Dissertation
submitted to the
Combined Faculties for the Natural Sciences and for Mathematics
of the Ruperto-Carola University of Heidelberg, Germany
for the degree of
Doctor of Natural Sciences

presented by Janos Binder
born in Budapest, Hungary
Oral-examination: .....

# INTEGRATION AND VISUALIZATION OF SCIENTIFIC BIG DATA TO AID SYSTEMS BIOLOGY RESEARCH

Referees: Prof Dr. Ursula Kummer
Dr. Jan Korbel

# Contents

# Acknowledgement

First and foremost, I would like to thank my supervisor Dr. Reinhard Schneider for giving me the opportunity to work in his lab and his mentorship throughout my PhD. He provided me with all of the freedom to pursue my research interests and he introduced me to his peers, which resulted fruitful international collaborations. Without his support the work presented in this thesis would not have been possible.

I would like to also thank to my collaborator, Prof. Lars Juhl Jensen, who trained me building new scientific resources. Through his advice and through his guidance I have learnt how to drive a successful scientific project.

Special thanks go to Dr. Kiran Patil, who served in my Thesis Advisory Committee, and also accepted to take the role of my second supervisor later during my PhD. He provided a lot of valuable input and I enjoyed the interactions with the members of his lab.

Many thanks goes to Prof. Ursula Kummer and Dr. Dietrich Rebholz-Schuhmann for their useful feedback and for motivating me at the annual Thesis Advisory Committee meetings.

I am grateful to Dr. Seán O'Donoghue for his visualization-related advices and his entertaining stories. Moreover he introduced me to the VIZBI community and gave me the opportunity to be a co-organizer of that meeting. A lot of good memories are associated to Venkata Satagopam through the cheerful discussions and his continuous support.

I would like to thank to Dr. Bernd Klaus, Gary Male and Dr. Matt Rogon for providing a valuable feedback to the thesis.

I had a very pleasant and entertaining work environment, which would not have been possible without sharing a room with nice and happy colleagues. I thank for this my colleagues in Heidelberg namely Afshin, Bernd, Clemens, Frederico, Maria, Matt and in Copenhagen namely Kalliopi and Thomas.

Big thanks go to my friends at EMBL, with whom we shared memorable moments

and had a cheerful time in Heidelberg. I enjoyed their humor, support and scientific knowledge. I hope that we keep in touch as we progress through our carrier.
My life partner, Uschi provided emotional support during my time at EMBL, and helped me to keep the faith, when I faced difficult times. I thank for her encouragement and for the wonderful times we have spent together.
Last, but not at least I thank for my family for giving me the continuous support and the education, which gave me the skills to start working as a scientist.

# Summary

Information on protein subcellular and tissue localization is important to understand the cellular functions of proteins. However getting such information is not trivial; one needs to consult model organisms database, to evaluate the results of high-throughput experiments, to read the ever-increasing literature and to use prediction tools, when no previous knowledge on localization is available. Collecting and integrating the necessary information is tedious and difficult to do, and there is a clear need for evidence integration efforts. In my thesis I explored a new way of integrating and presenting localization evidence for the scientific community.

First I discuss the COMPARTMENTS resource, which I developed in collaboration to provide a comprehensive view on localization of proteins. This resource integrates the above-mentioned sources and maps the evidence to common protein and localization identifiers. In addition we developed a text-mining pipeline to find localization-protein associations from the scientific literature. To facilitate comparison of the different types and sources of evidence, we assigned a confidence scoring system to the localization evidence. To provide a simple overview we visualize the evidence on a schematic of a cell. Finally we link the evidence to its source to provide more details to the users.

Large-scale analysis using the COMPARTMENTS resource is also possible with the bulk download files. I have illustrated its usefulness by identifying pairs of compartments that share a statistically significant number of human proteins and by showing that protein-protein interaction networks can be used to infer protein localization of interacting partners.

Later I present the TISSUES resource, which integrates evidence on tissue expression. The resource presents the evidence the same way as COMPARTMENTS, however it integrates more high-throughput experimental datasets. My contribution was to create reusable components; I created a simple graphical overview based on the type and the confidence score of the evidence. I have also improved the text-mining of

human tissues by filtering the underlying localization keywords.
Finally I study integration on identifier level through the example of disease databases. Ontologies are useful in data integration, however not all of them provide the same quality. Therefore we created a modified version of the text mining pipeline to map entries from the Online Mendelian Inheritance in Man (OMIM) to the Disease Ontology (DO). Moreover we built a collaboration with the team behind the ontology and they use these mappings as a basis for the next version. Overall this thesis provides novel solutions for integrating biological data at different levels.

# Zusammenfassung

Das Wissen über die Subzelluläre- und die Gewebelokalisation der Proteine ist wichtig um die zelluläre Funktion der Proteine zu verstehen. Dennoch ist es nicht einfach solche Information zu erhalten; es ist notwendig in den Datenbanken der Modelorganismen nachzuschlagen, die Ergebnisse der Hochdurchsatz-Experimente auszuwerten, die ständig zunehmende Literatur zu lesen und Vorhersagewerkzeuge zu nutzen, wenn über die Lokalisation nichts bekannt ist. Die Sammlung und die Integration der nötigen Informationen ist eine mühsame und schwierige Arbeit, und es gibt einen großen Bedarf für die Integration bestehender Erkenntnisse. In dieser Doktorarbeit erforschte ich einen neuen Weg, Informationen über die Lokalisation zu integrieren und für die wissenschaftliche Gemeinschaft darzustellen.

Zuerst behandle ich in meiner schriftlichen Ausarbeitung die COMPARTMENTS Datenbank, die ich in einer Kollaboration entwickelt habe, um eine Gesamtübersicht über die Lokalisation der Proteine bereitzustellen. Diese Datenbank integriert die obengenannte Quelle und verbindet sie mit einem Protein und dessen Lokalisation. Außerdem haben wir eine Text-Mining Pipeline entwickelt, um Lokalisation-Protein Assoziationen aus der wissentschaftlichen Literatur zu sammeln. Die verschiedenen Typen und die Quellen der Informationen sind mittels eines Confidence Scoring System vergleichbar. Wir visualisierten die Ergebnisse der Lokalisation in einer schematischen Abbildung einer Zelle, um einen einfachen Überblick zu geben. Am Ende verbindeten wir die Ergebnisse mit deren Datenquelle, um weitere Informationen für den User anzubieten.

Die gesamte COMPARTMENTS Datenbank kann auch heruntergeladen werden, was komplexere Analysen ermöglicht. Als Anwendungsbeispiel, zeige ich, wie man mit Hilfe der Datenbank Organell-Paare identifizieren kann, die viele Proteine gemeinsam haben. Außerdem zeige ich, dass man die Lokalisation der Proteine anhand eines Protein-Protein Interaktionsnetzwerks ableiten kann.

Später stelle ich die TISSUES Datenbank vor, die die Information über die Gewe-

belokalisation integriert. Sie stellt die Daten ähnlich wie COMPARTMENT dar, aber sie integriert mehrere Hochdurchsatz-Experiment Datensätze. Ich entwickelte wiederverwendbare Komponenten, zum Beispiel eine einfache graphische Abbildung, basierend auf dem Typ und dem Confidence Score der Information. Ich verbesserte auch das Text-Mining des menschlichen Gewebes durch die Filterung der grundlegenden Lokalisations-Stichwörter.
Abschließend untersuche ich die Integration der Informationen anhand verschiedener Krankheitsdatenbanken. Die Ontologie ist nützlich für die Datenintegration, aber nicht jede Ontologie hat die gleiche Qualität. Deswegen modifizierten wir die Text-Mining Pipeline, damit die Einträge aus „Online Mendelian Inheritance in Man (OMIM)„ auf „Disease Ontology (DO)„ abgebildet werden können. Das Team, dass die Ontologie entwickelt hat und mit dem wir eine Kollaboration ausgebaut haben, wird die neue Abbildung in die nächste Version einbauen. Diese Doktorarbeit zeigt neuartige Lösungen für die Integration biologischer Daten auf verschiedenen Ebenen.

# Abbrevations

| | |
|---|---|
| API | Application Programming Interface |
| ATP | Adenosine triphosphate |
| cDNA | Complementary DNA |
| DNA | Deoxyribonucleic acid |
| DO | Disease Ontology |
| ER | Entity Recognition |
| GFP | Green fluorescent protein |
| GO | Gene Ontology |
| GOCC | Gene Ontology cellular component namespace |
| ICD | International Classification of Diseases |
| IE | Information Extraction |
| OMIM | Online Mendelian Inheritance in Man |
| ORF | Open reading frame |
| MeSH | Medical Subject Headings |
| NCI | National Cancer Institute |
| NLP | Natural language processing |
| PAGE | Polyacrylamide gel electrophoresis |
| PCR | Polymerase chain reaction |
| PIR | Protein Information Resource |
| SDS | Sodium dodecyl sulfate |
| SNOMED CT | Systematized Nomenclature of Medicine Clinical Terms |
| SVG | Scalable Vector Graphics |
| WWW | World Wide Web |

# List of Figures

# List of Tables

# 1 Introduction

## 1.1 Prologue

In biology we continuously extend our understanding of how life works. When one has a disease, we want to understand what, when and how it went wrong in the body. For example, when we realize that our elderly relative cannot remember recent events, we ask questions like: does he have an Alzheimer's disease, if yes how can we cure it?

Usually one or more little parts (gene, transcript, protein) of the system cease to work in our cells, and it results in a malfunction (usually a disease). Unfortunately we do not always know which little parts went wrong, thus we cannot cure every disease, because we do not understand the whole biological machinery yet. Nowadays our understanding of the biological system, while still limited, increases through data interpretation. To have a view on the current knowledge, we need to consult multiple sources of information such as scientific papers, datasets and results of experiments. Despite the ongoing efforts of life scientists, it is impossible to keep up with the ever growing quantity of biological information. Thus there is a need for further big data integration, where the science community filters, combines and visualizes the biomedical information, so it is easier to understand biological systems. My research focus was to provide a solution, by creating such an integrative effort for the community.

In this thesis I present a one-stop shop resource, COMPARTMENTS, for localization of small cellular parts called proteins in a cell, which I developed in collaboration. We integrated localization evidence from well-known biological databases, results of

high-throughput experiments and results of prediction tools for protein localizations. Moreover we developed computational text mining tool to extract protein–subcellular localization associations from the biomedical literature. Additionally we unified heterogeneous evidence, translated it to confidence scores, and visualized the accumulated evidence in a simple web interface.

Then I present my developed techniques could be used for integrating other types of biological data. As an example I will demonstrate my contribution to a tissue localization database, called TISSUES. Later I describe an automated linking of disease databases. These databases have been used to describe and annotate biological data, and I developed a technique to link entries in the various databases, which is later manually checked by collaborators. Finally, I discuss how the databases can be used to analyze large-scale biological data and how it can aid biological research.

## 1.2 Acquiring information about protein localization in a cell

A cell consists of small "organs" called organelles, and they are responsible for specific functions. For example in eukaryotes mitochondria supply the cell with a molecule (called *ATP*), which is the source of chemical energy, which is later used in other organelles to build new molecules and structures. Most of the proteins in the mitochondrion are involved in a chain of chemical reactions, which contributes to the larger goal; the production of the chemical energy. Thus determining the subcellular localization of a protein is a key step towards understanding its cellular function.

It is also known, that mislocalization of proteins occurs in diseases (Quintana et al., 2006; Magzoub and Miranker, 2012; Gutekunst et al., 1998; Zhang et al., 2005). One recent therapeutic effort aimed to specifically target tumor cell's organelles. (Agemy et al., 2013).

In this section I will give an overview how one can retrieve information on the protein of interest. First, I will introduce a few experimental methods to unravel the subcellular localization of proteins. Then I will talk about model organism databases, and how the information is stored. Later I will show a few sequence-based prediction methods, which are useful when there is no annotated data available.

### 1.2.1 Experimental methods for verifying localizations

Most of the proteins are synthesized in the *endoplasmic reticulum*, and many proteins are exported to other organelles in the cells. These protein have a short target peptide (usually 3-70 amino-acid long) in their sequence and it is used by the cellular transport machine to deliver the protein into the right compartment. The short target peptide is either located in the middle of the sequence or in the latter case on the end of the protein (C- or N-terminus). When the protein reaches its final destination, the target peptide is cleaved.

There are a few high-throughput methods developed during the last decade to unravel localization of proteins. One example is the large-scale cDNA sequencing projects by Ziauddin and Sabatini (2001); Ozawa et al. (2003); Simpson et al. (2000). Since the underlying ideas were similar, I describe the work by Simpson et al. (2000). In this study a cDNA library was generated and the gene of interest was fused with the gene of a fluorescent protein and cloning vectors are prepared for expression (Figure 1.1). During the fusion with the fluorescent proteins the target peptide sequence might be destroyed, thus two types of clones were generated by fusing the target protein either at the 5' end or at the 3' end with the fluorescent marker protein. When the vectors are expressed in cells, the proteins glow in specific organelles (Figure 1.2).

A similar approach was performed by Huh et al. (2003). They systematically tagged 6,234 open-reading frames (ORFs) through oligonucleotide-directed homologous recombination. Afterwards they combined each ORF in a plasmid, followed by GFP tag at the 5' end (which is fused to the C-terminus of the target protein after translation). Then they transformed haploid yeast strains and tested them using genomic PCR whether they have integrated at the appropriate locus the GFP tag and the specific ORF. Roughly 75% of the yeast proteome (4,156 proteins) expressed GFP signals above the background level. After evaluating the GFP signals, the proteins were classified into twelve subcellular localization categories. The authors suggest that the false positive rate of this method is low. A modified approach has been carried out by Kumar et al. (2002). Instead of marking the genes with GFP, they used an epitope sequence for the fusion with the ORF. Then they used immunostaining to analyze the localization of proteins. The authors tried another method as well; they used transposons to integrate the ORFs into the genomic DNA, however it turned out to be less effective.

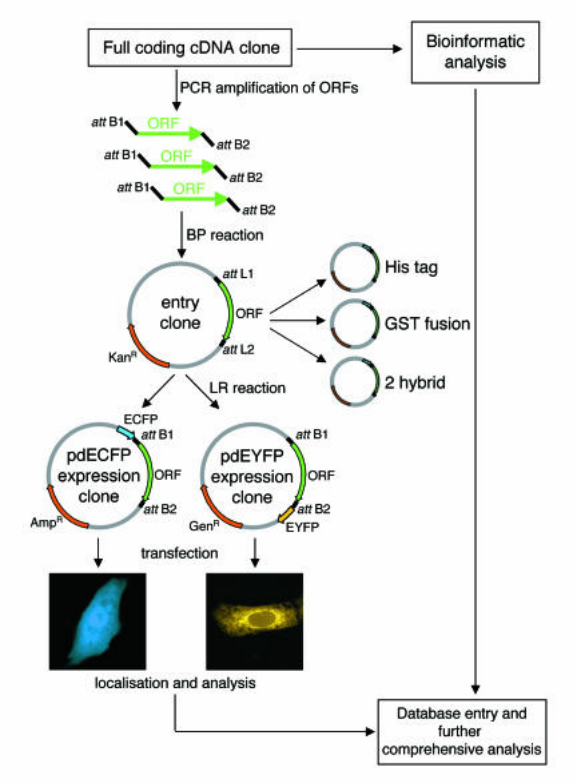



Figure 1.1: **Strategy for rapid systematic localization and functional characterization of proteins encoded by novel cDNA.**
The cDNAs encoding the proteins are amplified with designed primers and to the 5' and 3' ends *attB1* and *attB2* are also added. Later the products are recombined using the BP clonase into an entry clone, which functions as a base for expression vectors. Then a cyan and yellow fluorescent protein gene was recombined to the 5' and 3' end of the gene, respectively. These vectors were used to transfect cells and the localization of proteins were recorded. The results were later combined with data from bioinformatical analysis and other validation experiments (immunostaining) when it was appropriate. The figure has been taken from Simpson et al. (2000).

In proteomics studies (Andersen et al., 2002; Bell et al., 2001) first the organelles are collected, and the proteins are extracted and analyzed. These studies begin with the isolation of organelles through by sonication and centrifugation. Then the organelles are purified and the proteins are separated using (1D or 2D) SDS PAGE and identified using mass spectrometry. Peptide sequence tags were derived from the fragmentation spectra of the proteins and a peptide sequence tag search algorithm (Mann and Wilm, 1994; Küster et al., 2001) was used to identify proteins using the draft genome sequence. Andersen et al. (2002) could identify 271 nucleolar proteins and Bell et al.

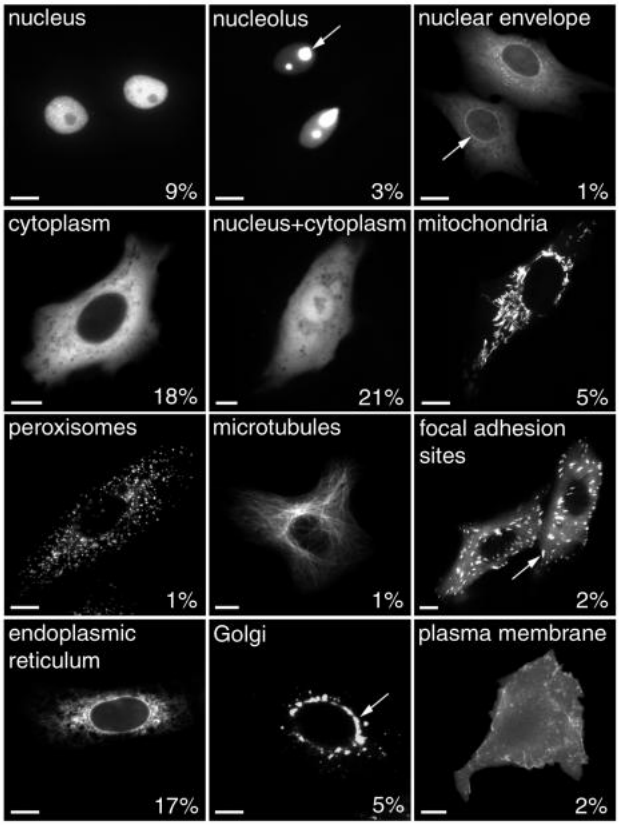



Figure 1.2: **cDNA-Cyan/Yellow Flourescent Protein fusion express and localize to a wide variety of intracellular compartments.**
The authors expressed the protein fusions originated from 107 entry clones in Vero cells and imaged the cells. This figure shows the localization of twelve categories, where the percentage refers to the part of individual expressed cDNA molecules in the shown categories. Golgi apparatus and plasma membrance (6%), other unknown localization (8%) and no expression (1%) are not shown. The figure has been taken from Simpson et al. (2000).

(2001) found 81 proteins localized in the Golgi.

Also verification experiments were necessary to validate the protocol (Andersen et al., 2002); for example the fractionation was controlled by embedding organelle in resin and examining the samples under an electron microscope. For further verifications the proteins are immunolabeled with antibodies and examined under a light-microscope. The protein contents of the unfractionated and the fractionated part is also compared using SDS-PAGE separation and probing them with specific antibodies. For the *in vivo* validation of these proteins Andersen et al. (2002) isolated the cDNA clones and

fused them with fluorescent proteins.

The Human Protein Atlas (HPA) (Uhlen et al., 2010) is an on-going effort to provide high-throughput experimental information on tissue and subcellular localization of proteins. The HPA team is generating antibodies against the whole human proteome. They use at least three cell lines for the immunoflourescence analysis of subcellular localization. Since 2010 (Uhlen et al., 2010) the project uses a new scoring scheme to reflect whether two or more antibodies stain the protein in the same compartment. The images of the experimental results are open to the scientific community. According to the release history, one of biggest challenges is the proper generation and staining of the antibodies.

### 1.2.2 Curated knowledge about subcellular localization

Biological databases play an important role in biology (Bourne, 2005). Model organism databases integrate a large part of the above mentioned high-throughput experimental datasets, and provide a summarized overview of proteins for the life science community such as: protein and gene names, their sequence and subcellular localization and their biological function. Highly qualified curators manually annotate proteins, which is tedious and expensive work (Poux et al., 2014), however it gives the resources a huge value. This includes reading the scientific papers and integrating the key points as annotations. For easier data integration and computational processing, these databases use common identifiers for proteins, and keywords for the annotations. They also import datasets from other web resources such as protein-protein interaction networks from STRING (Franceschini et al., 2013). In the next sections I will introduce some of the well-known model organism databases.

In 1965 Margaret (Oakley) Dayhoff published a book called *Atlas of Protein Sequence and Structure*, which was a collection of all known protein sequences at that time. The book was republished in several editions, and after her sudden death, the members of her former group have created the *Protein Information Resource (PIR)* (Wu et al., 2003) in 1984. Amos Bairoch created another database, *SwissProt* (Bairoch and Boeckmann, 1991) for protein sequences with different annotation priorities than PIR. During the upcoming years more sequence data was generated than SwissProt team could annotate, therefore they created *TrEMBL* (Bairoch and Apweiler, 1996), for those proteins, which were not curated in SwissProt. TrEMBL annotates proteins

automatically without human supervision. In 2002 teams behind PIR, SwissProt and TrEMBL united their efforts and created the *UniProt* Consortium.

