National and Kapodistrian University of Athens, Medical School

BSRC “Alexander Fleming” Institute, Bioinformatics and Integrative Biology Lab

Master of Science
“MSc Molecular Biomedicine”

Thesis

Literature mining and network analysis in Biology

Zafeiropoulou Sofia

20191049

Main supervisor: Dr. Georgios A. Pavlopoulos

Researcher B’ - BSRC Alexander Fleming

Athens
2021

Declaration of Originality of Diploma Thesis - Disclaimer

“I declare responsibly that this Diploma Thesis for my Masters Degree in Molecular Biomedicine - Full-time Degree in Department of Physiology in Medical School of Athens has been conducted by me and has not been submitted in another postgraduate or undergraduate degree in Greece or abroad. The sources I referred to for this thesis are listed, giving full references to the authors, including any sources that may have been used online. In any case of untrue or inaccurate content, I am subject to the consequences provided by the Master's Degree Program in Molecular Biomedicine and the provisions of Greek Law.”

The Declarant
Full Name: Zafeiropoulou Sofia
Registration Number: 20191049
Signature:

Υπεύθυνη Δήλωση Πρωτοτυπίας Διπλωματικής Εργασίας - ΑΠΟΠΟΙΗΣΗ ΕΥΘΥΝΗΣ

«Δηλώνω υπεύθυνα ότι η συγκεκριμένη Διπλωματική Εργασία για τη λήψη του μεταπτυχιακού τίτλου σπουδών του ΔΠΜΣ στη Μοριακή Βιοϊατρική: Μηχανισμοί Ασθενειών, Μοριακές και Κυτταρικές θεραπείες και Βιοκαινοτομία – Πλήρους Φοίτησης του Τμήματος Φυσιολογίας της Ιατρικής Σχολής Αθηνών, έχει συγγραφεί από εμένα προσωπικά και δεν έχει υποβληθεί ούτε έχει εγκριθεί στο πλαίσιο κάποιου άλλου μεταπτυχιακού ή προπτυχιακού τίτλου σπουδών, στην Ελλάδα ή στο εξωτερικό. Οι πηγές στις οποίες ανέτρεξα για την εκπόνηση της συγκεκριμένης διπλωματικής αναφέρονται στο σύνολό τους, δίνοντας πλήρεις αναφορές στους συγγραφείς, συμπεριλαμβανομένων και των πηγών που ενδεχομένως χρησιμοποιήθηκαν από το διαδίκτυο. Σε κάθε περίπτωση αναληθούς ή ανακριβούς δηλώσεως, υπόκειμαι στις συνέπειες που προβλέπονται στον Κανονισμό Σπουδών του Μεταπτυχιακού Προγράμματος Σπουδών στην Μοριακή Βιοϊατρική: Μηχανισμοί Ασθενειών, Μοριακές και Κυτταρικές θεραπείες και Βιοκαινοτομία και στις διατάξεις που προβλέπει η Ελληνική και Κοινοτική Νομοθεσία περί πνευματικής ιδιοκτησίας».

Η Δηλούσα
Ονοματεπώνυμο: Ζαφειροπούλου Σοφία
Αριθμός Μητρώου: 20191049
Υπογραφή:

Acknowledgements

First of all, I would like to express my sincere gratitude to the National and Kapodistrian University of Athens as well as to the Biomedical Sciences Research Center “Alexander Fleming” and especially Dr. George Kollias for accepting my application and letting me be part of this incredible Master’s program.

This thesis would not have been possible without my supervisor Dr. Georgios A. Pavlopoulos who offered me the opportunity to participate in his research group and get acquainted with bioinformatics. His dedicated support and guidance were an invaluable motivation throughout my MSc thesis.

Furthermore, I would like to say a special thank you to Dr. Fotis A. Baltoumas for his participation and assistance at every stage of this project as well as for his patient support, great encouragement and programming suggestions.

My sincere thanks also to Dr. Evangelos Karatzas, Dr. Savvas Paragkamian, Foteini Thanati, Dr. Ioannis Iliopoulos, Dr. Aristides G. Eliopoulos, Dr. Reinhard Schneider, Dr. Lars Juhl Jensen and Dr. Evangelos Pafilis for their collaboration in the project and insightful comments.

I also would like to express my very great appreciation to all of my respected professors and lecturers of this Master’s program and especially Dr. Christoforos Nikolaou for dedicating many hours of teaching as well as for his valuable and constructive suggestions.

