural and Functional Genomics*. http://doi.org/10.1007/s10969-008-9058-3

Suhrer, S. J., Wiederstein, M., Gruber, M., & Sippl, M. J. (2009). COPS - A novel workbench for explorations in fold space. *Nucleic Acids Research*, *37*(SUPPL. 2). http://doi.org/10.1093/nar/gkp411

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., … Von Mering, C. (2011). The STRING database in 2011: Functional interaction networks of proteins, globally integrated and scored. *Nucleic Acids Research*, *39*(SUPPL. 1). http://doi.org/10.1093/nar/gkq973

Taylor, W. R., Chelliah, V., Hollup, S. M., MacDonald, J. T., & Jonassen, I. (2009). Probing the "dark matter" of protein fold space. *Structure (London, England : 1993)*, *17*(9), 1244–52. http://doi.org/10.1016/j.str.2009.07.012

The UniProt Consortium. (2014). Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Research*, *42*(Database issue), D191-8. http://doi.org/10.1093/nar/gkt1140

Travaglini-Allocatelli, C., Ivarsson, Y., Jemth, P., & Gianni, S. (2009). Folding and stability of globular proteins and implications for function. *Current Opinion in Structural Biology*. http://doi.org/10.1016/j.sbi.2008.12.001

Travis, J. (2002). Biological Dark Matter Newfound RNA suggests a hidden complexity inside cells. *SCIENCE NEWS-WASHINGTON-*, *161*(2), 24–25. article.

Tringe, S. G., & Rubin, E. M. (2005). Metagenomics: DNA sequencing of environmental samples. *Nature Reviews. Genetics*, *6*(11), 805–814. http://doi.org/10.1038/nrg1709

Uversky, V. N., Oldfield, C. J., & Dunker, a K. (2008). Intrinsically disordered proteins in human diseases: introducing the D2 concept. *Annual Review of Biophysics*, *37*, 215–246. http://doi.org/10.1146/annurev.biophys.37.032807.125924

Vastrik, I., D'Eustachio, P., Schmidt, E., Gopinath, G., Croft, D., de Bono, B., … Stein, L. (2007). Reactome: a knowledge base of biologic pathways and processes. *Genome Biology*, *8*(3), R39. http://doi.org/10.1186/gb-2007-8-3-r39

Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F., & Jones, D. T. (2004). Prediction and functional analysis of native disorder in proteins from the. *J Mol*

Biol, *337*(3), 635–45. http://doi.org/10.1016/j.jmb.2004.02.002

Wellmann, B. (2012). Evaluation of sequence-to-structure alignments. thesis, University of Munich.

Wilhelm, M., Schlegl, J., Hahne, H., Moghaddas Gholami, A., Lieberenz, M., Savitski, M. M., … Kuster, B. (2014). Mass-spectrometry-based draft of the human proteome. *Nature*, *509*(7502), 582–7. http://doi.org/10.1038/nature13319

Wright, P. E., & Dyson, H. J. (1999). Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. *Journal of Molecular Biology*, *293*(2), 321–31. http://doi.org/10.1006/jmbi.1999.3110

Wu, C. H., Yeh, L. S., Huang, H., Arminski, L., Castro-Alvear, J., Chen, Y., … Barker, W. C. (2003). The protein information resource. *Nucleic Acids Research*. http://doi.org/10.1093/nar/gkg040

Yachdav, G., Kloppmann, E., Kajan, L., Hecht, M., Goldberg, T., Hamp, T., … Rost, B. (2014). PredictProtein--an open resource for online prediction of protein structural and functional features. *Nucleic Acids Research*, *49*(Web Server issue), W337-43. http://doi.org/10.1093/nar/gku366