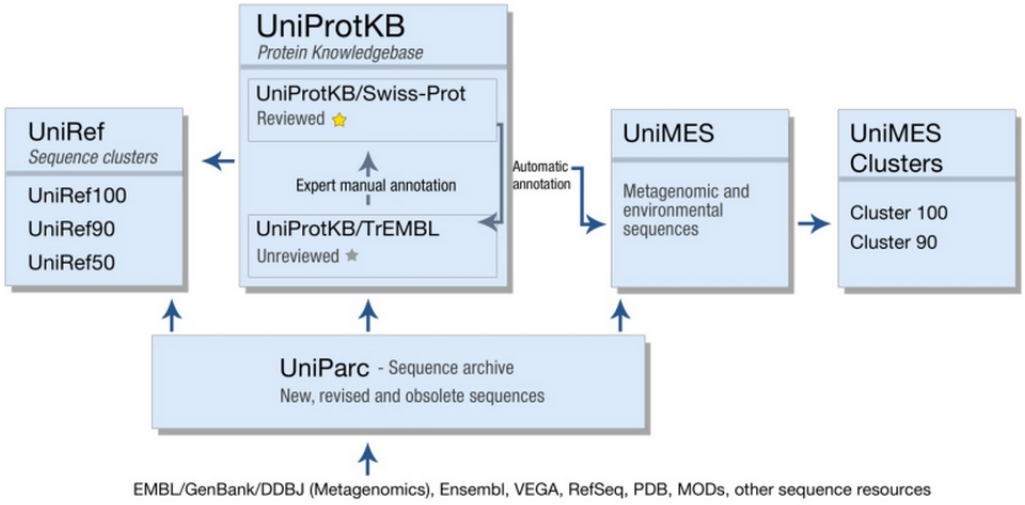



Figure 1.3: **Parts of UniProt databases.**
The UniProtKB collects functional information on proteins with annotations. The SwissProt part contains the high-quality manually annotated sequences, while sequences in TrEMBL are only analyzed and annotated computationally. Protein sequences might exist in various copies in the different databases, and UniParc imports them by filtering redundant copies and assigning them an unique identifier. UniMES focuses on environmental and metagenomic data, and UniRef databases are clustered sets of protein sequences from UniParc and UniProtKB. The figure has been taken from http://www.uniprot.org/help/about.

The *Universal Protein Resource (UniProt)* database stores protein sequences and their high-quality annotation data. The resource coverage is not limited to human proteins, but also covers other animal, plant, fungi, bacterial and viral proteins. The 2014_03 release of UniProtKB/SwissProt contains 542,782 manually annotated protein sequences, while TrEMBL stores 54,247,468 sequences. UniProt (Magrane and UniProt Consortium, 2011) integrates data from other resources such as MGI, SGD, FlyBase, WormBase (Eppig et al., 2011; Cherry et al., 2011; McQuilton et al., 2011; Harris et al., 2009) and it is extensively cross-referenced to dozens of external database (examples: STRING (Franceschini et al., 2013), EnsEMBL (Flicek et al., 2013), etc.) to add more and specialized information. To justify the annotations, they also provide the references to the underlying sources and to the scientific literature with a short description. The UniProt database is freely available for download in standardized file formats.

In 1997 the *Saccharomyces Genome Database (SGD) (Cherry et al., 1998; Costanzo et al., 2014)* was founded, which is a community database for budding yeast. It is the main information source for yeast researchers and stores a wide range of data such as yeast genome, proteome and other encoded features. Similarly to UniProt, it integrates high-quality manual annotations and results of high-throughput experiments. On top of that, they provide on their website a selection of bioinformatic tools to help the planning of experiments. Their data is also freely available for the scientific community.

The *Mouse Genome Database (MGI) (Blake et al., 1997)* was launched in 1996 to aid mice research by providing encyclopedic information about mouse genes and proteins. The mouse research is particularly important in biology, because it is used as a model of understanding human biology and diseases. It integrates similar types of evidence as UniProt and SGD, and it provides information about various mutations and their phenotypes. The website provides the possibility of ordering mice lines with the desired mutation. The other model organisms such as *Drosophila melanogaster* and *Caenorhabditis elegans* are covered by FlyBase (McQuilton et al., 2011) and WormBase (Harris et al., 2009) respectively. These databases are the central hubs of protein knowledge about these organisms and they use the same data structures as MGI and SGD.

### 1.2.3 Sequence-based prediction methods

Sometimes only the protein sequence is available, and there is no or very little annotation available. Fortunately using the sequence alone is enough to predict the localization of proteins. Imai and Nakai (2010) gave a good overview about the prediction techniques and methods. The protein sorting machinery of the cell secretes the protein to its final location using the information encoded in the sequences (e.g. target peptide). The first method was published by von Heijne (1986), where the method could predict the cleavage site of signal peptides. Afterwards more sorting signals were discovered and Nakai and Kanehisa (1992, 1991) developed PSORT, which has integrated the knowledge about sorting signals into one program. It could predict a few possible subcellular localizations based on the sequence and its type (whether it comes from an animal or a plant cell). Later dozens of prediction software were developed, which I will introduce in a later section.

How do these tools derive subcellular localization prediction from the sequence? They consider four different features according to Nakai (Imai and Nakai, 2010). First and the most obvious one is to look for the sorting signals, which makes the prediction feasible. Secondly they use homology information, because orthologous proteins share the ancestry thus it is very likely they are localized in same organelle. Indeed, Imai and Nakai (2010) used this feature alone as an objective test and assessed the power of the predictors by comparing it to a simple BLAST-based (Altschul et al., 1990) homology prediction. They used BLAST to compare sequences of proteins, where the subcellular localization was known. This simple method outperformed the predictors in redundant dataset (sequence similarity was allowed up to 90%). Third group of features are based on empirically correlated characteristics. These empirical features are based on the amino acid composition and captured by machine learning algorithms. Unfortunately these features do not always have a biological basis, however Andrade et al. (1998) attempted to give a biological explanation; for example extracellular proteins have more acidic surface residues, while nuclear proteins contains a large number of amino acids with basic surface residues. This is not suprising, because it helps to bind to the negatively charged phosphate backbone of the DNA. The last category of features are based on localization inference from external sources. These can be PPI data (Lee et al., 2008) or evolutionary information (Drawid and Gerstein, 2000). A limiting factor here is missing and incomplete information.

### 1.2.4 Tools for subcellular localization

Now I will explain how the well-known prediction methods work and incorporate these features.

One of the most well-known and oldest methods is PSORT (Nakai and Kanehisa, 1992, 1991), and it has been developed over the years (Bannai et al., 2002; Nakai and Horton, 1999; Horton and Nakai, 1997). The latest version is WoLF PSORT (Horton et al., 2007), which is accessible through a web interface and also available for download. It supports sequences from animals, plants and fungi, and the localization is based on sequence-derived signals (such as sorting signals and binding domains) and other features such as amino acid composition and length. A weighted $k$-nearest neighbor classifier uses these features in the prediction (Horton et al., 2006).

YLoc (Briesemeister et al., 2010a) is one of the state-of-the-art methods according to

the benchmarks (Goldberg et al., 2012; Briesemeister et al., 2010b). These methods uses naïve Bayes classifier that uses features like protein domains and motifs, GO annotations of close homologues and similar features as WoLF PSORT. Users can submit sequences from eukaryotes on their webpage, and they provide an explaination for their results to justify why the predicted compartment was favored and what were the biological reasons.

Another state-of-the-art method is LocTree2 (Goldberg et al., 2012), which can predict proteins of the three domains of life (Figure 1.4). They use a hierarchical system of support vector machine, which resembles the protein sorting machinery in cells. An online and newer version is available called LocTree3, which infers localization information by searching for the homologs with curated annotation. If there are no significantly similar homologs available, the method falls back to LocTree2.

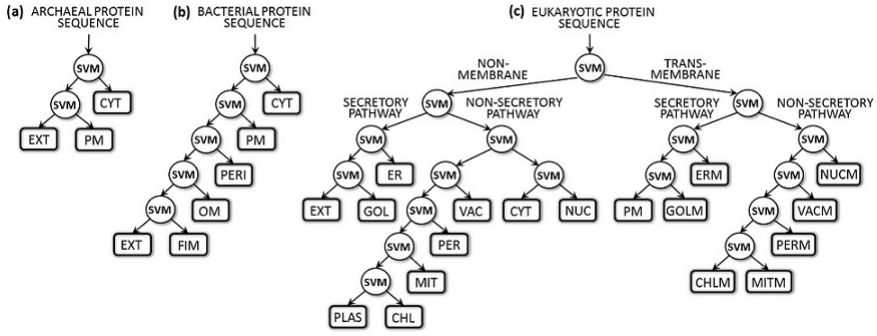



Figure 1.4: **Structure of LocTree2 prediction mechanism.**
The prediction architecture is different in archea, bacterial and eukaryotes respectively. Abbreviations: CHL, chloroplast; CHLM, chloroplast membrane; CYT, cytosol; ER, endoplasmic reticulum; ERM, endoplasmic reticulum membrane; EXT, extra-cellular; FIM, fimbrium; GOL, Golgi apparatus; GOLM, Golgi apparatus membrane; MIT, mitochondria; MITM, mitochondria membrane; NUC, nucleus; NUCM, nucleus membrane; OM, outer membrane; PERI, periplasmic space; PER, peroxisome; PERM, peroxisome membrane; PM, plasma membrane; PLAS, plastid; VAC, vacuole; VACM, vacuole membrane. The figure has been taken from Goldberg et al. (2012).

## 1.3 Integrating biomedical data

Combining biomedical data is not trivial. In this chapter I will describe, why ontologies are useful in integration and the basic concepts behind text mining.

### 1.3.1 Using ontologies in life sciences

It is well-known that many proteins are homologous and their functions and properties are shared across species. Thus there was a need to create a common "language" to describe the properties of proteins, therefore the GO Consortium (Ashburner et al., 2000) was founded more than a decade ago to develop controlled vocabulary for annotating proteins of eukaryotes and they decided to develop three ontologies, which I describe later.

The ontologies organize information into a hierarchy of concepts and the terms represent a domain of biological knowledge. For example the word *leg* represents an anatomical entity, but it can have another meaning in a different domain, a *leg part of* an object (e.g. *legs of a chair*). The anatomical entity, *leg* can be divided into a hierarchy, because *foot* is a *part of leg* and *toe*s are part of *foot* and *leg*. The *leg* is also a *limb* like an *arm*. Thus the hierarchy of term can be represented in a *Directed Acyclic Graph (DAG)*. The various types of relationships of the terms are also stored in the ontology e.g. *part of* or *is a*. Ontologies are extensively used, because they are:

**focused on different domains of biological knowledge** The first three ontologies (Ashburner et al., 2000) describe cellular components, biological processes and molecular functions. Later new ontologies were born to describe diseases (Schriml et al., 2011), chemical entities with biological interest (Degtyarenko et al., 2008; Hastings et al., 2013), plant anatomy and development (Avraham et al., 2008), human tissues (Gremse et al., 2011) and many others, those are available at the OBO foundry homepage (Smith et al., 2007).

**continuously developed by the life science community.** Every year new ontologies are published (Smith et al., 2007) and the users can submit revisions (Ashburner et al., 2000) as our knowledge about biology grows.

**using a hierarchy, where different types of relationships between the terms are allowed.** The links define how does one term relates to another such as *part of or is a* links. Some ontologies allow more fine-grained links such as *(positively/negatively regulates)*.

**linked to other resources.** The terms in the ontologies are mapped to external databases (e.g. Schriml et al. (2011)) to enlarge their scope. These mappings enables better

data integration and computational processing of biological datasets.

**processed easily by computational methods.** The controlled vocabulary and unique identifiers in the ontologies allow easy computional processing (Huang et al., 2009a,b; Huntley et al., 2009; Binns et al., 2009; Supek et al., 2011) . The tree-like structure of ontologies can be used for automated similarity measures.

**stable.** The identifiers are kept over the releases, and if an identifier is no longer used, it is regarded obsolete and it is de-linked from the other terms.

The most used ontology is probably the Gene Ontology (Ashburner et al., 2000), which covers three non-overlapping biological ontologies. The parts of a cell and its surronding are described by the *cellular component* ontology, *molecular function* terms represent enzymatic and molecular activities, while *biological process*es refer to a biological goal to which the protein or a gene contributes, such as cellular, tissue or organism level functions. These ontologies also describe relationship between terms(Consortium, 2010) such as *is_a* ,*part_of*, *has_part* or *(positively/negatively) regulates* (see Figure 1.5). The *is_a* defines, whether one term is a subtype of another one (e.g. *mitochondria* is an *organelle*) and if a term is a subset of another term the *is_part* relationship is used (e.g. *mitochondria* is a part of the *cell*). There are relationships on inter-ontology level, for example one or more *molecular functions* contribute to a *biological process* (Consortium, 2010; Hill et al., 2008) and the full version of Gene Ontology includes these links. There are other stripped down versions of the Gene Ontology: one which does not include the inter-ontology and the *has_part* links (which is the complement type of *is_a* relationship). Another version called GO Slim (Harris et al., 2004) even more reduced, because it does not contain the very detailed terms.

Curators use GO terms to annotate proteins in the model organism databases (Eppig et al., 2011; Cherry et al., 2011; McQuilton et al., 2011; Harris et al., 2009). These annotations come with source information, which can be literature reference, database reference or computational evidence (Smith et al., 2007; Consortium, 2004). They also include an evidence code, which is based on an experiment method, computational method or a statement in the literature. There were various studies to evaluate the effectiveness of the evidence codes (Jones et al., 2007; Skunca et al., 2012), or assigning a metric (Engelhardt et al., 2005; Buza et al., 2008) to them. There are also good

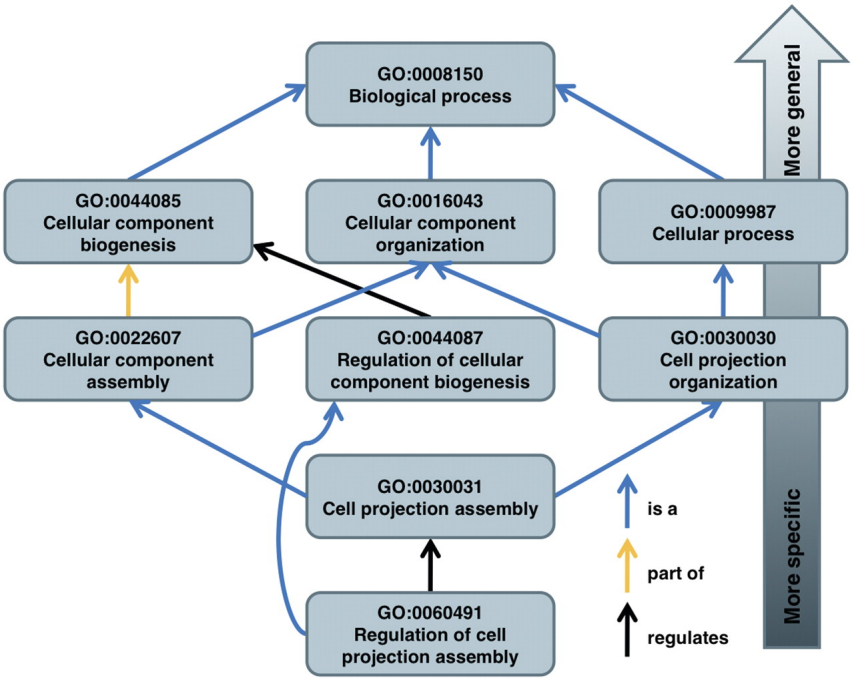



Figure 1.5: **The structure of the Gene Ontology.**
The ontology can be represented in a Direct Acyclic Graph (DAG), where the terms are the nodes and the relationship are the edges. The edges point to more general terms. The figure has been taken from du Plessis et al. (2011).

overviews available about the Gene Ontology by du Plessis et al. (2011); Jensen and Bork (2010).

In 2009 the community-driven Disease Ontology (Osborne et al., 2009; Du et al., 2009; Schriml et al., 2011) was created to integrate data on human inherited, developmental and acquired diseases. The ontology classifies diseases into a single ontology by unifying various vocabularies and terminologies. The ontology consists of currently eight main categories to represent: diseases of anatomical entity, cellular proliferation, infectious agent, mental health, and metabolism along with genetic diseases, medical disorders and syndromes. The terms are organized into a hierarchic tree-like structure by going from the more general diseases to specific ones. The terms integrate and are mapped to biomedical databases of ICD-9, ICD-10, MeSH, NCI's thesaurus, OMIM disease specific set (Amberger et al., 2011), and SNOMED CT.

The BRENDA Tissue Ontology (Gremse et al., 2011) was started in 2004 to create a classification of enzyme sources, which later became a structured encyclopedia

of tissue terms. The ontology currently covers anatomical structures, cell lines and types, and tissues from fungi, plants and higher level animals. It also includes disease-related and cancerous tissues. The terms are organized in a hierarchy, from the more general to the more specific terms, and it has three additional relationship types besides the general *is_a* and *part_of* relations. The *develops_from/derives_from* relationship shows from what was the tissue or cell was developed (e.g. myoma cell *develops_from/derives_from* muscle fibre), while the *related_to* relationship represent similar tissues (e.g. electroplax is *related_to* muscle fibre, the electroplax, which is a specialized muscle fiber, exists in electric eel and it generates a high electric potential instead of a contraction).

There is an ever-growing amount of environmental and metagenomics samples. Ecosystems biology researchers are interested in the physical environment of lifeforms. An interesting initiative is the Environmental Ontology (Buttigieg et al., 2013), which aims to classify environmental concepts. The ontology supports the annotation with a wide-range of terms including man-made objects to support the environment of remote measurement devices (such as *terrarium* or *hypodermic needle*). It categorizes the terms into the following classes: *environmental systems*, which has a placement in the environment ranging from oceans to gut lumens. Its subtypes; *biome* class represent ecologically similar communities of plants, animals and other organisms and *habitat* class refers to spatial region having environmental qualities which may sustain an organism or a community of organisms. There are also single entities (e.g. *human gut* or *coral reef*) in the ontology, which have a strong effect on the surronding space. Entities (e.g. *seamount*) belonging to causal "hubs" are described by *environmental features*, while masses and volumes (such as *soil*) are described by *environmental material* terms and climate conditions are covered by *environmental condition* branch.

The GO Evidence Codes (Schneider et al., 2009; Carbon et al., 2009) will be replaced later by Evidence Ontology terms. The ontology is still under development by the founders of GO and it will allow a more fine-grained evidence annotation. It distinguished *evidence*, which is “a type of information that is used to support an assertion” and *assertion method*, which is defined as “a means by which a statement is made about an entity” (whether the claim was made by a human or a computer).

### 1.3.2 Introduction to text mining

Despite the curators' very valued work of annotating databases, it is not possible to fully process the ever-growing scientific literature. Fortunately automated text-mining solutions can help to complement the work of human curators. These methods are capable of processing million of pages in a matter of hours and derive valuable information from it. The work of Soldatos (2009); Jensen et al. (2006); Rebholz-Schuhmann et al. (2012) gives an overview about the basics of text mining and I will reiterate it here.

Defining the scope(*corpus*) is important for successful mining of the literature. A naïve approach would be processing of all scientific literature available by today, however there might be no central repository available for every scientific research areas, thus it would be impossible to get every scientific paper. The currently available 23 million papers from biological and biomedical research are indexed by PubMed, only 10% is open-access, and the researcher institution needs to agree with the publishers to access subscription-only papers. Recently the funding bodies require researchers to publish their results under an open-access license and the latest developments allow better text mining of full text papers(Cohen et al., 2010). Fortunately abstracts can be download from PubMed and they provide a wealthy resource for text mining (Kuhn et al., 2014; Franceschini et al., 2013; Pafilis et al., 2013a). One can even further limit the scope to a set of documents from the scientific literature by using PubMed to search for a single biological term (e.g. gene). This querying step is called *information retrieval*. Another interesting source is the electronical health records (Blair et al., 2013; Eriksson et al., 2014). Mining this data can unravel unknown disease correlations, however the access is limited due to legal issues, and their language is different from scientific literature (Jensen et al., 2012).

Once the corpus has been selected, one needs to find biomedical entities that are mentioned within the text such as disease names, genes, proteins, species names etc. The *Entity Recognition (ER)* consists of two sub-tasks: first, recognizing names that refer to entities, and second, the unique identification of entities and mapping them unambiguously to database records. Let us have a look at an example sentence: "Alternatively, insufficient survival factor signaling (for instance inadequate levels of **interleukin-3** in lymphocytes or of **insulin-like growth factor 1/2** [**Igf1/2**] in epithelial cells) can elicit apoptosis through a **BH3**-only protein called **Bim**." (Hanahan

and Weinberg, 2011). A goal of ER would be to recognize all gene and protein names marked boldface in the sentence and map them to a widely used database identifier such as UniProtKB (Magrane and UniProt Consortium, 2011) accession numbers.

**Entity Recognition**. The ER is a very tough task, although it seems trivial at first glance. Early methods employed manually devised rules to look for specific postfix endings like *-ase* (Chang et al., 2004) as well as searching for nearby words like "gene" or "protein". These rules looked for typical characteristics of gene or protein names. However most of the modern methods use machine learning algorithms to identify protein or gene names based on their characteristic features.

Another approach is to create a *dictionary* based on a comphensive list of synonymous biomedical terms (e.g. IGF-1, IGFI, IGF1A, or Insulin-like growth factor I refer to the same protein). The terms in the dictionary will be matched against the documents. Algorithms are used to allow matching against the variations of the names e.g. IGF-1 or IGF1 will be both matched if dash symbol is ignored. One clear advantage that the terms are directly linked to a database reference in the dictionary, so the term recognition and the mapping is solved in one step. The dictionary can be extended further to increase the efficiency of Entity Recognition; using orthographical expansion with various rules such as converting Roman number to their Arabic counterpart and Greek letters to their latin counterparts. Abbreviating, permuting when possible, creating the pluralized form and removing unnecessary parts of the terms can also help (Pafilis et al., 2013a). There are several difficulties in Entity Recognition due to the lack of standardization of names. Some of the most common issues are:

1. Each protein or gene has usually more names e.g. Somatomedin-C is a synonym for Insulin-like growth factor I.
2. Some gene names are also used as common English words e.g. *homeless, crocodile*.
3. Abbreviated chemical terms might also mean protein names e.g. SDS.
4. Gene names that refer to unrelated genes in organisms e.g. Cdc2 in budding and in fission yeast.

To address this disambiguation methods have been developed (Gerner et al., 2010; Hanisch et al., 2005; Gaudan et al., 2005; Schijvenaars et al., 2005; Hur et al., 2009;

Pafilis et al., 2013a). Algorithms can also guess the right term by looking at the context (Tanabe and Wilbur, 2002; Fukuda et al., 1998). Another approach would be to create a *stopword* list to ignore ambiguous terms (Glenisson et al., 2003).