Last but not least, I would also like to express my deepest gratitude to my family for their unconditional support and continuous encouragement. Finally, I could not have completed this dissertation without the support of my friends, and especially Giannis Kassionis, who provided stimulating discussions as well as a number of helpful comments and suggestions.

Literature mining and network analysis in Biology

Zafeiropoulou Sofia

Thesis committee:

Supervisor: **Georgios A. Pavlopoulos**	*Member A':* **Christophoros Nikolaou**	*Member B':* **George Kollias**
Researcher B'	Researcher B'	Professor

Abstract

Undoubtedly, text constitutes an essential type of data within the biomedical field, as the ever-increasing knowledge is still mostly expressed and stored in the form of unstructured information. Vast numbers of biomedical and biology reports, providing valuable insights about new discoveries, constantly supplement the already overwhelming amount of the scientific literature. However, the inefficiency in retrieving and processing important information both quickly and accurately is an inevitable challenge when handling large quantities of unstructured data. Hence, the automated extraction and analysis of biomedical terms from documents is becoming an absolute necessity. The particular thesis presents OnTheFly$^{2.0}$, a web-based, versatile tool dedicated to the extraction and subsequent analysis of biomedical terms from individual files. More specifically, OnTheFly$^{2.0}$ supports different file formats, including plain texts, Office documents, PDF files or images, enabling simultaneous file handling. The integration of the EXTRACT tagging service allows the implementation of Named Entity Recognition (NER) for genes/proteins, chemical compounds, organisms, tissues, environments, diseases, phenotypes and Gene Ontology terms, as well as the generation of popup windows which provide concise, context related information about the identified term, accompanied by links to various databases. Once named entities, such as proteins, genes and chemicals are identified, they can be further explored via functional and publication enrichment analysis or be associated with diseases and protein domains reporting from protein family databases. Finally, visualization of protein-protein and protein-chemical associations is possible through the generation of interactive networks from the STRING and STITCH services, respectively. In order to demonstrate the potential and efficiency of OnTheFly$^{2.0}$, biomarkers of severe COVID-19 with clinical significance were retrieved from six published articles and combined in a meta-analysis case study. Interestingly, several inflammatory and senescence pathways that impact COVID-19 pathogenesis have been unraveled. OnTheFly$^{2.0}$ currently supports 197 species and is available at http://onthefly.pavlopouloslab.info.

Aim and objectives

The aim of this thesis is to present OnTheFly$^{2.0}$, a user-friendly, web-based application which not only applies NER pipeline for the identification of biological named entities in a collection of documents provided, but also enables further network and enrichment analyses. Text mining and biomedical entity extraction can be performed in a plethora of different file formats, including text documents, spreadsheets and image files, while the implementation of NER methods facilitates the recognition and retrieval of various biological and biomedical terms. A comprehensive dataset consisting of both extracted protein and chemical entities can be generated for functional enrichment analysis, related literature finding, associations with diseases and protein domain reporting from protein family databases. In addition to the previously mentioned analyses, protein-protein and protein-chemical interaction networks are generated via the STRING and the STITCH databases respectively, allowing the visualization of associations between the collected entities.

Table of contents

CHAPTER 1: Literature mining

The introduction of biomedical research to advanced high-throughput technologies and large-scale experiments has led to the emergence of the "*omics era*", characterized by an unbridled growth of heterogeneous collection of raw biological data, including genomic sequences, expression and metagenomic profiles as well as proteomic measurements [1]. A substantial proportion of the resulting biological discoveries and advancements is communicated mostly by means of scientific publications and reports, in an electronic text-based format, enabling the use of natural human language as a way to express and transmit information [2]. Currently, PubMed integrates more than 32 million publications, PubMed Central contains a total of 6 million articles, while there are over 27 million references in Medline as of April 2021.

However, the aforementioned search engines, whose aim is the facilitation of traditional information retrieval, are characterized by the keyword-based approaches, regularly resulting in a plethora of records, often not sorted by relevance [3]. Therefore, even though harnessing the easily accessible, yet vast, biomedical literature could reveal biomedical concepts and ultimately lead to the acquisition of new information, it is hardly surprising that manual handling and processing often becomes a tedious and error-prone task [2]. In an attempt to address the issues mentioned above, Natural Language Processing (NLP) techniques, including *Named Entity Recognition (NER)*, *Information Extraction (IE)*, *Question/Answer (QA)* and *Text Summarization (TS)* have emerged as a potential solution and serve as bridges between computers and human languages to enable efficient, systematic and automated discovery of knowledge [4].