### Scoring important terms in the corpus

The terms occur in various quantities in the text, the occurrence of a term (in a sentence, paragraph or in the full text) usually reflects how important the term is. Therefore many methods employ a scoring scheme which is proportional with the occurrence. Let us assume we have downloaded thousands of papers from Pubmed and are looking for the most relevant articles about "the Parkinson's disease". A starting point could be filtering out all articles, which do not contain all three words of "the", "Parkinson's" and "disease", but there would be still many documents left. If we count the number of times a terms occurs (called *term frequency*) in an article and then we summarize the counts. The method will incorrectly emphasize the articles with many occurrences of "the", because of its abundance in the English language. Therefore articles will not get enough weight with the meaningful "disease" or "Parkinson's" terms. "the" is clearly a bad term to distinguish non-relevant and relevant documents. Therefore many text mining methods incorporate an *inverse document frequency* factor to penalize terms, which occur very frequently in a set of articles. Numerically the *term frequency–inverse document frequency (tf-idf)* is a product of two statistics, the *inverse document frequency* and the terms frequency. There are various ways to calculate the *tf-idf* statistic, here I give one example:

$$tf(t_i, a) = \frac{f(t_i, a)}{max(f(t, a))} \tag{1.1}$$

Where $f(t_i, a)$ is the count of term $t_i$ and $max(f(t, a))$ is the count of the most occuring term in article $a$. The $max(f(t, a))$ reduces the bias towards the terms the longer documents, where the terms might receive a higher count. The following equation shows the *inverse document frequency* is measured.

$$idf(t_i, A) = log\left(\frac{N}{1 + \{a \in A : t_i \in a\}}\right) \tag{1.2}$$

$N$ denotes the total number of articles in the corpus and $\{a \in A : t_i \in a\}$ is the number of articles, where $t_i$ term was mentioned. The denominator is adjusted by adding 1 to

avoid division-by-zero if $t_i$ term does not exist in the corpus.

In many dictionary based methods $t_i$ can refer to an entity, which incorporate other synonyms. One should also consider, that if a term occurs too often, it should be ignored by adding it to stopwords.

### Information extraction from the corpus

Information Extraction (IE) aims to capture pre-defined types of fact – here relationship of biological terms. Going back to the example sentence, an IE system should capture that "**interleukin-3** or **insulin-like growth factor 1/2** protein interacts with **Bim**, which protein is necessary for the **apotosis** of the cell." To deduce relationships from biomedical literature, two fundamentally different solutions can be employed; co-occurence and natural-language processing (NLP).

**Natural Language Processing**. Natural Language Processing methods can extract facts from text such types of pathway interaction (Friedman et al., 2001) and protein-protein interactions (Šarić et al., 2006). These tools incorporate the analysis of syntax and semantics of the natural language (Rebholz-Schuhmann et al., 2012); first text is converted to 'tokens' by finding the boundaries of sentences and words, then the sentences are lexically analyzed to find part of speech tags (such as nouns, verbs and adverbs). Then the syntax tree is obtained to delineate nouns phrases (such as "insulin-like growth factor 1/2") and represent their relationship. Afterwards entity recognition methods are employed to semantically tag biomedical entities (for example, diseases and proteins) and other keywords (such as apotosis, cell division). In the last step sets of rules are used to extract relationships on the basis of syntax tree and the semantic labels. The standard English text and the language of biomedical literature is very different, thus limiting the applicability of general NLP parsers (particularly, the use of long, complex noun phrases). Another difficulty is to resolve intersentence relations linked by anaphoric relationships (e.g. use of the word 'it' or 'this' to refer to something stated before in the text). Luckily majority of the relationship are mentioned in one sentence (Ding et al., 2002). NLP methods are computationally intensive and a good curation of dictionaries and rules are needed (Jensen et al., 2006).

**Co-occurence**. The principle of the method is very simple, it looks for co-occuring entities within the same paragraph or abstracts, however many systems rely on co-

occuring entities only in the same sentence. If the two entities occur often together it means that they are somehow related, although the relationship is not known. Therefore many solutions employ a frequency-based scoring scheme to rank the co-occuring entities, because two-entities can be mentioned together without being related somehow. The methods based on co-occurrence can capture more relationship than NLP method, thus they usually yield a higher recall, however some of the relationship tends be false therefore the precision is worse. Co-occurence analysis is very versatile and it can be used to support new hipothesis generation (Ananiadou et al., 2006; Kell and Oliver, 2004).

One can limit co-occuring methods to harvest relationship for a specific type only, for example unraveling protein-protein interaction networks by looking for protein names (von Mering et al., 2005). Also it is possible to look for two different entity types such as mining protein-chemicals interactions (Kuhn et al., 2008, 2014). These relationships can be used for annotation basis in databases.

#### Mentioning a few tools for text mining

There are several well-known and widely used tools, which employ various aspects of the methods mentioned above. A starting point for literature search and probably the most used website for information retrieval is PubMed (or one of the derivative sites such as HubMed (von Isenburg, 2007) or UK PubMed Central (McEntyre et al., 2011), however these sites have difficulties dealing with abbrevations or synonyms. Upon query GoPubMed (Doms and Schroeder, 2005), searches for a term and by default its synonyms as well in the documents. The results are organized into a hierarchical structure where the results are categorized under concept labels from the Gene Ontology and from the MeSH thesaurus. By using this feature the user can easily navigate through the documents, thus reduced time is necessary for browsing through the search results. WhatIzIt (Rebholz-Schuhmann et al., 2008) is a biomedical entity recognition service, which can perform entity recognition on a text. The user can choose what kind of dictionary should be used (GO term or SwissProt terms, disease names etc.). Reflect (Pafilis et al., 2009; O'Donoghue et al., 2010) allows real-time tagging of webpages in the user's web browser, and protein, chemical and Wikipedia terms are highlighted. Clicking on the terms opens up a small popup window, with detailed information about the tagged term. OnTheFly (Pavlopoulos

et al., 2009; Pafilis et al., 2013b) converts Excel, Word and PDF documents to tagged web pages using the Reflect engine while keeping the layout. It provides advanced functionality such as the interaction network of proteins and chemical mentioned in the text. Several text-mining methods have been developed to automatically extract localization information from the biomedical abstracts (Cheng et al., 2008; Müller et al., 2004; Van Auken et al., 2012).

### 1.3.3 Integrating data from various sources

The different types and sources of information from high-throughput experimental data, model organism databases, sequence-based prediction methods and literature mining are complementary, it is important to combine them all to see the big picture. However, such a task is not trivial. The different experimental datasets and databases use various file formats and employ different identifiers and names for the same biomedical terms (proteins, subcellular localizations, diseases, tissues etc.) (Taylor, 2007). Therefore one needs to define a standard for assembling an integrated database by mapping data on common identifiers and keywords (Sterk et al., 2006). Many prediction tools (for example the sequence-based prediction tools for subcellular localization) have different web interfaces, the results consist of scores that are not directly comparable, and for genome-wide analyses one needs to generally hold a local installation of the software. Therefore it is thus time-intensive and difficult to gather and evaluate evidence (pertaining to involvement in diseases or to the subcellular localization) of a protein of interest. It is even more difficult when one needs to collect evidence on a large number of proteins.

Several one-stop shop resources on proteins and genes have been developed. NextProt (Lane et al., 2012; Gaudet et al., 2013) was initiated by the SwissProt/UniProtKB group and in addition to the information available in UniProtKB high-throughput datasets were integrated (such as high-quality proteomics, siRNA and 3D experimental, pathway datasets, population-related variant information and data of protein-protein and protein-drug interaction) and the web front-end support advanced powerful queries. GeneCards (Belinky et al., 2013) automatically mines and imports data from more than a hundred sources, resulting in a web-page for each of the human gene entries. Their goal is to integrate the fragments of information into GeneCards scattered over in specialized databases. MalaCards (Rappaport et al., 2013) has a similar architecture

as GeneCards, but the resource is focused on diseases. It is a search-able database of human diseases and their annotations, and the underlying information is derived from more than forty resources.

There were several attempts to address subcellular localization data integration according to Binder et al. (2014). DBSubLoc (Guo et al., 2004) was an early effort, which integrated annotations from knowledge bases such as SwissProt (Bairoch and Boeckmann, 1991), PIR (Wu et al., 2003) and the other major model organism databases.

Later on the databases begin to incorporate other types of data; manual annotations in eSLDB (Pierleoni et al., 2007) were complemented by sequence-based predictions, then experimental datasets were integrated into LOCATE (Sprenger et al., 2007), locDB (Rastogi and Rost, 2010), and SUBA3 (Tanz et al., 2012). Unfortunately the latest versions of the first three of these databases (DBSubLoc, eSLDB, and LOCATE) are more than five years old, and they seem to be abandoned resources. Because of the lack of updates they cannot be considered to reflect the current evidence. Although locDB and SUBA3 have been updated since 2011; however, between them these resources have limited species coverage by focusing on only on human and *Arabidopsis thaliana* proteins. In comparison the previously mentioned resources have been collecting subcellular localization evidence from multiple sources and depositing them into a single database, they generally do not solve the issue of putting the various types of evidence on a common confidence scale. One exception is the SUBA3 resource which covers exclusively *A. thaliana* proteins, and it assigns an overall confidence score; but, it is a tough task to trace these scores back to their source for the user.

# 2 Results

## 2.1 Unification and visualization of biological data

During my PhD, I developed an up-to-date one-stop shop for protein subcellular localization information in collaboration with Lars Juhl Jensen's group in Copenhagen. Also I contributed to a similar resource about tissue localization. What sets these resources apart is that they unify heterogeneous evidence, translate it to confidence scores, and visualize the accumulated evidence in a simple web interface. The underlying evidence is regularly updated based on annotations from manually curated knowledgebases, data from high-throughput screens, automatic text mining, and sequence-based predictions.

### 2.1.1 COMPARTMENTS: an unified subcellular localization evidence database

In this section I present the COMPARTMENTS (http://compartments.jensenlab.org) web resource, which is a slightly modified version of my paper (Binder et al., 2014). We have developed an automatically updated web resource to be able to provide up-to-date information on the subcellular localization of proteins from the major eukaryotic model organisms. In addition to integrating manually curated annotations, experimental data, and predictions, we use automatic text mining to extract associations from the biomedical literature. Unlike earlier resources, we address the challenge of making evidence comparable across types and sources by introducing a unified confidence scoring scheme. To further shield users from the heterogeneity of the

many evidence sources, we map all localization evidence onto Gene Ontology terms and visualize the combined results on an interactive schematic of a cell. All data is freely available for download to facilitate large-scale analyses. I further demonstrated that the dataset can be used for large scale analyses, and I showed that protein-protein interaction networks can be used to predict subcellular localizations.

### The COMPARTMENTS web resource

COMPARTMENTS holds subcellular localization information for 22,705 human and 6,696 yeast proteins and covers also other eukaryotes such as fruit fly, mouse and *C. elegans*. When querying the database for a protein of interest, the user is presented with an interactive schematic of a cell. These figures are color-coded according to the confidence of the evidence supporting each of eleven (twelve in case of plants) labeled compartments (Figure 2.4). Interactive tables provide the user with more fine-grained localization information and the source of the underlying evidence (Figure 2.1 and 2.2).

To provide a unified overview as described above, we map protein identifiers from the source databases to their corresponding identifiers in the STRING database (Franceschini et al., 2013), which for organisms in question come from Ensembl (Flicek et al., 2012). We similarly map all cellular compartments to their respective Gene Ontology cellular component terms (Ashburner et al., 2000). The labeled compartments are a subset of broad GO terms, much like GO Slims (Harris et al., 2004).

We further assign a confidence score to each piece of evidence to reflect that not all types and sources of localization information are equally reliable. To clearly signify that these should not be overinterpreted as probabilities, we use a scoring scheme that ranges from 1 star (lowest confidence) to 5 stars (highest confidence). The way that confidence scores are assigned varies between evidence channels as explained in the next section. The confidence scores are also the basis for the color-coding of the figures (Figure 2.3 and 2.4): the higher the confidence, the darker the shading of the compartment.

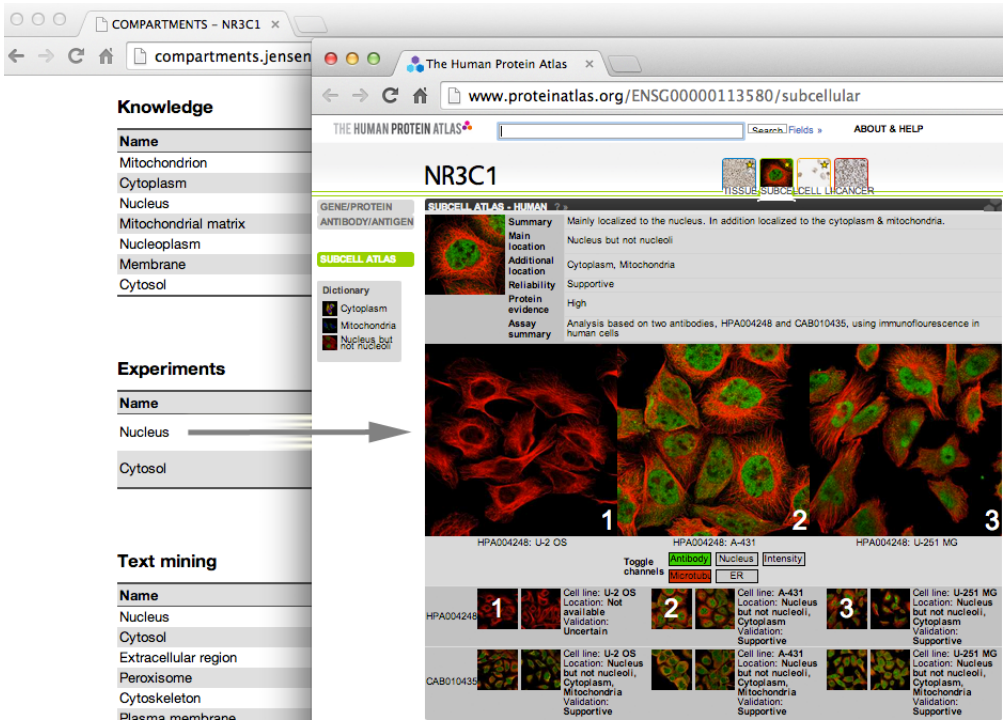
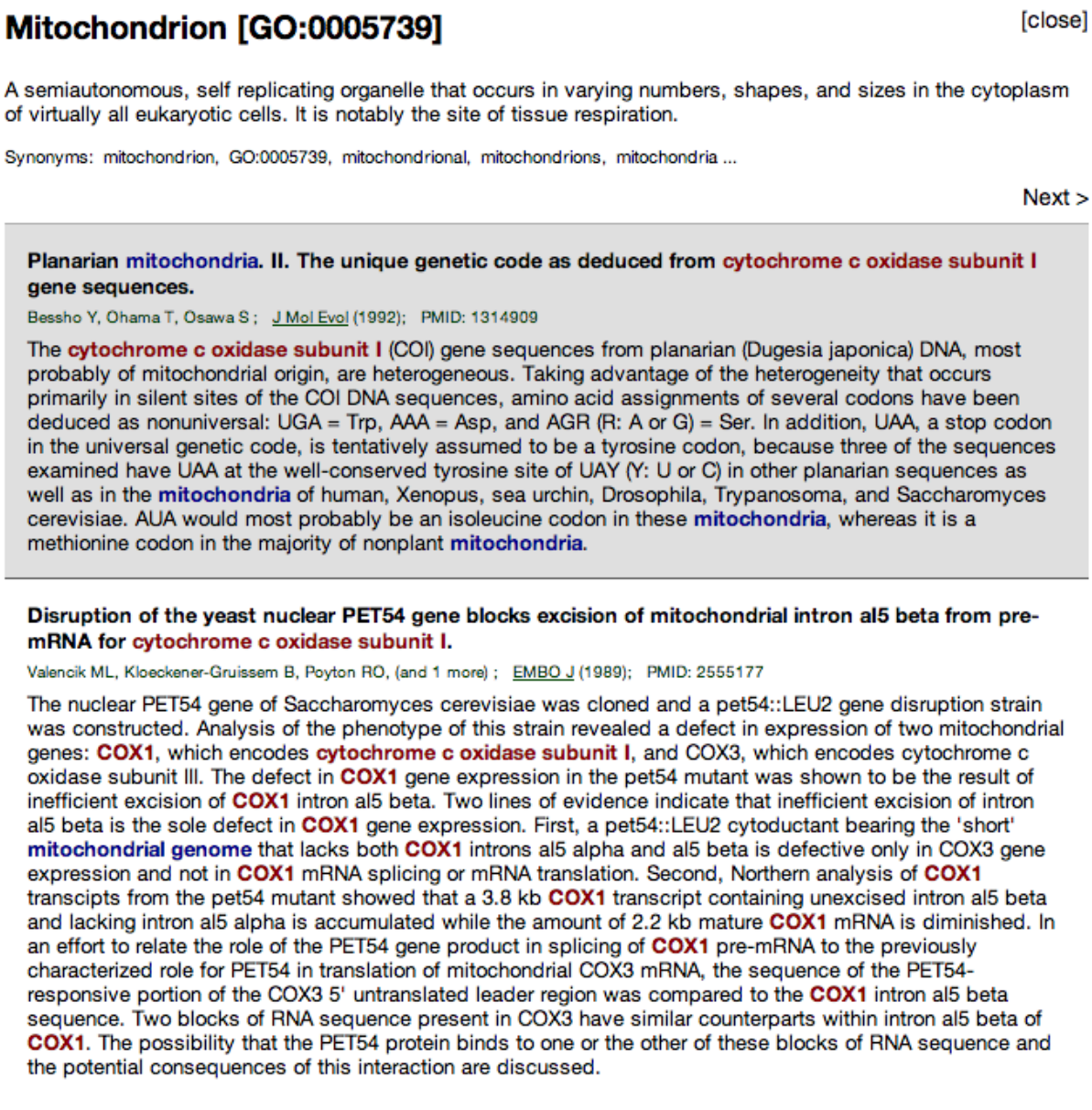



Figure 2.1: **Supporting information for underlying evidence.**
Clicking on evidence of interest, the user is directed to source database, where the evidence comes from. This feature allows the user to gather quickly specific information about the protein of interest.

### Evidence channels and sources

The evidence contained in COMPARTMENTS is logically partitioned into four channels. The first channel, called knowledge, is based on annotations from UniProtKB (Magrane and UniProt Consortium, 2011; UniProt Consortium, 2010) MGI (Eppig et al., 2011), SGD (Cherry et al., 2011), FlyBase (McQuilton et al., 2011) and WormBase (Harris et al., 2009). We assign confidence scores to these annotations based on the associated GO evidence codes (Schneider et al., 2009; Carbon et al., 2009), which encode whether the annotation is based on a peer-reviewed publication, an experimental dataset, or sequence similarity etc. (see Methods). The *knowledge* channel provides localization information on a total of 16864 human and 5909 yeast proteins.

The Human Protein Atlas (HPA) (Li et al., 2012) is an ongoing effort to experimentally validate the tissue expression and subcellular localization for the entire set of human proteins. The latter data is captured by the *experiments* channel and currently contains



**Mitochondrion [GO:0005739]** [close]

A semiautonomous, self replicating organelle that occurs in varying numbers, shapes, and sizes in the cytoplasm of virtually all eukaryotic cells. It is notably the site of tissue respiration.

Synonyms: mitochondrion, GO:0005739, mitochondrional, mitochondrions, mitochondria ...

Next >

**Planarian mitochondria. II. The unique genetic code as deduced from cytochrome c oxidase subunit I gene sequences.**

Bessho Y, Ohama T, Osawa S ; J Mol Evol (1992); PMID: 1314909

The **cytochrome c oxidase subunit I** (COI) gene sequences from planarian (Dugesia japonica) DNA, most probably of mitochondrial origin, are heterogeneous. Taking advantage of the heterogeneity that occurs primarily in silent sites of the COI DNA sequences, amino acid assignments of several codons have been deduced as nonuniversal: UGA = Trp, AAA = Asp, and AGR (R: A or G) = Ser. In addition, UAA, a stop codon in the universal genetic code, is tentatively assumed to be a tyrosine codon, because three of the sequences examined have UAA at the well-conserved tyrosine site of UAY (Y: U or C) in other planarian sequences as well as in the **mitochondria** of human, Xenopus, sea urchin, Drosophila, Trypanosoma, and Saccharomyces cerevisiae. AUA would most probably be an isoleucine codon in these **mitochondria**, whereas it is a methionine codon in the majority of nonplant **mitochondria**.

**Disruption of the yeast nuclear PET54 gene blocks excision of mitochondrial intron al5 beta from pre-mRNA for cytochrome c oxidase subunit I.**

Valencik ML, Kloeckener-Gruissem B, Poyton RO, (and 1 more) ; EMBO J (1989); PMID: 2555177

The nuclear PET54 gene of Saccharomyces cerevisiae was cloned and a pet54::LEU2 gene disruption strain was constructed. Analysis of the phenotype of this strain revealed a defect in expression of two mitochondrial genes: **COX1**, which encodes **cytochrome c oxidase subunit I**, and COX3, which encodes cytochrome c oxidase subunit III. The defect in **COX1** gene expression in the pet54 mutant was shown to be the result of inefficient excision of **COX1** intron al5 beta. Two lines of evidence indicate that inefficient excision of intron al5 beta is the sole defect in **COX1** gene expression. First, a pet54::LEU2 cytoductant bearing the 'short' **mitochondrial genome** that lacks both **COX1** introns al5 alpha and al5 beta is defective only in COX3 gene expression and not in **COX1** mRNA splicing or mRNA translation. Second, Northern analysis of **COX1** transcipts from the pet54 mutant showed that a 3.8 kb **COX1** transcript containing unexcised intron al5 beta and lacking intron al5 alpha is accumulated while the amount of 2.2 kb mature **COX1** mRNA is diminished. In an effort to relate the role of the PET54 gene product in splicing of **COX1** pre-mRNA to the previously characterized role for PET54 in translation of mitochondrial COX3 mRNA, the sequence of the PET54-responsive portion of the COX3 5' untranslated leader region was compared to the **COX1** intron al5 beta sequence. Two blocks of RNA sequence present in COX3 have similar counterparts within intron al5 beta of **COX1**. The possibility that the PET54 protein binds to one or the other of these blocks of RNA sequence and the potential consequences of this interaction are discussed.

Figure 2.2: **Text mining evidence viewer.**
The protein and subcellular localization terms are highlighted in the Medline abstracts. Therefore the user can see, which articles support the predicted protein-subcellular localization association in the text-mining evidence channel.

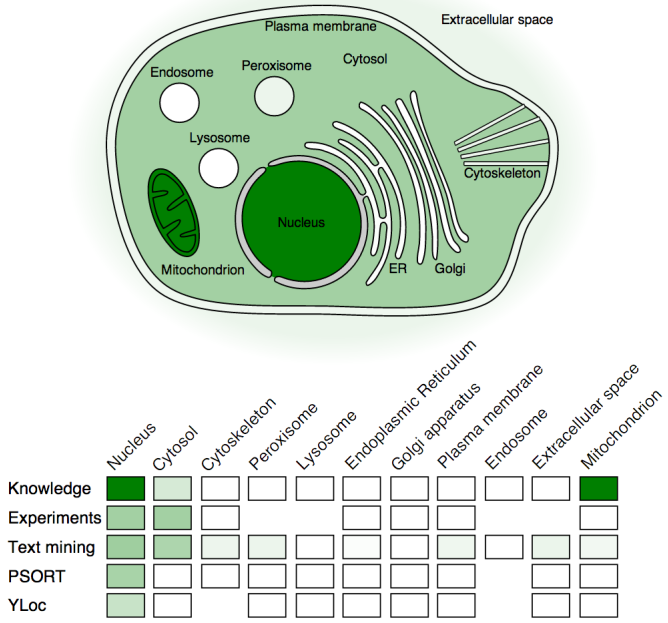



Figure 2.3: **Visualization of localization evidence.**
When querying the database for a protein, its localization is visualized on a schematic of a cell (top). We also graphically summarize the types of evidence supporting each compartment (bottom). The confidence of the evidence is color coded, ranging from light green for low confidence to dark green for high confidence. White indicates an absence of localization evidence. I created these figures, and they were used in a previous version of COMPARTMENTS.

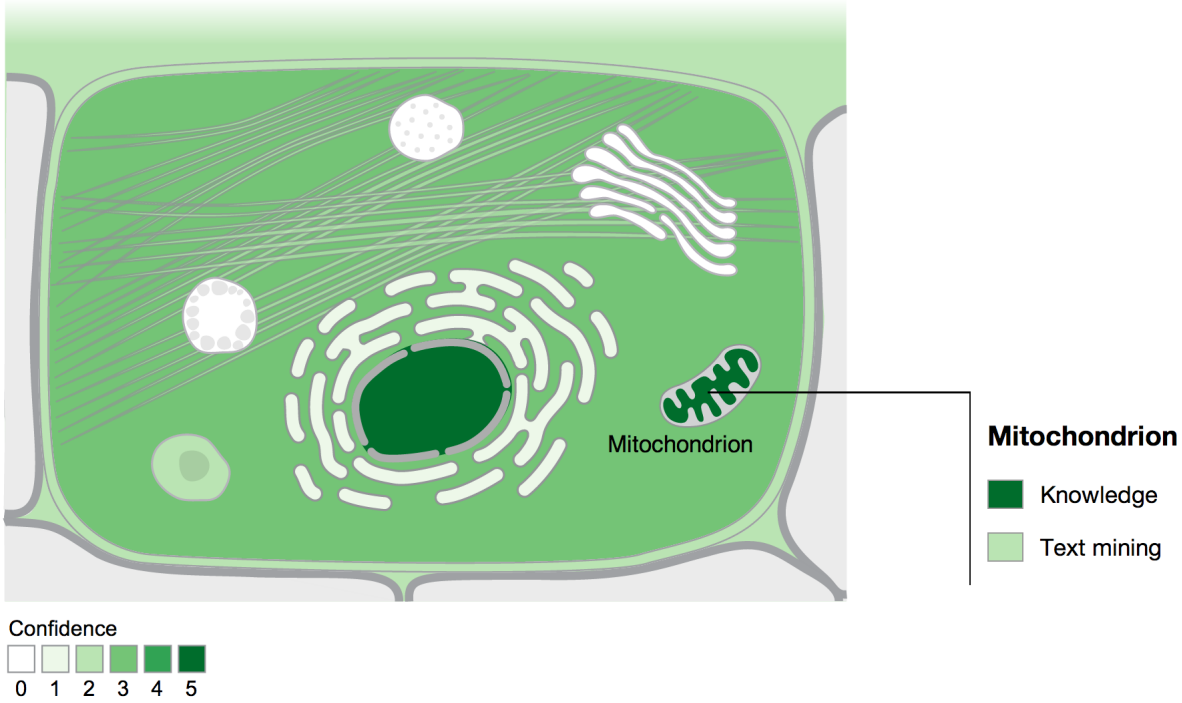



Figure 2.4: **New visualization of localization evidence.**
My collaborators, Seán I. O'Donoghue and Christian Stolte created a better figure. Besides the better look, another upgrade is, when the user hovers the cursor over a compartment, we also graphically summarize the types of evidence supporting this localization.

information on 9306 human proteins. The confidence scores of this channel are based on the antibody validation scores provided by HPA (Uhlen et al., 2010) (see Methods).

The third channel provides associations between proteins and subcellular localizations derived from automatic text mining of the abstracts in Medline. We used the dictionary of protein names from STRING (Franceschini et al., 2013) and created a dictionary of subcellular compartments from Gene Ontology (see Methods). We use a confidence scoring scheme, which is based on the fact that the more a protein and a cellular compartment are co-mentioned, the more likely the protein is to be localized to the compartment (see Methods). The *text-mining* channel currently contains putative localizations for 15304 human and 4144 yeast proteins.

Finally, the *predictions* channel contains pre-computed results from two sequence-based prediction methods, namely the well known WoLF PSORT (Horton et al., 2007) and the high-resolution version of YLoc (Briesemeister et al., 2010b,a). Published benchmarks (Goldberg et al., 2012; Briesemeister et al., 2010b) suggest that these methods are two of the best that cover many compartments, in particular for human proteins. Moreover, these and the other methods mentioned earlier were developed on overlapping training sets and thus cannot be considered independent evidence. The primary reason for including only two methods is thus to not present the user with a large number of redundant predictions. We applied both methods to 22523 human, 23443 mouse, 22938 rat, 14076 *Drosophila melanogaster*, 20158 *Caenorhabditis elegans*, 6697 *Saccharomyces cerevisiae*, and 31280 *Arabidopsis thaliana* protein sequences from STRING 9.1 (Franceschini et al., 2013). The output scores from each tool were transformed to make them comparable to other evidence in the database (see Methods).

The number of human and yeast proteins assigned to each of the eleven labeled compartments based on each of these evidence channels are summarized in Tables 2.1 and 2.2, respectively. The two sequence-based prediction tools both provide full coverage of the proteome and are therefore shown separately in the tables. For this reason we also leave out the prediction tools in Figure 2.5, which shows the overlap in terms of human proteins assigned to at least one compartment by knowledge, experiments, and text mining. This shows that integrating experimental and text-mining evidence increases the coverage by 11% additional human proteins. Even when more than one channel covers the same protein, this is not necessarily redundant information. Firstly,

the same protein can localize to multiple compartments, and the two evidence channels may not provide support for the same localization. Secondly, when two channels support the same localization of a protein, they typically provide complementary evidence of interest to the user. This is also why full coverage of the sequence-based prediction tools does not make the other evidence channels redundant; if a protein is predicted to have a certain localization, it is still of interest to the user if this also is supported by experiments or literature.

| | Knowledge | Experiments | Text mining | PSORT | YLoc |
|---|---|---|---|---|---|
| Nucleus | 6082 | 5848 | 2288 | 9600 | 5335 |
| Cytosol | 2538 | 4872 | 577 | 9128 | 4630 |
| Cytoskeleton | 1843 | 1215 | 1257 | 134 | — |
| Peroxisome | 124 | — | 240 | 315 | 262 |
| Lysosome | 386 | — | 262 | 5 | 120 |
| Endoplasmic reticulum | 1382 | 151 | 656 | 281 | 178 |
| Golgi apparatus | 1250 | 814 | 348 | 64 | 313 |
| Plasma membrane | 4440 | 1271 | 1515 | 3681 | 3815 |
| Endosome | 170 | — | 88 | — | — |
| Extracellular space | 2267 | — | 1528 | 4331 | 1625 |
| Mitochondrion | 1156 | 924 | 793 | 2008 | 871 |

Table 2.1: **Overview of the localization evidence for human proteins.**
We counted protein-compartment associations separately for each of the eleven labeled compartments and for each evidence channel. The only exception is the predictions channel, for which we show the results from the two sequence-based methods (PSORT and YLoc) separately. Dashes denote compartments for which a channel or prediction method cannot provide evidence.

### Benchmark of the text-mining pipeline

To assess the quality of the pairs extracted by text mining, we compared them against a benchmark set of 9,764 human and 3,834 yeast proteins having 12,232 and 4,530 high-confidence localization annotations, respectively. The benchmark set is derived from the evidence in the knowledge channel (see Methods). This shows that the method works well on the majority of the compartments Figure 2.6. The exceptions include the nucleus and – in case of human – the plasma membrane. The false positives for these compartments are predominantly due to functional associations captured by co-mentioning. For example, a protein involved in signal transduction can easily be functionally associated with both the plasma membrane and the nucleus without

| | Knowledge | Text mining | PSORT | YLoc |
|---|---|---|---|---|
| Nucleus | 2194 | 211 | 3870 | 1476 |
| Cytosol | 422 | 42 | 3242 | 1533 |
| Cytoskeleton | 231 | 108 | 44 | — |
| Peroxisome | 69 | 65 | 20 | 127 |
| Vacuole | 268 | 88 | 0 | 23 |
| Endoplasmic reticulum | 486 | 129 | 42 | 38 |
| Golgi apparatus | 236 | 75 | 12 | 57 |
| Plasma membrane | 457 | 135 | 775 | 350 |
| Endosome | 16 | 18 | — | — |
| Extracellular space | 94 | 69 | 302 | 624 |
| Mitochondrion | 1118 | 162 | 1486 | 422 |

Table 2.2: **Overview of the localization evidence for yeast proteins.**
For details refer to the caption of Table 2.1.

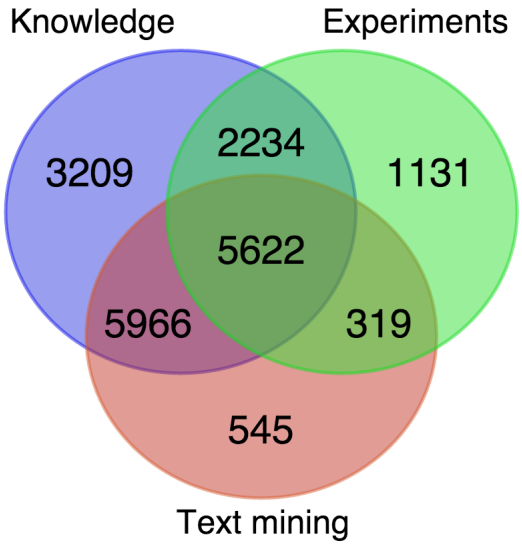



Figure 2.5: **Overlap between the knowledge, experimental, and text-mining evidence for human proteins.**
The Venn diagram shows the number of proteins with localization evidence from one or more of the three types of evidence. The two sequence-based prediction methods are not included as they are able to provide a prediction for any protein sequence.

being localized to either. The method also shows poor performance for the cytosol due to the experimental difficulty to distinguish proteins in the cytosol from those in, for example, vesicles. Consequently, many cytosolic proteins are conservatively annotated to the cytoplasm instead of the cytosol.

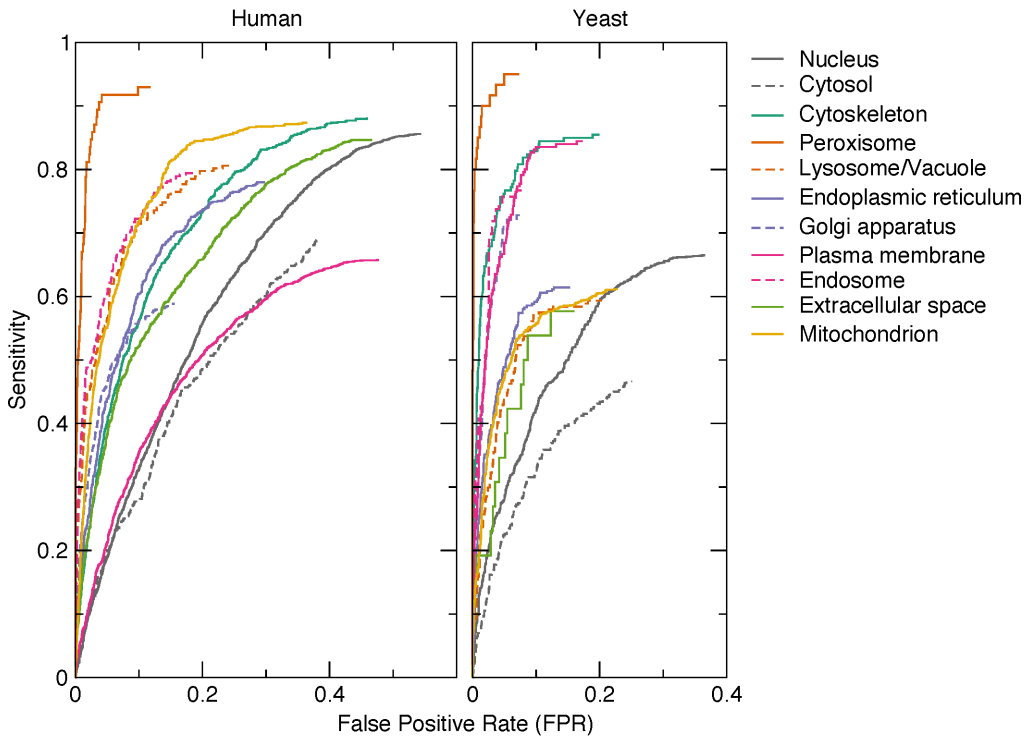



Figure 2.6: **Benchmark of text-mining results.**
The performance of the text-mining pipeline on human and yeast proteins is shown as Receiver Operating Characteristics (ROC) curves for each of eleven compartments. The curves do not intercept sensitivity = 1.0 and FPR = 1.0, because many of the protein-compartment pairs in the benchmark set are never found mentioned together in Medline, for which reason they have no text-mining score.

#### Linking compartments by overrepresentation of shared proteins

To illustrate the usefulness of COMPARTMENTS for large-scale studies, we identified pairs of compartments that share a statistically significant number of human proteins (Figure 2.7; see Methods). Notably, there were no borderline cases – all pairs of compartments were either highly significant after controlling for multiple testing, or they were not even significant before correction. The two compartments that share the most proteins are the cytosol and the nucleus, both of which also share many proteins with the cytoskeleton. Most of the remaining intracellular compartments form a highly connected network, except the extracellular space, the mitochondria, and the peroxisomes.

#### Using protein-protein interaction networks to predict subcellular localization

We assume, that protein-protein interaction could happen mostly in the same compartments due to the proximity of the molecules. To evaluate whether the hypothesis is true, we used the human protein-protein interaction networks from STRING 9.1 (Franceschini et al., 2013). We benchmarked it against a high-confidence annotations human protein annotations, which was derived from the knowledge channel. The results supported our hypotheses (Figure 2.8), however it is not surprising, because proteins in the same compartment can interact. One should take account that protein-protein interaction networks in STRING are also inferred by GO annotations from model organism databases. The initial version of the analysis was provided by Alberto Santos Delgado.

### 2.1.2 Improving the results in the TISSUES database

COMPARTMENTS (Binder et al., 2014) was the first resource, where we combined curated knowledge, high-throughput experimental results and text-mined data from the literature. However from the first moment, we were interested to integrate other biological data other than subcellular localization. To reduce the time of development a similar resource, we have built a general pipeline called *Mamba* framework. This effort also helped to easily maintain the resources.

The resource share this pipeline are: DISEASES (http://diseases.jensenlab.org), ORGANISMS (Pafilis et al., 2013a) and TISSUES (http://tissues.jensenlab.org). Many

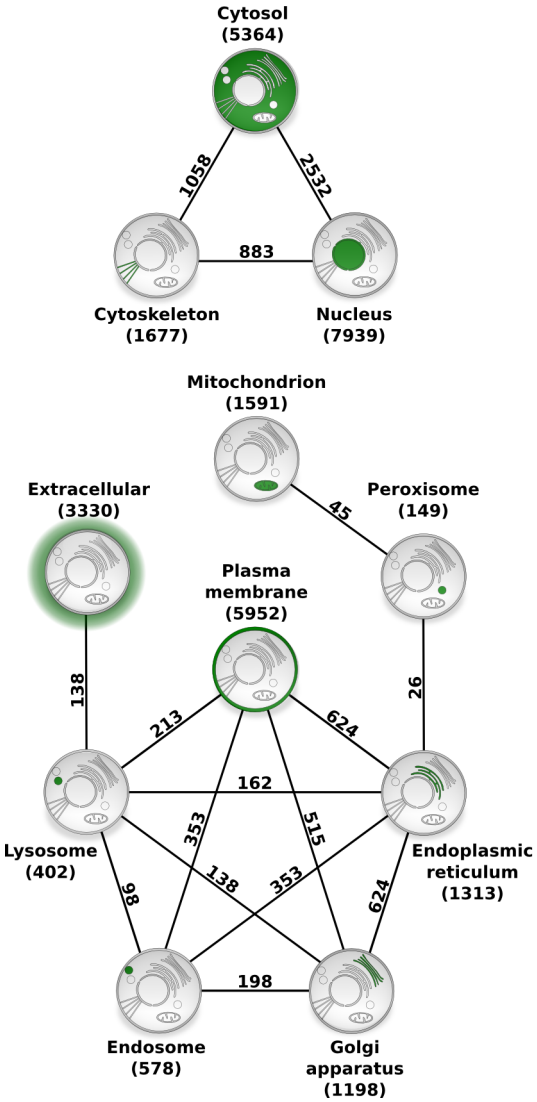



Figure 2.7: **Compartment relationships derived from shared proteins.** Illustrating the usefulness of COMPARTMENTS for global analysis of protein localization, we studied relationships between compartments. Each node represents a single compartment, which is highlighted in green. The number of proteins in the compartment is shown in parenthesis. We show an edge between two compartments whenever they share more proteins than expected at random (false discovery rate $< 0.1\%$). The number of proteins co-localized to the two compartments is shown next to the edge.

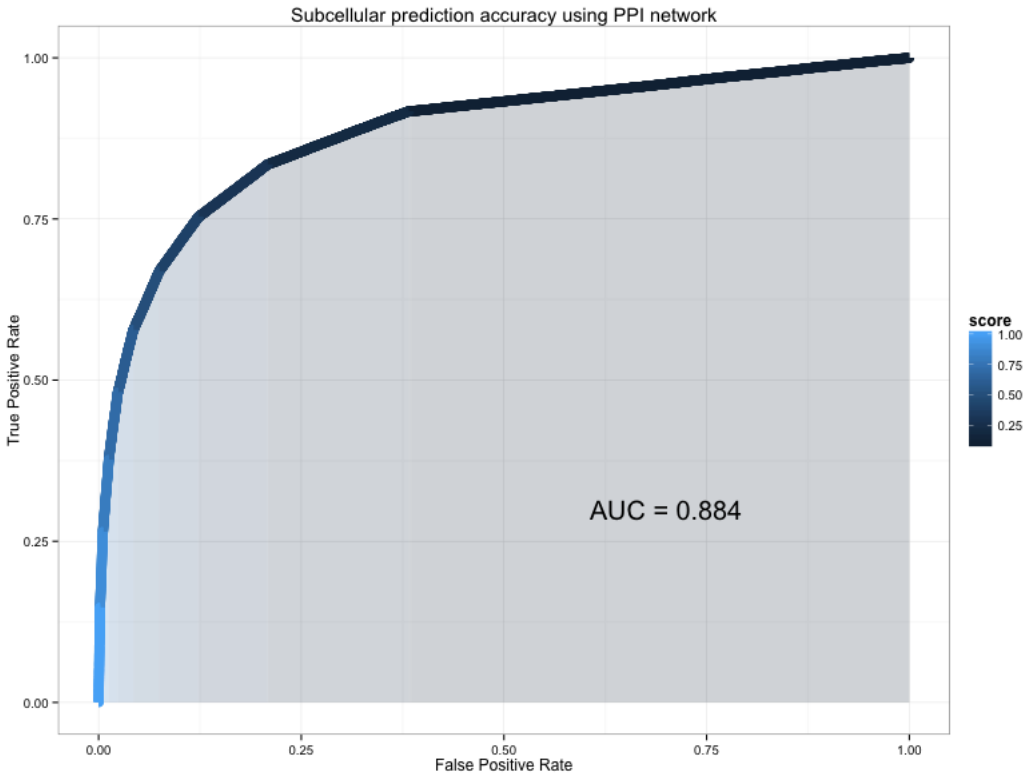



Figure 2.8: **Benchmark of protein-protein interaction to find subcellular localizations.**
The Receiver Operating Characteristics (ROC) shows the performance of predicting subcellular localizations from human protein-protein interaction network from STRING. *AUC* denotes the *Area Under Curve*, which is the probability that the prediction is correct. The predicted subcellular localizations predicted were scored with the number of interaction partners in a compartment divided with the number of interactions the protein had.

of the features are common in the background, but they are presented as a distinct services to the user. For example the text-mining engine contains all dictionaries for subcellular localizations, diseases, organism names and tissues and we run them regularly on PubMed to text-mine associations for all the resources, but we show only the respective results for each of the resources. Below I outline my contributions to this project.

### A general coloring mechanism for figures

Visualization is a main part of our resources to effectively communicate the scientific message with our users. Therefore we created a schematic of cell for COMPARTMENTS and a schematic of a human body for TISSUES. To support various devices, resolutions and zooming levels without loss of quality, we created the templates in SVG, which is a vector-based graphical file format. This also allowed us to easily visualize the location of every protein (see Methods). Therefore I developed a module, which can color the templates for every protein. Examples can be seen in Figure 2.3, Figure 2.4 or in Figure 2.9. For faster performance, the results can be stored in an SQL database, and the colored templates can be easily embedded on a webpage.

### Visualizing the contribution of different evidence channels

COMPARTMENTS and TISSUES resources integrate heterogeneous evidence, and it is important to our users to track the type and source of the evidence. Thus I created an overview figure to present the existence of an evidence in the main categories and visualize the different sources of evidence. Small squares are ordered in a table by larger structures and they are colored based on confidence scores. Below the figure, the user is able to retrieve more detailed information summarized in tables.