1.1 Natural Language Processing (NLP) techniques in bioinformatics

1.1.1 Named Entity Recognition (NER)

The vagueness in human communication due to colloquialisms, abbreviations and misspellings, as well as the difficulty in defining precisely the concept of *term*, are often responsible for the challenges in analysis of natural language. This variability is inevitably reflected in the scientific literature, as local dialects, jargons and idiosyncratic nomenclatures are shaping a heterogenous text corpora [4]. Despite these impediments, natural language processing (NLP), a research subfield of artificial intelligence in computer science, aims to develop computational models that simulate human linguistic abilities, by utilizing machine learning algorithms and programs [5]. Typically, semantic, syntax and content information are the main aspects of a text, thus syntactic and semantic analysis constitute the predominant techniques, that coupled with more articulated activities, such as the utilization of lexical resources (e.g. lexicons, vocabularies, thesauri, ontologies) are employed in order to accomplish NLP tasks [5]. *Machine Translation* (*MT*), *Information Extraction* (*IE*), *Information Retrieval* (*IR*), *Automatic Text Summarization* (*ATS*), *Question-Answering System*, *Parsing*, *Sentiment Analysis*, *Natural Language Understanding* (*NLU*) and *Natural Language Generation* (*NLG*) comprise the main tasks in NLP which ultimately produce a semantic representation of data retrieved from unstructured corpora [6].

Particularly, *Information Extraction (IE)* refers to a widely used NLP technique which aims at the automatic identification of structured information within unstructured or semi-structured textual data [7]. Since generally IE constitutes one of the primary stages in the analysis pipeline, the effective conversion of unstructured collected data and their efficient incorporation into machine readable databases determines the performance of the higher level tasks [7,8]. The decomposition of IE into various separate subtasks, including *Named Entity Recognition (NER)*, *Relation Extraction (RE)* and *Coreference Resolution (CR)*, illustrates the functionality of IE systems [6].

One of the first steps towards information retrieval, question answering and co-reference resolution is *Named Entity Recognition (NER)*, also referred to as entity identification or entity extraction [9]. NER consists a fundamental NLP subtask -especially in the semantic part- which automatically identifies and extracts selective information (named entities) within an unstructured text corpus, classifying them into predefined categories [9]. *Named Entity Recognition and Classification (NERC)* is an alternative, equally utilized term of NER, pinpointing the subsequent classification of identified entities [9]. The spectrum of the detected entities used for term normalization can be significantly broad, depending on the

field of interest, and may include words such as person names, cities, time expression as well as scientific or technical terms. In the biomedical domain, NER is usually referred to as *Biomedical Named Entity Recognition (BioNER)* and is considered an indispensable tool for the extraction of biomedically relevant named entities, including genes/proteins, chemicals/drugs, diseases, adverse effects, mutations/SNPs, species, tissues, metabolites or pathways [10].

Notably, approaches to develop a NER system can be categorized into four classes: *Rule-based*, *Dictionary-based*, *Machine learning (ML)-based* and *Hybrid* [10] (*Figure 1*). However, it is not unusual for a NER system to combine more than one of the aforementioned categories. Firstly, *Rule-based approaches* typically use a set of manually crafted linguistic, grammatical and syntactic rules and patterns in combination with dictionaries to extract and classify named entities [10,11]. A major advantage of this approach concerns the employment of context-based rules and domain specific features to distinguish with sufficient accuracy multiple named entities [11]. However, the high cost of rule maintenance by experts, the lack of portability and adaptability across different disciplines as well as the requirement of available resources are considered the main limitations of Rule-based NER systems [12].

Dictionary-based NER approach relies on lexicon resources, utilizing stored lists of terms in dictionaries for entity identification in given texts [11]. These systems recognize and extract specific entities either after searching the more relevant term in the dictionary or upon implementing *string-matching* algorithms (exact matching and flexible or approximate matching) [10]. They are particularly effective in queries, where the available contexts could be very limited due to their high precision [13]. The completeness, quality, constant updating and maintenance of the provided dictionary is a prerequisite for the robustness of the dictionary-based approaches, whose apparent simplicity can sometimes be outweighed by two major limitations [10]. Firstly, the usefulness of these approaches is significantly degraded in cases of spelling errors or/and variations in the text, while the generation of new words or the identification of the same term as different depending on the context is quite frequent [11,13].

Furthermore, the *Machine learning (ML)-based approaches*, as the name indicates, depend on machine learning algorithms and statistical models to recognize specific entities in a document, without requiring a rule set or listed terms in a dictionary [10,11]. More specifically, Machine learning (ML) is a branch of artificial intelligence (AI) focused on automating the learning of systems and improving their accuracy and efficiency over time by implementing sophisticated algorithms and statistics on large-scale datasets [14]. Typically, ML approaches can be divided into three main categories; *Supervised*, *Semi-supervised* and *Unsupervised* learning.

The main purpose of *supervised learning based* approaches is to train learning algorithms to map the relationship between a given set of inputs and outputs in order to be able to distinguish negative from positive examples and ideally output a prediction from unseen data [10,15]. Example labeled instances of named entities, referred to as training data, are a prerequisite to construct a statistical model for effective algorithm training. In recent years, statistical methods based on supervised learning, including Hidden Markov Model (HMM) [16], Support Vector Machine (SVM) [17] and Maximum Entropy Markov Model (MEMM) [16], have received the most research interest.

However, obtaining reliable and unbiased labeled training data from a large, heterogeneous dataset is usually an expensive, laborious and time-consuming task. Therefore, the implementation of *semi-supervised learning* approach could be proposed as a promising solution to the previously mentioned problems, due to its partial independence from labeled data [10]. More precisely, semi-supervised learning allows the algorithms to learn by using a small amount of labeled training examples (seed) for subsequent tagging of the unlabeled data provided [15]. The training of the system and the generation of more labeled examples based on the algorithm's outcome are continuous processes, until the acquisition of an adequate accuracy [15]. "Bootstrapping" is considered as one of the most prominent methods for improving the efficiency of semi-supervised learning algorithms [18].

In *unsupervised learning* the model is trained without any labeled or classified data, thus is mainly proposed in cases where text annotation is extremely limited or absent [12]. The goal of these programs is to generate representation models from data that can be utilized for three main tasks: clustering, association, and dimensionality reduction [12].

Finally, *the Hybrid method* implements a combination of multiple NER approaches, including rule-based and machine learning methods, harnessing their advantages in order to perform more accurately and provide better results.

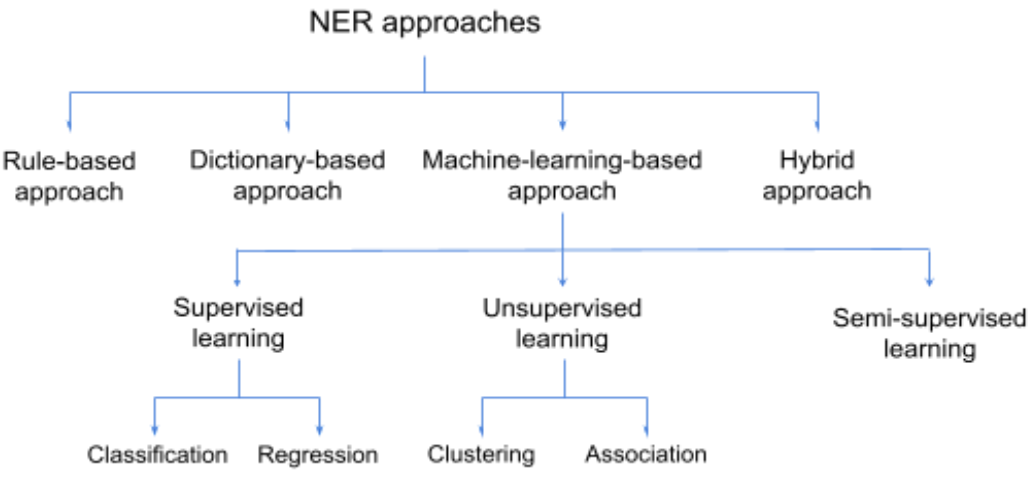


Figure 1: Types of NER approaches

1.1.2 Co-occurrence analysis

The interpretation of results or the use of certain data analysis techniques is undoubtedly challenging without accessing the meaning of words. Therefore, the detection of semantic similarity and relatedness between two ontologies involves mainly the application of co-occurrence methodologies which focus on the analysis of paired data, existing on the same collection unit [19,20]. Apart from punctuation signs, all the words and numbers are included in the analysis, evaluating both their number of paired presence and their concordances [21].