### Ensuring organism specific text-mining results in TISSUES

We extensively use ontologies in our resources, because they allow different levels of annotation. Also we found a few mature ontologies, we used as the basis for the dictionaries for text mining. In TISSUES we use the BRENDA Tissue Ontology (Gremse et al., 2011), which differentiate many tissues and allows broad and narrow tissue annotations for the proteins. To our knowledge, the terms are not categorized into

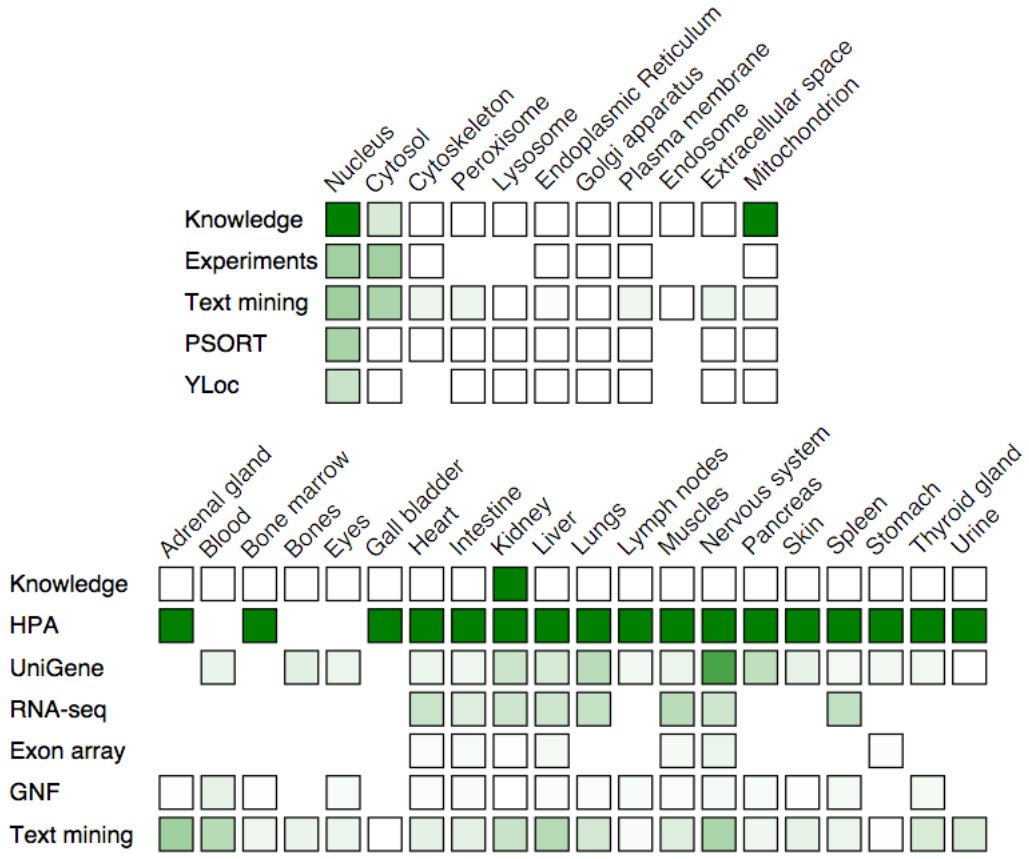



Figure 2.9: **Localization of human glucocortoid receptor, NR3C1 in COMPARTMENTS and TISSUES.**
Through the example of NR3C1 we demonstrate the effectiveness of the overview figure. The user can see the evidence channels and the broad localization categories. The darkness of the green coloring correlates to the strength of the evidence. This information provides an overview to the user, and down on the webpage he/she can get a detailed information on the evidence.

species in the current version, therefore many human proteins were falsely associated with non-human tissues by the text-mining engine on the Medline abstracts. Therefore I went through the BRENDA Tissue Ontology and assigned to every term in which model organisms do they exists. By rigorously annotating them, we have on first hand human proteins are only associated with human tissues, and on the other hand we made sure that the resource can be easily extended to support other organisms than humans.

## 2.2 Using ontologies in data integration

The ontologies describe various biological properties and they are used in the annotation of model organisms (Eppig et al., 2011; Harris et al., 2009; Magrane and UniProt Consortium, 2011; McQuilton et al., 2011) since the first release of Gene Ontology (Ashburner et al., 2000). The ontologies are structured into a direct acyclic graph, which allows to track back narrower to broader annotations (e.g. *mitochondrial inner membrane* is a part of *mitochondrion*). During the last decade other ontologies have been created, where the more mature ones are listed by OBO Foundry (Smith et al., 2007). We used ontologies in our projects such as the Gene Ontology (Ashburner et al., 2000), the BRENDA Tissue Ontology (Gremse et al., 2011) or the Disease Ontology (Schriml et al., 2011). Disease Ontology was very effective to text mine the literature, however it is not used for annotation purposes. We decided to set up a collaboration with Disease Ontology team to extend the ontology and make it a more mature resource. First we added missing diseases, then we linked the entries to another database.

### 2.2.1 Introducing a mapping pipeline to Disease Ontology

Currently, there are 1,388 OMIM pages mapped to 749 DO terms. We have development an automated mapping pipeline (see Figure 2.10) to link 3,095 OMIM pages (Amberger et al., 2009) to 829 DO (Schriml et al., 2011) terms. The pipeline is made general in order to make it applicable for other ontologies. My collaborators went through the provided mapping file and used them as a bases for manual curation. By linking the entries to OMIM, we could also add new genetic diseases to Disease Ontology.

To increase the recall of the mapping, we incorporated an effective text mining program described by Pafilis et al. (2013a). Compared to a simple string matching, the method helped to incorporate the synonyms of diseases and meanwhile being as specific as possible (For example catching diseases caused by different mutations on the same gene: *Charcot-Marie-Tooth disease type I–IV*).

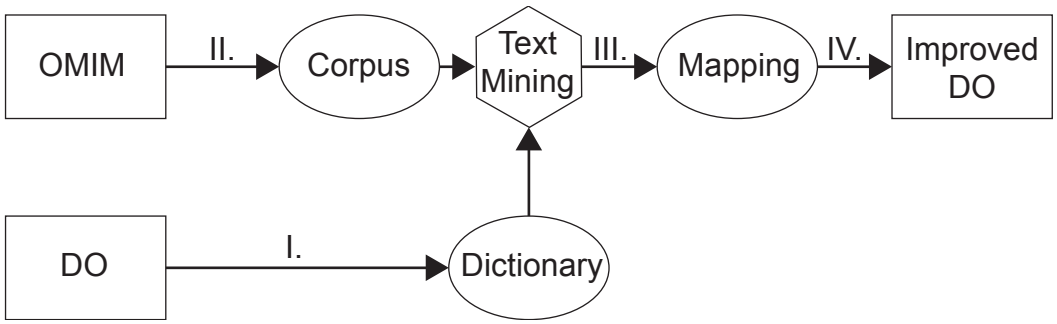



Figure 2.10: **The steps of creation of an improved Disease Ontology.**
Here I show the important steps of creating the mapping using a "pseudo" text mining pipeline. The numbers denote the following parts: I. deriving corpus from OMIM (see Methods), II. creation of a dictionary from Disease Ontology (see Methods) III. development of an efficient "scoring" to choose the best mapping and IV. manual curation by collaborators.

### Converting the Disease Ontology to a dictionary

We need to define the set of diseases that we will tag for mapping. This is called dictionary creation and the terms within are used to tag important expressions in the text. We have chosen to create a dictionary from the Disease Ontology, because the structure of the ontology ensures that we can create a broader mapping, when a strict mapping is not possible (e.g. various types of *Alzheimer disease 1-17* in OMIM should be mapped to the broader *Alzheimer disease* in Disease Ontology). We could further exploit the ontology by using the underlying tree structure (Schriml et al., 2011), where the broader terms are the parent terms and specific terms are the leaves. While creating the dictionary, we removed some Disease Ontology specific "features" from the term names (such as: *(morphologic abnormality)*) or introduced synonyms of *disorder* and *syndrome* for every *disease*. We created general to make sure that the dictionaries can be used in other text mining projects (e.g. DISEASES (http://diseases.jensenlab.org)) besides the mapping.

### Creating a corpus from OMIM

We defined the set of document, which we use to map every of them to a single disease. Therefore we prepared a corpus from OMIM and we ignored the parts, which were genes and obsolete entries. Moreover we have processed the titles of the disease pages in corpus to increase the chances of a correct mapping (see Methods).

To increase the possibility of a specific mapping, the corpus includes only the titles of diseases. We decided to ignore the descriptions, because it introduces a number of issues: I. one needs a scoring scheme, which can score enough high the correct disease from the description (however we implemented in a scoring scheme, which we dropped later, see Methods), II. how one chooses between the possible matches from the title and the description and III. it would make unnecessary difficult to handle complex diseases (such as *Cerebellar ataxia, mental retardation, and dysequilibrium syndrome*). We also ignored the complex diseases for the time being.

# 3 Methods

## 3.1 Assembling the COMPARTMENTS database

This section explains how the COMPARTMENTS database was built and the text was mainly taken from Binder et al. (2014).

### 3.1.1 Visualization of protein subcellular localization

For visualization purposes, we selected a set of commonly used localizations, including the cytosol and all major organelles. Each of these represents a GO term, and all evidence for more fine-grained localizations is projected onto these through *is_a* and *part_of* relationships. In case of multiple lines of evidence for the same localization, we always select the strongest. We subsequently present the evidence by color-coding a schematic of a cell. We have developed separate figures for animal, fungal, and plant cells to account for differences in their cell structure; for example, animal cells have no cell wall, and only plants have chloroplasts.

### 3.1.2 Assembly of the knowledge and experiments channels

We imported subcellular localization annotations from comments and database cross-reference fields of UniProtKB. We map these to the corresponding Ensembl identifiers using the STRING alias file (Franceschini et al., 2013) and GO terms using the UniProtKB controlled vocabulary of subcellular localizations. For *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster* and *Mus musculus* we imported

cellular component Gene Ontology annotations from their respective model organism database (Eppig et al., 2011; Cherry et al., 2011; McQuilton et al., 2011; Harris et al., 2009).

For the knowledge channel we assigned the highest score of 4 stars for annotations with the following evidence codes: CURATED, IDA, TAS, and NAS. We assigned 3 stars to the evidence codes PROBABLE, EXP, IPI, IMP, IGI, IEP, ISS, ISO, ISA, ISM, IBA, IBD, IKR, IMR, IRD, and IC. We assigned 2 stars to the less reliable evidence codes POTENTIAL, IGC, and IEA, while BY SIMILARITY, RCA and NR are assigned only 1 star. Because we consider some sources to be more reliable than others, we upgraded annotations from UniProtKB and the model organism databases by one star, resulting in a maximum score of 5 stars for the knowledge channel.

We also imported subcellular localization data from the Human Protein Atlas (HPA) (Uhlen et al., 2010; Fagerberg et al., 2013), which uses Ensembl identifiers, and manually mapped their locations to the corresponding GO terms. HPA uses two scoring schemes to classify the quality of its data. When a protein has been stained using two or more antibodies, HPA provides a reliability score based on the similarity of the staining patterns obtained with the different antibodies and the agreement with published literature. This scale has four levels of reliability: High (4 stars), Medium (3 stars), Low (2 stars), and Very low (1 star). When only a single antibody has been used for staining, we instead make use of the validation score provided by HPA. This scale has three levels: Supportive (3 stars), Uncertain (1 star), and Non-supportive (not imported).

### 3.1.3 Text mining of Medline abstracts

We used the protein dictionary from STRING 9.1 (Franceschini et al., 2013) and created a dictionary of names of subcellular localizations from the cellular component terms of the Gene Ontology (Ashburner et al., 2000). To improve the protein dictionary we discarded protein names that conflict with names of Gene Ontology terms. Furthermore, we blocked frequently occurring ambiguous names, such as acronyms, thereby greatly improving the precision. This was done through manual inspection of all protein and localization names giving rise to more than 2000 matches in Medline.

We matched these dictionaries against all Medline abstracts using an efficient named

entity recognition engine described elsewhere (Pafilis et al., 2013a). To score the co-occurring proteins and localizations, we used the text-mining scoring scheme of STRING 9.1 (Franceschini et al., 2013), which is a weighted count ($C(P,L)$) for each pair of protein $P$ and for localization $L$:

$$C(P,L) = \sum_{k=1}^{n} w_s \delta_{sk}(P,L) + w_a \delta_{ak}(P,L) \tag{3.1}$$

where $n$ is the number of abstracts, $w_s = 0.2$ and $w_a = 3$ are the weights for co-occurrence within the same sentence and within the same abstract, respectively. If $P$ and $L$ are mentioned together in a sentence or in abstract $k$, the delta functions $\delta_{ak}(P,L)$ and $\delta_{sk}(P,L)$ are 1, and 0 otherwise. Thus, an abstract that mentions $P$ and $L$ in the same sentence will give a score contribution of $w_s + w_a$ whereas an abstract that mentions them in different sentences will give a score contribution of $w_a$ only. The co-occurrence score ($S(P,L)$) is defined as:

$$S(P,L) = C(P,L)^{\alpha} \left( \frac{C(P,L)C(\bullet,\bullet)}{C(P,\bullet)C(\bullet,L)} \right)^{1-\alpha} \tag{3.2}$$

where $C(P,\bullet)$, $C(\bullet,L)$ and $C(\bullet,\bullet)$ are the sums over localizations paired with protein $P$, over all proteins from the same organism paired with localization $L$ and over all pairs of proteins from the same organism and localizations, respectively. The weighting factor $\alpha$ is 0.6. All parameters in the scoring scheme ($w_s$, $w_a$, and $\alpha$) were optimized to maximize the agreement between protein-protein co-occurrence scores and KEGG pathways (Franceschini et al., 2013).

The text-mining score depends on number of pairs identified in Medline abstracts, which changes as Medline grows. We therefore convert the scores into z-scores ($Z(P,L)$) to get a more robust measure. The observed distribution is a mixture of two, one from low-scoring random pairs and second from high-scoring, biologically meaningful pairs. The former is modeled as a Gaussian where the mean is equal to the mode of the observed distribution, which empirically coincides with the 40th percentile. The variance of the background is estimated from the difference between the 20th and the 40th percentiles. The final confidence score, stars, is the z-score/2, limited to a maximum of four.

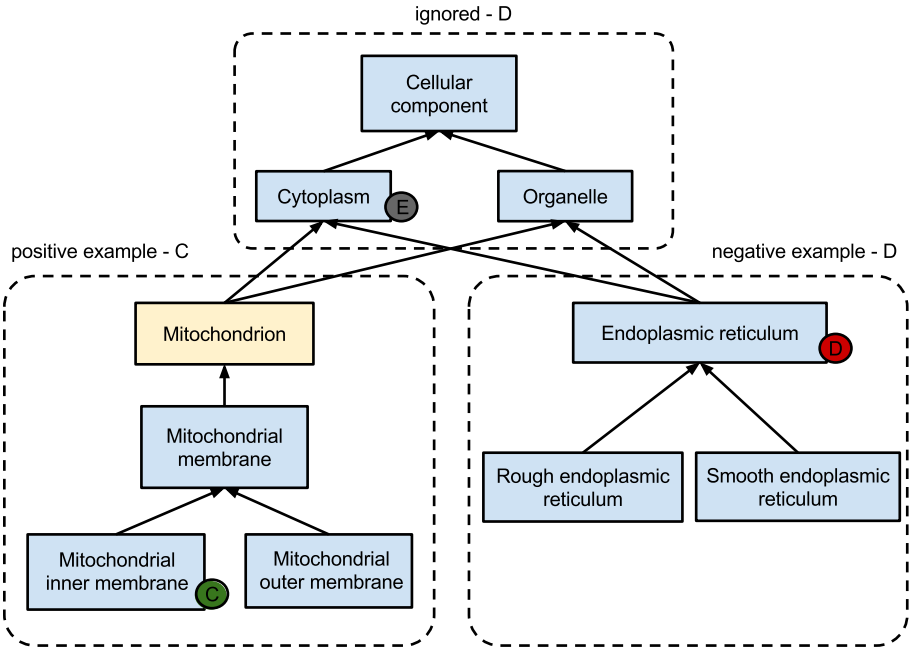



Figure 3.1: **Counting positive and negative examples.**
Through counting the mitochondrial proteins, we show how the benchmarking was done. Protein *C* was annotated being in the *Mitochondrial inner membrane* in the benchmark set, thus it could be backtracked to *Mitochondrion*, so we counted it as a positive example. Protein *D* was annotated being in the *Endoplasmic reticulum* so we added as a negative example. Protein *E* is known to be in the cytoplasm, but it also means that it might be in the *Mitochondrion* and in *Endoplasmic reticulum*, so we ignored it.

### 3.1.4 Construction of text-mining benchmark set

We constructed a high-quality benchmark set based on the knowledge channel. The positive examples are pairs of proteins and compartments supported by 5-star evidence. The negative examples are pairs of proteins and compartments for which there is no evidence suggesting that the protein is in the compartment and 5-star evidence for the protein being in a different compartment (Figure 3.1). The compartments considered for the benchmark set are the eleven subcellular localizations used in the overview figure, and all evidence for more specific localizations have been backtracked to this level. The benchmark set is available for download from the COMPARTMENTS web resource.

### 3.1.5 Construction of benchmark set for prediction based on protein-protein interaction

We derived the same benchmark set as for the text-mining. We used protein-protein interactions from STRING and for each protein we marked it in which compartment are they localized according to the benchmark set. The protein's (up to ten) highest-scoring interaction partners were also labeled with this compartment. Afterward we compared the labeled proteins to the benchmark set; if it is indeed localized in that compartment it was counted a true positive otherwise it was a false positive.

### 3.1.6 Scoring of sequence-based predictions

The WoLF PSORT and YLoc-HighRes methods were selected for prediction of subcellular localization. We pre-computed predictions for the entire set of protein sequences for human, mouse, rat, *Drosophila melanogaster*, *Caenorhabditis elegans*, *Saccharomyces cerevisiae* and *Arabidopsis thaliana* in STRING 9.1. We converted all scores to stars to make them comparable to other evidence types; the maximum number of stars that can be assigned to a sequence-based prediction is 3. This ensures that prediction scores cannot exceed the scores of reliable manual annotations, experiments or text mining.

PSORT (Horton et al., 2006) predicts localization based on various sequence-derived features such as sorting signals, binding domains and amino acid composition. These are used by a weighted $k$-nearest neighbor classifier. The output scores ($n$) roughly correspond to the number of the $k$ nearest neighbors from the training set that are annotated with each localization. We convert these scores to stars ($s_{PSORT}$) using the following formula:

$$s_{PSORT} = 3\frac{n}{k} \tag{3.3}$$

YLoc (Briesemeister et al., 2010b) is a naïve Bayes classifier that uses features similar to those of PSORT combined with GO annotations of close homologues. We found that most of the posterior probabilities from YLoc are very close to either 0 or 1. To differentiate between the probabilities close to 1 when converting them to stars, we

transform them using the following heuristic function:

$$s_{YLoc} = 3(1 - \sqrt[4]{1-p}) \tag{3.4}$$

where $s_{YLoc}$ is the stars derived from a YLoc prediction, $p$ is the prediction probability that the protein is localized in the given compartment. ($p < 0.2$ is ignored). This formula ensures that probabilities close to 1 become distinguishable when converted to stars: $p = 0.8 \rightarrow$ 1 star, $p = 0.99 \rightarrow$ 2 stars, $p = 0.999 \rightarrow$ 2.5 stars, $p = 1.0 \rightarrow$ 3 stars.

### 3.1.7 Statistical analysis of compartments sharing proteins

From the unified dataset we extracted localization information on human proteins with more than two stars to disregard weak text-mining and prediction evidence. The retrieved dataset comprised 18692 unique human proteins with 29493 links to compartments; 20021 were supported by curated knowledge, 4841 by high-throughput experimental evidence, 1468 by text mining, and 15788 by sequence-based predictions. We counted the number of proteins shared between any two compartments. To assess if this is higher than expected we compared the counts to a null model that assumed no correlation between any compartments. To this end we generated 1,000,000 random datasets in which links between proteins and compartments were permuted, thereby preserving the number of links per protein and per compartment. We computed a p-value for each pair of compartments as the fraction of random datasets resulting in a count greater than or equal to the observed count. Finally, we defined the statistically significant compartment pairs by imposing a false discovery rate of 0.1% using the Benjamini-Hochberg method (Benjamini and Hochberg, 1995).

## 3.2 Contribution to the TISSUES database

In this section I outline the technical aspects of my contributions to the TISSUES project. I will describe, how does the presentation of the figures work, how are the figures colored and how do we filter out the non-human tissues from the BRENDA Tissues Ontology (Gremse et al., 2011).

First I convert the scores to colors using the following equation, where $score_{max} = 5$ is the maximum confidence score, $score_{min} = 0$ is the minimum confidence score and

$score$ is the confidence score of the evidence at a given localization:

$$w = \frac{{score_{max}}^2 - score^2}{{score_{max}}^2 - {score_{min}}^2} \tag{3.5}$$

The $w$ weight also ensures that higher confidence scores receive a higher weight due to the quadratic function. Using the weight we calculated the colors with the following equation:

$$\mathbf{c} = \mathbf{c}_{\mathbf{score_{min}}}\, w + \mathbf{c}_{\mathbf{score_{max}}}\,(1 - w) \tag{3.6}$$

$\mathbf{c}_{\mathbf{score_{min}}}$ is the color (usually white) at zero confidence score and $\mathbf{c}_{\mathbf{score_{max}}}$ is the color (usually dark green) at the highest confidence score. Using the $w$ weight the equation creates the $\mathbf{c}$ color between $\mathbf{c}_{\mathbf{score_{min}}}$ and $\mathbf{c}_{\mathbf{score_{max}}}$.

After we calculated for every protein the color of the evidence for every tissue or subcellular localization based on the scores and stored the results in a PostgreSQL table.

The web framework called *blackmamba*, developed in Python, renders webpages using a PostgreSQL database and colors the figure templates. We designed our templates in SVG, because the images are vector-based and the format is a text-file, thus it is easy to automate the coloring the subcellular locations or tissues. My script finds the subcellular localizations or tissues in the figure and color them based on the evidence, and finally insert it as an element in the Document Object Model of the webpage.

```
<path title="plasma membrane" fill="#cccccc" />
<path title="cytosol" fill="#ffffff" />
<g title="cytoskeleton" fill="#cccccc" >
    ...
</g>
```

Figure 3.2: **Coloring the evidence in an SVG.**
One strength of SVGs that they are XML based and it is easy to manipulate them. The script look for specific titles in the file and changes the *fill* color based on the evidence.

As my another contribution to the project, I created a filtering of tissues. It works following way: Where a tissue does not exist an organisms (e.g. *swim bladder* in humans), we assigned those NCBI taxonomy ID (Federhen, 2012) to that tissue term.

Due the nature of the text mining some human proteins are associated with non-human tissues, but thanks to the filter these tissues are filtered out in a post-processing step, thus these false associations are not show on the webpage.

## 3.3 Developing tools to map OMIM to Disease Ontology

In this section I describe the technical details of the mapping. First I explain the rules for dictionary creation from the ontology, then I discuss, how did we convert OMIM to a corpus and finally I present the one-to-one mapping procedure from text mining.

### 3.3.1 Methods to build a dictionary from Disease Ontology

One term has a name and it might have synonyms as well in the ontology. We used a generalized script to convert the ontology into two files: one contains identifier name pairs, the other stores the parent-child relationships. To increase the coverage, we extended orthographically names and synonyms of the identifier name pairs. We used specific expansion rules for the Disease Ontology discussed in the table below: such as creating an orthographic variant by changing Roman numbers to Arabic numbers (such as for *congenital disorder of glycosylation type I and II*).

| Rule | Example |
|---|---|
| Removal of genitive case | *Alzheimer's disease* |
| Strip punctuations, quotes and parenthesis | *total, mature senile cataract* |
| Remove unnecessary word after deficiency | *coenzyme Q10 deficiency disease* |
| Roman to Arabic numbers | *congenital disorder of glycosylation type I and II* |
| Permuting disease, disorder and syndrome | *gallbladder disease* and *gallbladder disorder* |
| Permuting familial and hereditary | *familial retinoblastoma* |

Table 3.1: **The rules for orthographic expansion for the dictionary creation.**

After long testing we found these rules sufficient to cover most of the disease names found in OMIM and other corpuses.

### 3.3.2 Creating a corpus from OMIM database

The corpus creation consists of two parts. First we filtered out every OMIM page marked as a gene (symbol *) or moved into another entry (symbol ^) and we ignored the disease descriptions. Then continued with the titles left from OMIM disease pages. We kept the two categories of titles (main title, and alternative titles) to create a better mapping. We invented a new category for complex and less described diseases where the syndromes are listed (such as *knuckle pads, leukonychia, and sensorineural deafness*), becuase these diseases cannot be mapped to a single disease in OMIM. Acronyms were splitted from the titles and treated as a new title.

We prepared further the titles for the text mining by stripping commas, punctuations and quotes. Roman numbers and complex types (such as *type IIA*) were converted to arabic numbers without the trailing subtype (which are marked with a letter). OMIM writes the adjectives after the disease (such as *heart disease, congenital*) so we have moved them in the front. When more adjectives were used, we permuted them (e.g. *diabetes and deafness, maternally inherited* converted to *maternally inherited diabetes and deafness* and *inherited maternally diabetes and deafness* respectively). We had an old mapping scheme where we included the disease descriptions as well, however we do not utilize it for the being.

### 3.3.3 Using specific rules to derive a one-to-one mapping

After we ran the text-mining, we usually had more Disease Ontology candidate terms to map them to an OMIM page, we employed a "scoring scheme" to select the most proper mapping. We ignored the matches coming from the body of an OMIM page. We applied various rules until only one mapping was left (see Table 3.2 for details).

OMIM pages mapped to `"disease"` or to `"syndrome"` were excluded from the mapping, and it meant that those diseases are missing from the Disease Ontology. Therefore we used the excluded pairs to add more diseases to the Disease Ontology.

In this paragraph I explain, how did the old mapping worked and what were the issues. We counted how often does a term occur in a disease page and we included the disease descriptions. We corrected the counts per page by counting the parent terms as well and down-weighting terms that are mentioned in more OMIM pages by

| Level | Rule |
|---|---|
| 1st | Removing duplicated terms |
| 2nd | Discarding parent terms |
| 3rd | Keeping the matches from the first character |
| 4th | Discarding matches from alternative titles |
| 5th | Keeping match with the most words |

Table 3.2: **Applying filtering rules until one candidate is left.**
First terms tagged more than once are removed from matches. To increase the specificity of the mapping, we remove any terms, which children are also tagged. If there are still more matches left, we keep only those, that matches from the very first character in the title. This step ensures the selection of specific diseases. Some diseases in OMIM have multiple titles, and in such cases we keep the match from the main title. In very cases there are still multiple matches left and to get a one-to-one mapping we employ a heurestic rule, namely we map to the term with more words, because they are more specific.

using the term frequency–inverse document frequency. To emphasize the importance the matches from the main title and the alternative titles, we tripled and doubled the counts in these parts respectively. Finally for every OMIM term, we have chosen the highest-scoring term. Unfortunately there have been two major issues with this scoring: I. it tended to create more general mapping, even when a specific term existed in DO. II. it could not handle correctly the complex and less described diseases.

# 4 Discussion

Retrieving existing knowledge about proteins and other biological entities is a key step in biomedicine research. One needs to invest time in reading the scientific literature, examine previous experimental results and use computational tools, which can predict the answer to the biological question. While a lot of annotation resources exist the coverage as well as quality are lacking, thus there is a need for higher coverage of annotation of the biological data.

Fortunately there are many tools and databases available which address the above mentioned issues. Text mining collects meaningful biological associations from thousands of articles within the matter of minutes, while data integration provides a single view on different domains of knowledge. Ontologies have been created to annotate biological entities on different levels.

However some tools and databases are difficult to use or obsolete and no longer maintained. A large number of databases does not truly solve the data integration challenge, which include the scoring of biological data based on the different types and sources of information. The tools and databases presented in this thesis have tried to address these challenges by incorporating up-to-date data integration framework focusing on subcellular and tissue localization and diseases.

## 4.1 Providing an unified view of subcellular localization using COMPARTMENTS

The COMPARTMENTS resource unifies complementary evidence on protein localization from curated knowledge, high-throughput experiments, text mining and sequence-based prediction methods. We developed a self-updating pipeline, which regularly imports the data from model organism databases and from experimental datasets and processes the recent literature for associations without any human involvement.

We go beyond merely integrating many sources of evidence into a single database by mapping all pieces of evidence onto the same set of identifiers and carefully assigning them comparable confidence scores. We derived these through a combination of manual inspection of each evidence source, a previous study of the reliabilities of Gene Ontology evidence codes (Skunca et al., 2012), the benchmark results for the text-mining pipeline, and score distributions for the sequence-based prediction methods. This allows the users to compare and assess the different evidence based on the confidence score.

### 4.1.1 Accessing the web site of COMPARTMENTS

The primary aim of COMPARTMENTS web interface is to provide the user with a simple overview of the localization of a protein of interest without losing the connection to the underlying evidence. The resources disambiguates the queries by accepting the common names and various identifiers and provide a list of possible candidate organisms for the protein of interest. The overview is provided through a schematic of a cell, which is color-coded based on the strongest evidence supporting each compartment. This visualization is interactive, vector-based, supports a wide-range of devices and allows the user to see which evidence channels support a particular compartment and how strongly. Evidence exploration can be followed up with additional information about the origin of the evidence.

For the knowledge and experiments channels, the tables link out to the external databases from which the evidence was obtained. For text mining, the table gives access to a text viewer that shows the literature sources from abstracts in which the

protein and localization are co-mentioned, highlighting the terms that were recognized. The website itself incorporates the latest HTML5 technologies and supports a wide range of devices including smartphones and tablets.

COMPARTMENTS is the first resource to integrate subcellular localization evidence from manually curated annotations, high-throughput screens, and sequence-based predictions with automatic text mining for all major model organisms. To avoid the common problem of bioinformatics databases not being maintained, we have – from the beginning – designed the resource to be automatically kept up-to-date with the constant changes in source databases and literature. We address the challenge of making it easy for users to comprehend the heterogeneous evidence by projecting it onto a common reference both in terms of protein and compartment identifiers and in terms of reliability scores. This is complemented by the web interface, which provides an intuitive, interactive graphical overview of the unified evidence and simple tables with more detailed information, including links to the original sources. The website was built using an in-house developed general framework and it has been effectively used in similar resources. We also make the unified evidence available as bulk download files to facilitate large-scale computational studies of protein localization and integration with omics datasets.

### 4.1.2 Function of shared proteins shuffling between compartments

Demonstrating the usefulness of COMPARTMENTS for large-scale analyses, we derived a network of compartments, which is highly consistent with established knowledge on protein trafficking. We describe here 18 highly interconnected pairs of compartments sharing a high number of proteins as seen in Figure 2.7.

The pair of compartments, that shares the largest group of proteins, is the cytosol and the nucleus. The observation is not surprising, since nucleocytosolic protein transport is long known to orchestrate vital cell functions, such as the cell cycle and transcription regulation (Yoneda, 2000). However, the biological mechanisms that control the protein shuttling between nucleus and cytosol are currently in the focus of investigations (Sedek and Strous, 2013). Interestingly, both nucleus and cytosol exclusively share a significantly high number of proteins with the cytoskeleton. The majority of these proteins are involved in cell cycle regulation processes (e.g. centrosome organization, chromosome segregation and nuclear division) and the

cytoskeleton organization, representing successfully the highly dynamic collaboration between these compartments during cell cycle progression (Heng and Koh, 2010).

A further interesting relationship that emerged is the one that the peroxisomes have with the endoplasmic reticulum and the mitochondria, which is formed by proteins mainly involved in fatty acid metabolic and lipid biosynthetic processes. Noteworthy, while the functional cooperation between the peroxisomes and the endoplasmic reticulum in metabolic regulation is well studied (Dirkx et al., 2005; Braverman and Moser, 2012), their connection to mitochondria was not revealed until recently (Camões et al., 2009; Schrader and Yoon, 2007). However, further investigation of the proteins forming the functional bridges between peroxisomes and its interactors is expected to provide a valuable means to trace the underlying mechanisms of this yet poorly explored interplay (Islinger et al., 2012).

In contrast to the exclusive relationships described above, the rest of the intracellular compartments form a highly connected network. Proteins shared between the plasma membrane, endosomes and lysosomes, as well as the ones between lysosomes and the extracellular matrix are mainly involved in processes such as immune response and phagocytosis, reflecting their long known cooperation in the endocytic trafficking pathway (Watts, 2001, 2012). Respectively, proteins shared between the endoplasm reticulum, Golgi apparatus and plasma membrane mainly participate in protein processing such as protein metabolism, membrane proteolysis and phospholipid biosynthesis, supporting their well described collaboration on the exocytic pathway (Mellman and Warren, 2000). Last, further connections between the Golgi apparatus, endosomes and lysosomes represent the complex cross-talk between these two major trafficking pathways (Le Roy and Wrana, 2005).

Considering the fact that the protein environment in the cell is critical for the protein function, highly interconnected compartments can give valuable insight or even predict the functional role of the proteins identified in these compartments. Additionally, such information can support putative functional or physical interactions between proteins uncovering unreported protein collaborations. Because COMPARTMENTS uses the same protein identifiers as the STRING database (Franceschini et al., 2013), it also facilitates large-scale analysis of protein localization in the context of interaction networks.

### 4.1.3 Effectiveness of prediction of subcellular localization using protein-protein interactions

We have shown that if we understand the interaction network of proteins and some proteins are known to be localized in a specific compartment, then we can infer with high probability that the other proteins are also located in the same compartment. This is not surprising because in many cases physical proximity is necessary for protein-protein interaction (Fields and Song, 1989), thus it is likely that the proteins are localized in the same compartment. Indeed protein-protein interaction studies showed previously that interacting proteins are likely to be in the same compartment (Huh et al., 2003; von Mering et al., 2002).

With this study we showed that one can easily combine our database with other databases for large scale studies. One can also make predictions based on their localization as we did know predicting the localization of a protein using the interaction network.

### 4.1.4 Future directions

COMPARTMENTS is a user-friendly database, which provides an overview about subcellular localization of any given protein. Although I have developed the first version of this resource, it could be easily extended to support a broader range of studies or the resource could be tighter incorporated into bioinformatics pipeline. In these sections I give an overview about the possible future avenues.

#### Support for large-scale studies

Currently the community can carry out large studies by parsing the bulk download files. Although we showed that using only these files one can carry out powerful studies, using them also requires a solid knowledge of scripting and bioinformatics tools. Therefore the life science community could benefit if one could do simple large-scale studies using only the web site.

One obvious feature would be the possibility of querying for multiple proteins, but this task is not trivial and it needs significant amount of planning and development. Such features could especially aid high-throughput studies such as proteomics research.

For example it will be possible to submit a list of proteins to answer questions related to their localization. One question is whether the proteins are enriched in a specific compartment. The enrichment could be further customized by selecting the type and source of evidence (e.g. selecting the most reliable evidence from the knowledge channel). Another feature would visualize protein-protein interaction networks using STRING (Franceschini et al., 2013) and its payload mechanism to show where these proteins are localized. This can further help to see how proteins in different compartments interact.

### Increased evidence and species coverage

The recent version of COMPARTMENTS supports only seven organisms, but more species can be easily included (such as including proteins from *Danio rerio* or *Schizosaccharomyces pombe*). There are many challenges, whether a model organism database is available. The same name can refer to two completely different genes in within species (e.g. *Cdc2* in *Schizosaccharomyces pombe* and in *Saccharomyces cerevisiae*), and a better disambiguation module is needed for this task in the text mining pipeline.

The resource can also cover more evidence as it becomes available. We can add an experimental evidence channel for organisms other than human, if similar screening studies like Human Protein Atlas (Uhlen et al., 2010) become available. We can also incorporate more sources of evidence if the quality is high enough e.g. another source for human high-throughput experimental results. The set of predictions tools can be changed as well if better ones are developed.

### Developing a curation tool for database annotation

The evidence for protein-subcellular localization associations are supported by a number of Medline abstracts. We show these abstracts where the protein and the localization term are highlighted in the text. A curation tool based on this abstract viewer could be created, and it would show uncurated assocations to aid the annotiation process. It can also include full-text articles, and the highlighting feature improves the curators' ability to catch the evidence for the annotation.

**Accessing COMPARTMENTS through an API and integration into other services**

More tools can be integrated into our database, if we open the resource for external developers. This can be done by developing an Application Programming Interface (API) to retrieve the results programmatically. The schematic of the cell can be easily integrated into existing web-based frameworks. It can be easily manipulated to meet the needs of the tools, because it is a vector based image.

The API will be developed in conjunction with the new version of Reflect (O'Donoghue et al., 2010; Pafilis et al., 2009). One of the many improvements is refreshing the underlying database, from where Reflect retrieves the data. The new version will include the protein dictionary from COMPARTMENTS, and the overview of subcellular localization in the popup will be based on the COMPARTMENTS dataset and figure. Thus the Reflect will be capture more biomedical terms on the websites.

## 4.2 Using the general components of COMPARTMENTS and TISSUES

COMPARTMENTS (Binder et al., 2014), DISEASES, TISSUES and ORGANISMS (Pafilis et al., 2013a) share the most of the same software components and this framework reduced the time of development and the complexity of the resources. These components are reusable and they can be easily used to create new resources in order to effectively integrate and visualize biomedical data by today. Here I discuss how the scientific community will benefit from my contributions to this framework.

### 4.2.1 Creating a human-specific text mining pipeline

The tissue protein association mined from Medline abstracts were not species specific. The terms in the BRENDA tissue ontology (Gremse et al., 2011) are not organism specific, and all terms were included in the dictionary creation. Many protein names are the same in human and other species. The text-mining pipeline does not always recognize, what species the text is about and then by default it tags human proteins. Unfortunately this gave rise to incorrect associations with non-human tissues.

Therefore I solved the problem of false association by creating a filter for the tissue

dictionary. This filter efficiently removes non-human tissues from the final results, therefore it increases the quality of the text-mining results. These results can also be a basis later for annotation of proteins.

### 4.2.2 Providing a general overview of evidence channels

It is important to show where the original biological data was derived. Therefore I created an efficient overview figure for COMPARTMENTS and TISSUES, which shows where the evidence comes from, how strong it is, and to what bigger category does it belongs to. To support a wide-range of devices, the figure is vector-based and it is text-based to populate easily with information in bioinformatics processes.

### 4.2.3 Future directions

COMPARTMENTS (Binder et al., 2014), DISEASES, TISSUES and ORGANISMS (Pafilis et al., 2013a) can be easily extended simultaneously by developing new components. Also the resources can be further integrated to unravel more biological associations. Here I outline a few possible avenues.

#### Opening a one-stop shop for protein data

We have a resource for subcellular localizations, tissue and disease associations. Integrating these data we can create a knowledge-base about proteins. Upon that we could create a mobile or tablet application, which is useful if one needs to gather information quickly about a protein at a conference or at a workshop. Some parts are already shared, for example text-mining engine collects associations in one run, but the results are presented in the above mentioned databases.

It is still a challenge, how to present the results to the user or what would be a clear visualization for the evidence. Also we could create an integrated platform for high-throughput experiments, where one can not only see if a protein is enriched in a subcellular compartment, but whether these proteins are also connected somehow to a disease.

#### Mining tissue associations for more species

We made the filter to support more organisms and text-mining pipeline can easily include associations from other species. However we do not have high-quality annotated and experimental data for organisms other than human yet. Should that data become available, we could integrate protein-tissue associations for other model organisms.

## 4.3 Linking more disease related data to Disease Ontology

Ontologies were used for data integration in COMPARTMENTS and in TISSUES, because they allow different levels of annotation. Moreover the dictionaries derived from the ontologies are very effective for text mining of biological associations. Unfortunately some of the ontologies contain only a limited number of terms and not all of them are cross-linked to other popular biological databases, therefore their usability is limited. However there are clear advantages of using ontologies in biological data integration and improving them could accelerate the integration progresses.

To unleash the full potential of our text-mining engine, we created an efficient mapping pipeline, which links OMIM (Amberger et al., 2011) documents to Disease Ontology terms (Schriml et al., 2011). Since OMIM is widely used for curation processes, this integration effort allows easier and automatic annotation of proteins using Disease Ontology terms. Moreover as a "by-product" of the mapping process, we found several hereditary diseases, which do not exist in the Disease Ontology. Hopefully these will be added to the next release.

The current version of Disease Ontology performs well in text mining of Medline abstracts (see DISEASES), and through the mapping we could increase evidence from knowledge. By adding more term to the ontology, we can mine more disease-protein associations from the abstracts. A clear advantage of the mapping is that the Disease Ontology allows both narrow and broad annotations thanks to the tree-like structure.

In this section we elaborate what are the advantages of this automated mapping pipeline and how does it help manual integration efforts. Finally we discuss how can it be extended to satisfy further goals of the bioinformatics community.

### 4.3.1 Mapping datasets automatically

Manual integration of ontologies and biological databases is a time-consuming and difficult task because they evolve over time. By developing an automated solution, we addressed the mapping entries of new versions of these databases. We also created specific rules adopted for this task, which ensures high precision of the mappings.

Many maladies contain words of "disease" or "syndrome", which are root terms in the Disease Ontology. When the mapping could only link the documents to these entries meant that these maladies do not exist in Disease Ontology thus we could extend it with these terms.

### 4.3.2 Building up international collaborations

Disease Ontology is a community effort and we wanted to give back to the community. We have sent the results of the automated mappings to the team behind the ontology and it served as a basis for their annotation efforts. The manual inspection increased also the quality of the mappings, because the automated pipeline cannot outperform the wisdom of the curators. We hope that the pipeline will be part of the annotation process of the Disease Ontology in the future.

### 4.3.3 Future directions

For the time being the mapping is focused on two specific databases, however we envision that it can be coverted to a general mappinf and extended to support more datasets. Here I outline how over text-mining based integration can be used in other projects.

#### Automatic Disease Ontology annotation of UniProt or Ensembl entries

UniProt and Ensembl annotate proteins and genes using OMIM identifiers, however a significant problem of OMIM lies in the restrictive licensing terms. Therefore disease descriptions from OMIM cannot be taken without signing an agreement and it can limit the commercial usability of derivative products. Fortunately UniProt contains beside the annotations a short description of the diseases with a more relaxed license,

which can be reused in Disease Ontology descriptions.

Also the descriptions can be used for the text-mining based mapping pipeline, and we can automatically annotate the proteins in UniProt with Disease Ontology terms. Another approach is to automatically map the OMIM annotations in UniProt and Ensembl by using the mapping file generated by the pipeline.

### Integrating more disease-related databases

Disease Ontology links to many datasets about human diseases, it is however manually curated. This could be largely automatized, if the pipeline supported other sources. Since many of the rules are general, we think that it could be possible to integrate some of these sources (e.g. ICD-9, ICD-10, MeSH, NCI's thesaurus and SNOMED CT entries).

# 5 Conclusions

In this thesis I presented a work, that is focused on integrating and presenting biomedical information to the life science community. These resources are targeting at a broad audience, from wet lab biologist to bioinformatics experts.

Presenting the data in a user friendly way was one of the driving principles. This goal was achieved in easy-to-use querying interface and a summarized results page supporting multiple devices, which make the resources attractive for simple biologist users. The reuse of data is encouraged by providing the underlying dataset in standardized formats.

To widen the scope of the resources and extend their capabilities, developing reusable components was also one of the main goals. For text-mining and data integration we have reused and fine-tuned existing components from other highly-used scientific methods.

Keeping the resources up-to-date is tedious work, and resources often become abandoned and outdated, when the main authors leaves. Therefore we developed a robust self-updating pipeline, which automatically pulls in the recent updates. In comparison with many other resource, this database keeps itself up-to-date thanks to the automated self-updating pipeline.

COMPARTMENTS – to our knowledge – is the first resource, which integrates subcellular localization evidence for all model organisms, derived from curated knowledge, high-throughput experimental data, associations based on automated text-mining and results of sequence-based predictions. The resource is equipped with a user-

friendly interface and graphical summaries, which makes it attractive to a broader audience including students, teachers and researchers with an interest in molecular biology.

The resource solved the challenging task of identifier disambiguation for data integration from various sources, while hiding the complex underlying machinery from the user. The in-house text-mining pipeline complements the manually curated knowledge by capturing valuable protein-subcellular localization associations from millions of scientific articles.

To make sure that the scientific community can assess the underlying localization evidence, COMPARTMENTS uses a common confidence scoring system. We have carefully scored the different types and sources of evidence and for every protein there is an overview figure created to get an immediate understanding about the localization information. Moreover the resource also presents specific detailed textual evidence to provide the necessary details for scientific research. Finally every piece of information are supported by linkouts to its origins in order to easily track the source of the underlying evidence.

The database is also available for download to allow large-scale analyses of the proteome. These analyses help to understand how protein-protein interaction networks, protein function are connected with subcellular localization.

Many pieces of the pipeline behind COMPARTMENTS have been reused to build a database, called TISSUES, about protein expressions in human tissues. This resource employs a powerful graphical overview to give a quick glance of tissue expression based on different types and sources of evidence. The text-mining pipeline has been tailored to mine human-specific tissue-protein associations from the scientific literature. In contrary to the previous resource, this one integrates multiple high-throughput experimental datasets. In summary TISSUES, provides a unified view on tissue expression based on curated data, experimental results and text-mining.

Ontologies were used to integrate evidence in the previously mentioned resources, because they allow different levels of annotation. However they have varying qualities, which limits the possibility of building a resource based on them. Therefore a modified-version of the text-mining pipeline (OMIM2DO) mapped terms of OMIM to Disease Ontology terms. A clear advantage is that the mapping is extended when

a new release comes out, and it also suggests new terms to the ontology. Finally we have built up a new international collaboration, and the mapping serves as a basis of manual curation.

Both COMPARTMENTS and TISSUES tackled the challenge of integrating and assigning a common scoring system for evidence from various sources. These resources provide a one-stop-shop for subcellular and tissue localization information and support large analysis of the whole proteome. OMIM2DO solved the task on another level by improving the fundamentals of data integration.

This software paved the way for enabling the creation of a single knowledge-base of proteins by collecting evidence such as their subcellular localization, tissue expression, involvement in diseases from multiple sources. The underlying evidence and visualization can be incorporated into third party tools for example for high-throughput analysis.

# Bibliography


Agemy, L., Kotamraju, V. R., Friedmann-Morvinski, D., Sharma, S., Sugahara, K. N., and Ruoslahti, E. (2013). Proapoptotic peptide-mediated cancer therapy targeted to cell surface p32. *Molecular therapy: the journal of the American Society of Gene Therapy*, **21**, 2195–2204. doi:10.1038/mt.2013.191.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990). Basic local alignment search tool. *Journal of molecular biology*, **215**, 403–410. doi:10.1016/S0022-2836(05)80360-2.

Amberger, J., Bocchini, C., and Hamosh, A. (2011). A new face and new challenges for online mendelian inheritance in man (OMIM®). *Human Mutation*, **32**, 564–567. doi:10.1002/humu.21466.

Amberger, J., Bocchini, C. A., Scott, A. F., and Hamosh, A. (2009). McKusick's online mendelian inheritance in man (OMIM®). *Nucleic Acids Research*, **37**, D793–D796. doi:10.1093/nar/gkn665.

Ananiadou, S., Kell, D. B., and Tsujii, J.-i. (2006). Text mining and its potential applications in systems biology. *Trends in Biotechnology*, **24**, 571–579. doi:10.1016/j.tibtech.2006.10.002.

Andersen, J. S., Lyon, C. E., Fox, A. H., Leung, A. K., Lam, Y. W., Steen, H., Mann, M., and Lamond, A. I. (2002). Directed proteomic analysis of the human nucleolus. *Current Biology*, **12**, 1–11. doi:10.1016/S0960-9822(01)00650-9.

Andrade, M. A., O'Donoghue, S. I., and Rost, B. (1998). Adaptation of protein surfaces to subcellular location. *Journal of molecular biology*, **276**, 517–525. doi:10.1006/jmbi.1997.1498.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M., and Sherlock, G. (2000). Gene ontology: tool for the unification of biology. *Nature Genetics*, **25**, 25–29. doi:10.1038/75556.

Avraham, S., Tung, C.-W., Ilic, K., Jaiswal, P., Kellogg, E. A., McCouch, S., Pujar, A., Reiser, L., Rhee, S. Y., Sachs, M. M., Schaeffer, M., Stein, L., Stevens, P., Vincent, L., Zapata, F., and Ware, D. (2008). The plant ontology database: a community resource for plant structure and developmental stages controlled vocabulary and annotations. *Nucleic Acids Research*, **36**, D449–D454. doi:10.1093/nar/gkm908.

Bairoch, A., and Apweiler, R. (1996). The SWISS-PROT protein sequence data bank and its new supplement TREMBL. *Nucleic acids research*, **24**, 21–25.

Bairoch, A., and Boeckmann, B. (1991). The SWISS-PROT protein sequence data bank. *Nucleic acids research*, **19 Suppl**, 2247–2249.

Bannai, H., Tamada, Y., Maruyama, O., Nakai, K., and Miyano, S. (2002). Extensive feature detection of n-terminal protein sorting signals. *Bioinformatics*, **18**, 298–305. doi:10.1093/bioinformatics/18.2.298.

Belinky, F., Bahir, I., Stelzer, G., Zimmerman, S., Rosen, N., Nativ, N., Dalah, I., Iny Stein, T., Rappaport, N., Mituyama, T., Safran, M., and Lancet, D. (2013). Non-redundant compendium of human ncRNA genes in GeneCards. *Bioinformatics (Oxford, England)*, **29**, 255–261. doi:10.1093/bioinformatics/bts676.

Bell, A. W., Ward, M. A., Blackstock, W. P., Freeman, H. N. M., Choudhary, J. S., Lewis, A. P., Chotai, D., Fazel, A., Gushue, J. N., Paiement, J., Palcy, S., Chevet, E., Lafrenière-Roula, M., Solari, R., Thomas, D. Y., Rowley, A., and Bergeron, J. J. M. (2001). Proteomics characterization of abundant golgi membrane proteins. *Journal of Biological Chemistry*, **276**, 5152–5165. doi:10.1074/jbc.M006143200.

Benjamini, Y., and Hochberg, Y. (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society. Series B. Methodological*, **57**, 289–300.

Binder, J. X., Pletscher-Frankild, S., Tsafou, K., Stolte, C., O'Donoghue, S. I., Schneider, R., and Jensen, L. J. (2014). COMPARTMENTS: unification and visualization of pro-

tein subcellular localization evidence. *Database: the journal of biological databases and curation*, **2014**, bau012. doi:10.1093/database/bau012.

Binns, D., Dimmer, E., Huntley, R., Barrell, D., O'Donovan, C., and Apweiler, R. (2009). QuickGO: a web-based tool for gene ontology searching. *Bioinformatics (Oxford, England)*, **25**, 3045–3046. doi:10.1093/bioinformatics/btp536.

Blair, D. R., Lyttle, C. S., Mortensen, J. M., Bearden, C. F., Jensen, A. B., Khiabanian, H., Melamed, R., Rabadan, R., Bernstam, E. V., Brunak, S., Jensen, L. J., Nicolae, D., Shah, N. H., Grossman, R. L., Cox, N. J., White, K. P., and Rzhetsky, A. (2013). A nondegenerate code of deleterious variants in mendelian loci contributes to complex disease risk. *Cell*, **155**, 70–80. doi:10.1016/j.cell.2013.08.030.

Blake, J. A., Richardson, J. E., Davisson, M. T., and Eppig, J. T. (1997). The mouse genome database (MGD). a comprehensive public resource of genetic, phenotypic and genomic data. *Nucleic Acids Research*, **25**, 85–91. doi:10.1093/nar/25.1.85.

Bourne, P. (2005). Will a biological database be different from a biological journal? *PLoS Comput Biol*, **1**, e34. doi:10.1371/journal.pcbi.0010034.

Braverman, N. E., and Moser, A. B. (2012). Functions of plasmalogen lipids in health and disease. *Biochimica et Biophysica Acta (BBA) - Molecular Basis of Disease*, **1822**, 1442–1452. doi:10.1016/j.bbadis.2012.05.008.

Briesemeister, S., Rahnenfuhrer, J., and Kohlbacher, O. (2010a). YLoc–an interpretable web server for predicting subcellular localization. *Nucleic Acids Research*, **38**, W497–W502. doi:10.1093/nar/gkq477.

Briesemeister, S., Rahnenführer, J., and Kohlbacher, O. (2010b). Going from where to why—interpretable prediction of protein subcellular localization. *Bioinformatics*, **26**, 1232–1238. doi:10.1093/bioinformatics/btq115.

Buttigieg, P. L., Morrison, N., Smith, B., Mungall, C. J., Lewis, S. E., and $author.lastName, a. f. (2013). The environment ontology: contextualising biological and biomedical entities. *Journal of Biomedical Semantics*, **4**, 43. doi:10.1186/2041-1480-4-43.

Buza, T. J., McCarthy, F. M., Wang, N., Bridges, S. M., and Burgess, S. C. (2008). Gene ontology annotation quality analysis in model eukaryotes. *Nucleic acids research*, **36**, e12. doi:10.1093/nar/gkm1167.

Camões, F., Bonekamp, N. A., Delille, H. K., and Schrader, M. (2009). Organelle dynamics and dysfunction: A closer link between peroxisomes and mitochondria. *Journal of inherited metabolic disease*, **32**, 163–180. doi:10.1007/s10545-008-1018-3.

Carbon, S., Ireland, A., Mungall, C. J., Shu, S., Marshall, B., and Lewis, S. (2009). AmiGO: online access to ontology and annotation data. *Bioinformatics*, **25**, 288–289. doi:10.1093/bioinformatics/btn615.

Chang, J. T., Schütze, H., and Altman, R. B. (2004). GAPSCORE: finding gene and protein names one word at a time. *Bioinformatics (Oxford, England)*, **20**, 216–225.

Cheng, D., Knox, C., Young, N., Stothard, P., Damaraju, S., and Wishart, D. S. (2008). PolySearch: a web-based text mining system for extracting relationships between human diseases, genes, mutations, drugs and metabolites. *Nucleic Acids Research*, **36**, W399–W405. doi:10.1093/nar/gkn296.

Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A., Dwight, S. S., Hester, E. T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S., and Botstein, D. (1998). SGD: saccharomyces genome database. *Nucleic Acids Research*, **26**, 73–79. doi:10.1093/nar/26.1.73.

Cherry, J. M., Hong, E. L., Amundsen, C., Balakrishnan, R., Binkley, G., Chan, E. T., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S. R., Fisk, D. G., Hirschman, J. E., Hitz, B. C., Karra, K., Krieger, C. J., Miyasato, S. R., Nash, R. S., Park, J., Skrzypek, M. S., Simison, M., Weng, S., and Wong, E. D. (2011). Saccharomyces genome database: the genomics resource of budding yeast. *Nucleic Acids Research*, **40**, D700–D705. doi:10.1093/nar/gkr1029.

Cohen, K. B., Johnson, H. L., Verspoor, K., Roeder, C., and Hunter, L. E. (2010). The structural and content aspects of abstracts versus bodies of full text journal articles are different. *BMC Bioinformatics*, **11**, 492. doi:10.1186/1471-2105-11-492.

Consortium, G. O. (2004). The gene ontology (GO) database and informatics resource. *Nucleic Acids Research*, **32**, D258–D261. doi:10.1093/nar/gkh036.

Consortium, T. G. O. (2010). The gene ontology in 2010: extensions and refinements. *Nucleic Acids Research*, **38**, D331–D335. doi:10.1093/nar/gkp1018.

Costanzo, M. C., Engel, S. R., Wong, E. D., Lloyd, P., Karra, K., Chan, E. T., Weng, S., Paskov, K. M., Roe, G. R., Binkley, G., Hitz, B. C., and Cherry, J. M. (2014). Saccha-

romyces genome database provides new regulation data. *Nucleic Acids Research*, **42**, D717–D725. doi:10.1093/nar/gkt1158.

Degtyarenko, K., Matos, P. d., Ennis, M., Hastings, J., Zbinden, M., McNaught, A., Alcántara, R., Darsow, M., Guedj, M., and Ashburner, M. (2008). ChEBI: a database and ontology for chemical entities of biological interest. *Nucleic Acids Research*, **36**, D344–D350. doi:10.1093/nar/gkm791.

Ding, J., Berleant, D., Nettleton, D., and Wurtele, E. (2002). Mining MEDLINE: abstracts, sentences, or phrases. In *Proceedings of the pacific symposium on biocomputing*, vol. 7, (pp. 326—337).

Dirkx, R., Vanhorebeek, I., Martens, K., Schad, A., Grabenbauer, M., Fahimi, D., Declercq, P., Van Veldhoven, P. P., and Baes, M. (2005). Absence of peroxisomes in mouse hepatocytes causes mitochondrial and ER abnormalities. *Hepatology*, **41**, 868–878. doi:10.1002/hep.20628.

Doms, A., and Schroeder, M. (2005). GoPubMed: exploring PubMed with the gene ontology. *Nucleic Acids Research*, **33**, W783–W786. doi:10.1093/nar/gki470.

Drawid, A., and Gerstein, M. (2000). A bayesian system integrating expression data with sequence patterns for localizing proteins: comprehensive application to the yeast genome. *Journal of molecular biology*, **301**, 1059–1075. doi:10.1006/jmbi.2000.3968.

Du, P., Feng, G., Flatow, J., Song, J., Holko, M., Kibbe, W. A., and Lin, S. M. (2009). From disease ontology to disease-ontology lite: statistical methods to adapt a general-purpose ontology for the test of gene-ontology associations. *Bioinformatics*, **25**, i63–i68. doi:10.1093/bioinformatics/btp193.

du Plessis, L., Skunca, N., and Dessimoz, C. (2011). The what, where, how and why of gene ontology–a primer for bioinformaticians. *Briefings in bioinformatics*, **12**, 723–735. doi:10.1093/bib/bbr002.

Engelhardt, B. E., Jordan, M. I., Muratore, K. E., and Brenner, S. E. (2005). Protein molecular function prediction by bayesian phylogenomics. *PLoS computational biology*, **1**, e45. doi:10.1371/journal.pcbi.0010045.

Eppig, J. T., Blake, J. A., Bult, C. J., Kadin, J. A., Richardson, J. E., and the Mouse Genome Database Group (2011). The mouse genome database (MGD): comprehensive resource for genetics and genomics of the laboratory mouse. *Nucleic Acids Research*, **40**, D881–D886. doi:10.1093/nar/gkr974.

Eriksson, R., Werge, T., Jensen, L. J., and Brunak, S. (2014). Dose-specific adverse drug reaction identification in electronic patient records: Temporal data mining in an inpatient psychiatric population. *Drug Safety*, **37**, 237–247. doi:10.1007/s40264-014-0145-z.

Fagerberg, L., Oksvold, P., Skogs, M., Älgenäs, C., Lundberg, E., Pontén, F., Sivertsson, A., Odeberg, J., Klevebring, D., Kampf, C., Asplund, A., Sjöstedt, E., Al-Khalili Szigyarto, C., Edqvist, P.-H., Olsson, I., Rydberg, U., Hudson, P., Ottosson Takanen, J., Berling, H., Björling, L., Tegel, H., Rockberg, J., Nilsson, P., Navani, S., Jirström, K., Mulder, J., Schwenk, J. M., Zwahlen, M., Hober, S., Forsberg, M., von Feilitzen, K., and Uhlén, M. (2013). Contribution of antibody-based protein profiling to the human chromosome-centric proteome project (c-HPP). *Journal of Proteome Research*, **12**, 2439–2448. doi:10.1021/pr300924j.

Federhen, S. (2012). The NCBI taxonomy database. *Nucleic acids research*, **40**, D136–143. doi:10.1093/nar/gkr1178.

Fields, S., and Song, O.-k. (1989). A novel genetic system to detect protein-protein interactions. *Nature*, **340**, 245–246. doi:10.1038/340245a0.

Flicek, P., Ahmed, I., Amode, M. R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fairley, S., Fitzgerald, S., Gil, L., Garcia-Giron, C., Gordon, L., Hourlier, T., Hunt, S., Juettemann, T., Kahari, A. K., Keenan, S., Komorowska, M., Kulesha, E., Longden, I., Maurel, T., McLaren, W. M., Muffato, M., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., Riat, H. S., Ritchie, G. R. S., Ruffier, M., Schuster, M., Sheppard, D., Sobral, D., Taylor, K., Thormann, A., Trevanion, S., White, S., Wilder, S. P., Aken, B. L., Birney, E., Cunningham, F., Dunham, I., Harrow, J., Herrero, J., Hubbard, T. J. P., Johnson, N., Kinsella, R., Parker, A., Spudich, G., Yates, A., Zadissa, A., and Searle, S. M. J. (2012). Ensembl 2013. *Nucleic Acids Research*, **41**, D48–D55. doi:10.1093/nar/gks1236.

Flicek, P., Amode, M. R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S., Gil, L., Girón, C. G., Gordon, L., Hourlier, T.,

Hunt, S., Johnson, N., Juettemann, T., Kähäri, A. K., Keenan, S., Kulesha, E., Martin, F. J., Maurel, T., McLaren, W. M., Murphy, D. N., Nag, R., Overduin, B., Pignatelli, M., Pritchard, B., Pritchard, E., Riat, H. S., Ruffier, M., Sheppard, D., Taylor, K., Thormann, A., Trevanion, S. J., Vullo, A., Wilder, S. P., Wilson, M., Zadissa, A., Aken, B. L., Birney, E., Cunningham, F., Harrow, J., Herrero, J., Hubbard, T. J. P., Kinsella, R., Muffato, M., Parker, A., Spudich, G., Yates, A., Zerbino, D. R., and Searle, S. M. J. (2013). Ensembl 2014. *Nucleic acids research*. doi:10.1093/nar/gkt1196.

Franceschini, A., Szklarczyk, D., Frankild, S., Kuhn, M., Simonovic, M., Roth, A., Lin, J., Minguez, P., Bork, P., von Mering, C., and Jensen, L. J. (2013). STRING v9.1: protein-protein interaction networks, with increased coverage and integration. *Nucleic acids research*, **41**, D808–D815. doi:10.1093/nar/gks1094.

Friedman, C., Kra, P., Yu, H., Krauthammer, M., and Rzhetsky, A. (2001). GENIES: a natural-language processing system for the extraction of molecular pathways from journal articles. *Bioinformatics (Oxford, England)*, **17 Suppl 1**, S74–82.

Fukuda, K., Tamura, A., Tsunoda, T., and Takagi, T. (1998). Toward information extraction: identifying protein names from biological papers. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, (pp. 707–718).

Gaudan, S., Kirsch, H., and Rebholz-Schuhmann, D. (2005). Resolving abbreviations to their senses in medline. *Bioinformatics*, **21**, 3658–3664. doi:10.1093/bioinformatics/bti586.

Gaudet, P., Argoud-Puy, G., Cusin, I., Duek, P., Evalet, O., Gateau, A., Gleizes, A., Pereira, M., Zahn-Zabal, M., Zwahlen, C., Bairoch, A., and Lane, L. (2013). neXtProt: organizing protein knowledge in the context of human proteome projects. *Journal of Proteome Research*, **12**, 293–298. doi:10.1021/pr300830v.

Gerner, M., Nenadic, G., and Bergman, C. M. (2010). LINNAEUS: a species name identification system for biomedical literature. *BMC Bioinformatics*, **11**, 85. doi:10.1186/1471-2105-11-85.

Glenisson, P., Antal, P., Mathys, J., Moreau, Y., and De Moor, B. (2003). Evaluation of the vector space representation in text-based gene clustering. *Pacific Symposium on Biocomputing. Pacific Symposium on Biocomputing*, (pp. 391–402).

Goldberg, T., Hamp, T., and Rost, B. (2012). LocTree2 predicts localization for all domains of life. *Bioinformatics*, **28**, i458–i465. doi:10.1093/bioinformatics/bts390.

Gremse, M., Chang, A., Schomburg, I., Grote, A., Scheer, M., Ebeling, C., and Schomburg, D. (2011). The BRENDA tissue ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Research*, **39**, D507–D513. doi:10.1093/nar/gkq968.

Guo, T., Hua, S., Ji, X., and Sun, Z. (2004). DBSubLoc: database of protein subcellular localization. *Nucleic Acids Research*, **32**, D122–D124. doi:10.1093/nar/gkh109.

Gutekunst, C.-A., Li, S.-H., Yi, H., Ferrante, R. J., Li, X.-J., and Hersch, S. M. (1998). The cellular and subcellular localization of huntingtin-associated protein 1 (HAP1): comparison with huntingtin in rat and human. *The Journal of Neuroscience*, **18**, 7674–7686.

Hanahan, D., and Weinberg, R. A. (2011). Hallmarks of cancer: the next generation. *Cell*, **144**, 646–674. doi:10.1016/j.cell.2011.02.013.

Hanisch, D., Fundel, K., Mevissen, H.-T., Zimmer, R., and Fluck, J. (2005). ProMiner: rule-based protein and gene entity recognition. Report Suppl 1, BioMed Central Ltd.

Harris, M. A., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G. M., Blake, J. A., Bult, C., Dolan, M., Drabkin, H., Eppig, J. T., Hill, D. P., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J. M., Christie, K. R., Costanzo, M. C., Dwight, S. S., Engel, S., Fisk, D. G., Hirschman, J. E., Hong, E. L., Nash, R. S., Sethuraman, A., Theesfeld, C. L., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S. Y., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E. M., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T., White, R., and Gene Ontology Consortium (2004). The gene ontology (GO) database and informatics resource. *Nucleic acids research*, **32**, D258–261. doi:10.1093/nar/gkh036.

Harris, T. W., Antoshechkin, I., Bieri, T., Blasiar, D., Chan, J., Chen, W. J., De La Cruz, N., Davis, P., Duesbury, M., Fang, R., Fernandes, J., Han, M., Kishore, R., Lee, R., Muller, H.-M., Nakamura, C., Ozersky, P., Petcherski, A., Rangarajan, A., Rogers,

A., Schindelman, G., Schwarz, E. M., Tuli, M. A., Van Auken, K., Wang, D., Wang, X., Williams, G., Yook, K., Durbin, R., Stein, L. D., Spieth, J., and Sternberg, P. W. (2009). WormBase: a comprehensive resource for nematode research. *Nucleic Acids Research*, **38**, D463–D467. doi:10.1093/nar/gkp952.

Hastings, J., de Matos, P., Dekker, A., Ennis, M., Harsha, B., Kale, N., Muthukrishnan, V., Owen, G., Turner, S., Williams, M., and Steinbeck, C. (2013). The ChEBI reference database and ontology for biologically relevant chemistry: enhancements for 2013. *Nucleic Acids Research*, **41**, D456–D463. doi:10.1093/nar/gks1146.

Heng, Y.-W., and Koh, C.-G. (2010). Actin cytoskeleton dynamics and the cell division cycle. *The International Journal of Biochemistry & Cell Biology*, **42**, 1622–1633. doi:10.1016/j.biocel.2010.04.007.

Hill, D. P., Smith, B., McAndrews-Hill, M. S., and Blake, J. A. (2008). Gene ontology annotations: what they mean and where they come from. *BMC bioinformatics*, **9 Suppl 5**, S2. doi:10.1186/1471-2105-9-S5-S2.

Horton, P., and Nakai, K. (1997). Better prediction of protein cellular localization sites with the k nearest neighbors classifier. *Proceedings / ... International Conference on Intelligent Systems for Molecular Biology ; ISMB. International Conference on Intelligent Systems for Molecular Biology*, **5**, 147–152.

Horton, P., Park, K.-J., Obayashi, T., Fujita, N., Harada, H., Adams-Collier, C., and Nakai, K. (2007). WoLF PSORT: protein localization predictor. *Nucleic Acids Research*, **35**, W585–W587. doi:10.1093/nar/gkm259.

Horton, P., Park, K.-J., Obayashi, T., and Nakai, K. (2006). Protein subcellular localization prediction with WoLF PSORT. *Proceedings of the 4th Annual Asia Pacific Bioinformatics Conference APBC06*, (pp. pp. 39–48).

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009a). Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. *Nucleic acids research*, **37**, 1–13. doi:10.1093/nar/gkn923.

Huang, D. W., Sherman, B. T., and Lempicki, R. A. (2009b). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. *Nature protocols*, **4**, 44–57. doi:10.1038/nprot.2008.211.

Huh, W.-K., Falvo, J. V., Gerke, L. C., Carroll, A. S., Howson, R. W., Weissman, J. S., and O'Shea, E. K. (2003). Global analysis of protein localization in budding yeast. *Nature*, **425**, 686–691. doi:10.1038/nature02026.

Huntley, R. P., Binns, D., Dimmer, E., Barrell, D., O'Donovan, C., and Apweiler, R. (2009). QuickGO: a user tutorial for the web-based gene ontology browser. *Database*, **2009**, bap010–bap010. doi:10.1093/database/bap010.

Hur, J., Schuyler, A. D., States, D. J., and Feldman, E. L. (2009). SciMiner: web-based literature mining tool for target identification and functional enrichment analysis. *Bioinformatics*, **25**, 838–840. doi:10.1093/bioinformatics/btp049.

Imai, K., and Nakai, K. (2010). Prediction of subcellular locations of proteins: Where to proceed? *Proteomics*, **10**, 3970–3983. doi:10.1002/pmic.201000274.

Islinger, M., Grille, S., Fahimi, H. D., and Schrader, M. (2012). The peroxisome: an update on mysteries. *Histochemistry and cell biology*, **137**, 547–574. doi:10.1007/s00418-012-0941-4.

Jensen, L. J., and Bork, P. (2010). Ontologies in quantitative biology: a basis for comparison, integration, and discovery. *PLoS biology*, **8**, e1000374. doi:10.1371/journal.pbio.1000374.

Jensen, L. J., Saric, J., and Bork, P. (2006). Literature mining for the biologist: from information retrieval to biological discovery. *Nature Reviews Genetics*, **7**, 119–129. doi:10.1038/nrg1768.

Jensen, P. B., Jensen, L. J., and Brunak, S. (2012). Mining electronic health records: towards better research applications and clinical care. *Nature Reviews Genetics*, **13**, 395–405. doi:10.1038/nrg3208.

Jones, C. E., Brown, A. L., and Baumann, U. (2007). Estimating the annotation error rate of curated GO database sequence annotations. *BMC Bioinformatics*, **8**, 170. doi:10.1186/1471-2105-8-170.

Kell, D. B., and Oliver, S. G. (2004). Here is the evidence, now what is the hypothesis? the complementary roles of inductive and hypothesis-driven science in the post-genomic era. *BioEssays: news and reviews in molecular, cellular and developmental biology*, **26**, 99–105. doi:10.1002/bies.10385.

Kuhn, M., Mering, C. v., Campillos, M., Jensen, L. J., and Bork, P. (2008). STITCH: interaction networks of chemicals and proteins. *Nucleic Acids Research*, **36**, D684–D688. doi:10.1093/nar/gkm795.

Kuhn, M., Szklarczyk, D., Pletscher-Frankild, S., Blicher, T. H., von Mering, C., Jensen, L. J., and Bork, P. (2014). STITCH 4: integration of protein-chemical interactions with user data. *Nucleic Acids Research*, **42**, D401–D407. doi:10.1093/nar/gkt1207.

Kumar, A., Agarwal, S., Heyman, J. A., Matson, S., Heidtman, M., Piccirillo, S., Umansky, L., Drawid, A., Jansen, R., Liu, Y., Cheung, K.-H., Miller, P., Gerstein, M., Roeder, G. S., and Snyder, M. (2002). Subcellular localization of the yeast proteome. *Genes & development*, **16**, 707–719. doi:10.1101/gad.970902.

Küster, B., Mortensen, P., Andersen, J. S., and Mann, M. (2001). Mass spectrometry allows direct identification of proteins in large genomes. *Proteomics*, **1**, 641–650. doi:10.1002/1615-9861(200104)1:5<641::AID-PROT641>3.0.CO;2-R.

Lane, L., Argoud-Puy, G., Britan, A., Cusin, I., Duek, P. D., Evalet, O., Gateau, A., Gaudet, P., Gleizes, A., Masselot, A., Zwahlen, C., and Bairoch, A. (2012). neXtProt: a knowledge platform for human proteins. *Nucleic Acids Research*, **40**, D76–D83. doi:10.1093/nar/gkr1179.

Le Roy, C., and Wrana, J. L. (2005). Clathrin- and non-clathrin-mediated endocytic regulation of cell signalling. *Nature Reviews Molecular Cell Biology*, **6**, 112–126. doi:10.1038/nrm1571.

Lee, K., Chuang, H.-Y., Beyer, A., Sung, M.-K., Huh, W.-K., Lee, B., and Ideker, T. (2008). Protein networks markedly improve prediction of subcellular localization in multiple eukaryotic species. *Nucleic acids research*, **36**, e136. doi:10.1093/nar/gkn619.

Li, J., Newberg, J. Y., Uhlén, M., Lundberg, E., and Murphy, R. F. (2012). Automated analysis and reannotation of subcellular locations in confocal images from the human protein atlas. *PLoS ONE*, **7**, e50514. doi:10.1371/journal.pone.0050514.

Magrane, M., and UniProt Consortium (2011). UniProt knowledgebase: a hub of integrated protein data. *Database*, **2011**, bar009. doi:10.1093/database/bar009.

Magzoub, M., and Miranker, A. D. (2012). Concentration-dependent transitions govern the subcellular localization of islet amyloid polypeptide. *The FASEB Journal*, **26**, 1228–1238. doi:10.1096/fj.11-194613.

Mann, M., and Wilm, M. (1994). Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Analytical Chemistry*, **66**, 4390–4399. doi:10.1021/ac00096a002.

McEntyre, J. R., Ananiadou, S., Andrews, S., Black, W. J., Boulderstone, R., Buttery, P., Chaplin, D., Chevuru, S., Cobley, N., Coleman, L.-A., Davey, P., Gupta, B., Haji-Gholam, L., Hawkins, C., Horne, A., Hubbard, S. J., Kim, J.-H., Lewin, I., Lyte, V., MacIntyre, R., Mansoor, S., Mason, L., McNaught, J., Newbold, E., Nobata, C., Ong, E., Pillai, S., Rebholz-Schuhmann, D., Rosie, H., Rowbotham, R., Rupp, C. J., Stoehr, P., and Vaughan, P. (2011). UKPMC: a full text article resource for the life sciences. *Nucleic Acids Research*, **39**, D58–D65. doi:10.1093/nar/gkq1063.

McQuilton, P., St. Pierre, S. E., Thurmond, J., and the FlyBase Consortium (2011). FlyBase 101 - the basics of navigating FlyBase. *Nucleic Acids Research*, **40**, D706–D714. doi:10.1093/nar/gkr1030.

Mellman, I., and Warren, G. (2000). The road taken: Past and future foundations of membrane traffic. *Cell*, **100**, 99–112. doi:10.1016/S0092-8674(00)81687-6.

Müller, H.-M., Kenny, E. E., and Sternberg, P. W. (2004). Textpresso: An ontology-based information retrieval and extraction system for biological literature. *PLoS Biol*, **2**, e309. doi:10.1371/journal.pbio.0020309.

Nakai, K., and Horton, P. (1999). PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends in Biochemical Sciences*, **24**, 34–35. doi:10.1016/S0968-0004(98)01336-X.

Nakai, K., and Kanehisa, M. (1991). Expert system for predicting protein localization sites in gram-negative bacteria. *Proteins*, **11**, 95–110. doi:10.1002/prot.340110203.

Nakai, K., and Kanehisa, M. (1992). A knowledge base for predicting protein localization sites in eukaryotic cells. *Genomics*, **14**, 897–911.

Osborne, J. D., Flatow, J., Holko, M., Lin, S. M., Kibbe, W. A., Zhu, L. J., Danila, M. I., Feng, G., and Chisholm, R. L. (2009). Annotating the human genome with disease ontology. *BMC Genomics*, **10**, S6. doi:10.1186/1471-2164-10-S1-S6.

Ozawa, T., Sako, Y., Sato, M., Kitamura, T., and Umezawa, Y. (2003). A genetic approach to identifying mitochondrial proteins. *Nature Biotechnology*, **21**, 287–293. doi: 10.1038/nbt791.

O'Donoghue, S. I., Horn, H., Pafilis, E., Haag, S., Kuhn, M., Satagopam, V. P., Schneider, R., and Jensen, L. J. (2010). Reflect: A practical approach to web semantics. *Web Semantics: Science, Services and Agents on the World Wide Web*, **8**, 182–189. doi: 10.1016/j.websem.2010.03.003.

Pafilis, E., Frankild, S. P., Fanini, L., Faulwetter, S., Pavloudi, C., Vasileiadou, A., Arvanitidis, C., and Jensen, L. J. (2013a). The SPECIES and ORGANISMS resources for fast and accurate identification of taxonomic names in text. *PLoS ONE*, **8**, e65390. doi:10.1371/journal.pone.0065390.

Pafilis, E., O'Donoghue, S. I., Jensen, L. J., Horn, H., Kuhn, M., Brown, N. P., and Schneider, R. (2009). Reflect: augmented browsing for the life scientist. *Nature Biotechnology*, **27**, 508–510. doi:10.1038/nbt0609-508.

Pafilis, E., Pavlopoulos, G., Satagopam, V., Papanikolaou, N., Horn, H., Arvanitidis, C., Jensen, L., Schneider, R., and Iliopoulos, I. (2013b). OnTheFly 2.0: A tool for automatic annotation of files and biological information extraction. In *2013 IEEE 13th International Conference on Bioinformatics and Bioengineering (BIBE)*, (pp. 1–4). doi:10.1109/BIBE.2013.6701679.

Pavlopoulos, G. A., Pafilis, E., Kuhn, M., Hooper, S. D., and Schneider, R. (2009). OnTheFly: a tool for automated document-based text annotation, data linking and network generation. *Bioinformatics*, **25**, 977–978. doi:10.1093/bioinformatics/btp081.

Pierleoni, A., Martelli, P. L., Fariselli, P., and Casadio, R. (2007). eSLDB: eukaryotic subcellular localization database. *Nucleic Acids Research*, **35**, D208–D212. doi: 10.1093/nar/gkl775.

Poux, S., Magrane, M., Arighi, C. N., Bridge, A., O'Donovan, C., Laiho, K., and The UniProt Consortium (2014). Expert curation in UniProtKB: a case study on dealing with conflicting and erroneous data. *Database*, **2014**, bau016–bau016. doi:10.1093/database/bau016.

Quintana, C., Bellefqih, S., Laval, J. Y., Guerquin-Kern, J. L., Wu, T. D., Avila, J., Ferrer, I., Arranz, R., and Patiño, C. (2006). Study of the localization of iron, ferritin, and hemosiderin in alzheimer's disease hippocampus by analytical microscopy at the subcellular level. *Journal of Structural Biology*, **153**, 42–54. doi:10.1016/j.jsb.2005.11.001.

Rappaport, N., Nativ, N., Stelzer, G., Twik, M., Guan-Golan, Y., Iny Stein, T., Bahir, I., Belinky, F., Morrey, C. P., Safran, M., and Lancet, D. (2013). MalaCards: an integrated compendium for diseases and their annotation. *Database*, **2013**, bat018–bat018. doi:10.1093/database/bat018.

Rastogi, S., and Rost, B. (2010). LocDB: experimental annotations of localization for homo sapiens and arabidopsis thaliana. *Nucleic Acids Research*, **39**, D230–D234. doi:10.1093/nar/gkq927.

Rebholz-Schuhmann, D., Arregui, M., Gaudan, S., Kirsch, H., and Jimeno, A. (2008). Text processing through web services: calling whatizit. *Bioinformatics*, **24**, 296–298. doi:10.1093/bioinformatics/btm557.

Rebholz-Schuhmann, D., Oellrich, A., and Hoehndorf, R. (2012). Text-mining solutions for biomedical research: enabling integrative biology. *Nature Reviews Genetics*, **13**, 829–839. doi:10.1038/nrg3337.

Schijvenaars, B. J., Mons, B., Weeber, M., Schuemie, M. J., Mulligen, E. M. v., Wain, H. M., and Kors, J. A. (2005). Thesaurus-based disambiguation of gene symbols. *BMC Bioinformatics*, **6**, 149. doi:10.1186/1471-2105-6-149.

Schneider, M., Lane, L., Boutet, E., Lieberherr, D., Tognolli, M., Bougueleret, L., and Bairoch, A. (2009). The UniProtKB/Swiss-Prot knowledgebase and its plant proteome annotation program. *Journal of Proteomics*, **72**, 567–573. doi:10.1016/j.jprot.2008.11.010.

Schrader, M., and Yoon, Y. (2007). Mitochondria and peroxisomes: are the 'big brother' and the 'little sister' closer than assumed? *BioEssays: news and reviews in molecular, cellular and developmental biology*, **29**, 1105–1114. doi:10.1002/bies.20659.

Schriml, L. M., Arze, C., Nadendla, S., Chang, Y.-W. W., Mazaitis, M., Felix, V., Feng, G., and Kibbe, W. A. (2011). Disease ontology: a backbone for disease semantic integration. *Nucleic Acids Research*, **40**, D940–D946. doi:10.1093/nar/gkr972.

Sedek, M., and Strous, G. J. (2013). SUMOylation is a regulator of the translocation of jak2 between nucleus and cytosol. *Biochemical Journal*, **453**, 231–239. doi: 10.1042/BJ20121375.

Simpson, J. C., Wellenreuther, R., Poustka, A., Pepperkok, R., and Wiemann, S. (2000). Systematic subcellular localization of novel proteins identified by large-scale cDNA sequencing. *EMBO reports*, **1**, 287–292. doi:10.1093/embo-reports/kvd058.

Skunca, N., Altenhoff, A., and Dessimoz, C. (2012). Quality of computationally inferred gene ontology annotations. *PLoS Computational Biology*, **8**. doi:10.1371/journal.pcbi.1002533.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., OBI Consortium, Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.-A., Scheuermann, R. H., Shah, N., Whetzel, P. L., and Lewis, S. (2007). The OBO foundry: coordinated evolution of ontologies to support biomedical data integration. *Nature biotechnology*, **25**, 1251–1255. doi: 10.1038/nbt1346.

Soldatos, T. (2009). *Automated extraction of gene set function and drug side effect information from biomedical literature*. Ph.D. thesis, University of Dundee, Dundee.

Sprenger, J., Lynn Fink, J., Karunaratne, S., Hanson, K., Hamilton, N. A., and Teasdale, R. D. (2007). LOCATE: a mammalian protein subcellular localization database. *Nucleic Acids Research*, **36**, D230–D233. doi:10.1093/nar/gkm950.

Sterk, P., Kersey, P. J., and Apweiler, R. (2006). Genome reviews: standardizing content and representation of information about complete genomes. *Omics: a journal of integrative biology*, **10**, 114–118. doi:10.1089/omi.2006.10.114.

Supek, F., Bošnjak, M., Škunca, N., and Šmuc, T. (2011). REVIGO summarizes and visualizes long lists of gene ontology terms. *PloS one*, **6**, e21800. doi:10.1371/journal.pone.0021800.

Tanabe, L., and Wilbur, W. J. (2002). Tagging gene and protein names in biomedical text. *Bioinformatics (Oxford, England)*, **18**, 1124–1132.

Tanz, S. K., Castleden, I., Hooper, C. M., Vacher, M., Small, I., and Millar, H. A. (2012). SUBA3: a database for integrating experimentation and prediction to define the

SUBcellular location of proteins in arabidopsis. *Nucleic Acids Research*, **41**, D1185–D1191. doi:10.1093/nar/gks1151.

Taylor, C. F. (2007). Standards for reporting bioscience data: a forward look. *Drug Discovery Today*, **12**, 527–533. doi:10.1016/j.drudis.2007.05.006.

Uhlen, M., Oksvold, P., Fagerberg, L., Lundberg, E., Jonasson, K., Forsberg, M., Zwahlen, M., Kampf, C., Wester, K., Hober, S., Wernerus, H., Björling, L., and Ponten, F. (2010). Towards a knowledge-based human protein atlas. *Nature Biotechnology*, **28**, 1248–1250. doi:10.1038/nbt1210-1248.

UniProt Consortium (2010). The universal protein resource (UniProt) in 2010. *Nucleic Acids Research*, **38**, D142–D148. doi:10.1093/nar/gkp846.

Van Auken, K., Fey, P., Berardini, T. Z., Dodson, R., Cooper, L., Li, D., Chan, J., Li, Y., Basu, S., Muller, H.-M., Chisholm, R., Huala, E., Sternberg, P. W., and the WormBase Consortium (2012). Text mining in the biocuration workflow: applications for literature curation at WormBase, dictyBase and TAIR. *Database*, **2012**, bas040. doi:10.1093/database/bas040.

von Heijne, G. (1986). A new method for predicting signal sequence cleavage sites. *Nucleic acids research*, **14**, 4683–4690.

von Isenburg, M. (2007). HubMed. *Journal of the Medical Library Association*, **95**, 95–97.

von Mering, C., Jensen, L. J., Snel, B., Hooper, S. D., Krupp, M., Foglierini, M., Jouffre, N., Huynen, M. A., and Bork, P. (2005). STRING: known and predicted protein-protein associations, integrated and transferred across organisms. *Nucleic acids research*, **33**, D433–437. doi:10.1093/nar/gki005.

von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S. G., Fields, S., and Bork, P. (2002). Comparative assessment of large-scale data sets of protein–protein interactions. *Nature*, **417**, 399–403. doi:10.1038/nature750.

Watts, C. (2001). Antigen processing in the endocytic compartment. *Current Opinion in Immunology*, **13**, 26–31. doi:10.1016/S0952-7915(00)00177-1.

Watts, C. (2012). The endosome–lysosome pathway and information generation in the immune system. *Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics*, **1824**, 14–21. doi:10.1016/j.bbapap.2011.07.006.

Wu, C. H., Yeh, L.-S. L., Huang, H., Arminski, L., Castro-Alvear, J., Chen, Y., Hu, Z., Kourtesis, P., Ledley, R. S., Suzek, B. E., Vinayaka, C. R., Zhang, J., and Barker, W. C. (2003). The protein information resource. *Nucleic Acids Research*, **31**, 345–347. doi:10.1093/nar/gkg040.

Yoneda, Y. (2000). Nucleocytoplasmic protein traffic and its significance to cell function. *Genes to cells: devoted to molecular & cellular mechanisms*, **5**, 777–787.

Zhang, L., Shimoji, M., Thomas, B., Moore, D. J., Yu, S.-W., Marupudi, N. I., Torp, R., Torgner, I. A., Ottersen, O. P., Dawson, T. M., and Dawson, V. L. (2005). Mitochondrial localization of the parkinson's disease related protein DJ-1: implications for pathogenesis. *Human Molecular Genetics*, **14**, 2063–2073. doi:10.1093/hmg/ddi211.

Ziauddin, J., and Sabatini, D. M. (2001). Microarrays of cells expressing defined cDNAs. *Nature*, **411**, 107–110. doi:10.1038/35075114.

Šarić, J., Jensen, L. J., Ouzounova, R., Rojas, I., and Bork, P. (2006). Extraction of regulatory gene/protein networks from medline. *Bioinformatics*, **22**, 645–650. doi:10.1093/bioinformatics/bti597.

# Publications

Binder, J. X., Pletscher-Frankild, S., Tsafou, K., Stolte, C., O'Donoghue, S. I., Schneider, R., and Jensen, L. J. (2014). COMPARTMENTS: unification and visualization of protein subcellular localization evidence. *Database: the journal of biological databases and curation*, **2014**, bau012. doi:10.1093/database/bau012.

Budd A., Dinkel H., Corpas M., Fuller J. C., Rubinat L., Devos D. P., Khoueiry P. H., Förstner K. U., Georgatos F., Rowland F., Sharan M., Binder J. X., Grace T., Traphagen K., Gristwood A., and Wood N. T. (2014). 10 simple rules for organizing an unconference. *Submitted*.

Schriml L. M., Arze C., Nadendla S., Chang Y. W., Mazaitis M., Binder, J. X., Pletscher-Frankild, S., Felix V., Feng G., and Kibbe W. A. (2014). Disease Ontology: Towards an integrated disease knowledgebase. *In preparation*.