Notably, the use of the collected textual data by machine learning models requires their numerical representation, also known as *text vectorization*, a fundamental process of machine learning for the data analysis. Thus, data extraction is the basis for the generation of co-occurrence matrices or word vectors, which subsequently produce tables or graphs, enabling the analysis process [22]. More specifically, co-occurrence matrix is a square, asymmetric matrix, whose aim is to provide the associations between notions. It presents the number of times ($n_{i,j}$) where word *j* is inside a unit of text which contains word *i*, taking into account the total number of identified bioentities [21] (*Figure 2b*). Weighted and directed graphs are usually generated by these matrices where the vertices correspond to terms and the values $n_{i,j}$ are the weights of the edges [21] (*Figure 2c*).

Particularly, co-occurrence networks are graphic visualizations of associations between entities that appear in text data [23]. Each vertice (V) of the corresponding graph G = (V, E) represents a term/bioentity, while edges (E) depict semantic relationships between these entities, implying a functional relevance as well . Co-occurrence is often defined based on desired criteria, such as the segmentation of a text. Hence, co-occurrence networks can be generated for terms identified in any unit of text, including sentence, paragraph, section or even abstract [24]. In case of *sentence-based* co-occurrence, the connections are created between paired entities that exclusively appear in the same sentence, while *abstract-based* co-occurrence requires the coexistence of terms in the same abstract text [24]. Both of the aforementioned network types are typically undirected graphs which represent the paired presence of biomedical terms within a specified text corpus, inferring a potential action between nodes. However, *semantic co-occurrence networks* depict the functionality between two entities via the interactions of nodes with multiple and directed edges [24].

Nonetheless, it is not uncommon for graphs to contain a great number of interconnected nodes, resulting from large data volumes which are produced by co-occurrence analysis. As a consequence, networks inevitably become dense and indecipherable, complicating the extraction of meaningful information. To this end, the implementation of high-quality clustering algorithms and visualization techniques by

specialized tools aims to accomplish an organization of functionally related terms in order to decrease the complexity of the graph and facilitate the interpretation of the results [24].

a

Abstract

FoxM1 is an oncogenic Forkhead transcription factor that is overexpressed in ovarian cancer. However, the mechanisms by which FoxM1 is deregulated in ovarian cancer and the extent to which FoxM1 can be targeted in ovarian cancer have not been reported previously. In this study, we showed that MDM2 inhibitor Nutlin-3 upregulated p53 protein and downregulated FoxM1 expression in several cancer cell lines with wild type TP53 but not in cell lines with mutant TP53. FoxM1 downregulation was partially blocked by cycloheximide or actinomycin D, and pulse-chase studies indicate Nutlin-3 enhances FoxM1 mRNA decay. Knockdown of p53 using shRNAs abrogated the FoxM1 downregulation by Nutlin-3, indicating a p53-dependent mechanism. FoxM1 inhibitor, thiostrepton, induces apoptosis in cancer cell lines and enhances sensitivity to cisplatin in these cells. Thiostrepton downregulates FoxM1 expression in several cancer cell lines and enhances sensitivity to carboplatin in vivo. Finally, FoxM1 expression is elevated in nearly all (48/49) ovarian tumors, indicating that thiostrepton target gene is highly expressed in ovarian cancer. In summary, the present study provides novel evidence that both amorphic and neomorphic mutations in TP53 contribute to FoxM1 overexpression and that FoxM1 may be targeted for therapeutic benefits in cancers. *PMID: 25426548*

b

	FoxM1	NMD2	Nutlin-3	p53	cycloheximide	actinomycin D	thiostrepton	cisplatin	carboplatin
FoxM1	1	1	1	1	1	1	1	1	1
NMD2	1	1	1	1	0	0	0	0	0
Nutlin-3	1	1	1	1	1	1	0	0	0
p53	1	1	1	1	0	0	0	0	0
cycloheximide	1	0	1	0	0	0	0	0	0
actinomycin D	1	0	1	0	0	0	0	0	0
thiostrepton	1	0	0	0	0	0	0	0	0
cisplatin	1	0	0	0	0	0	0	0	0
carboplatin	1	0	0	0	0	0	0	0	0

c

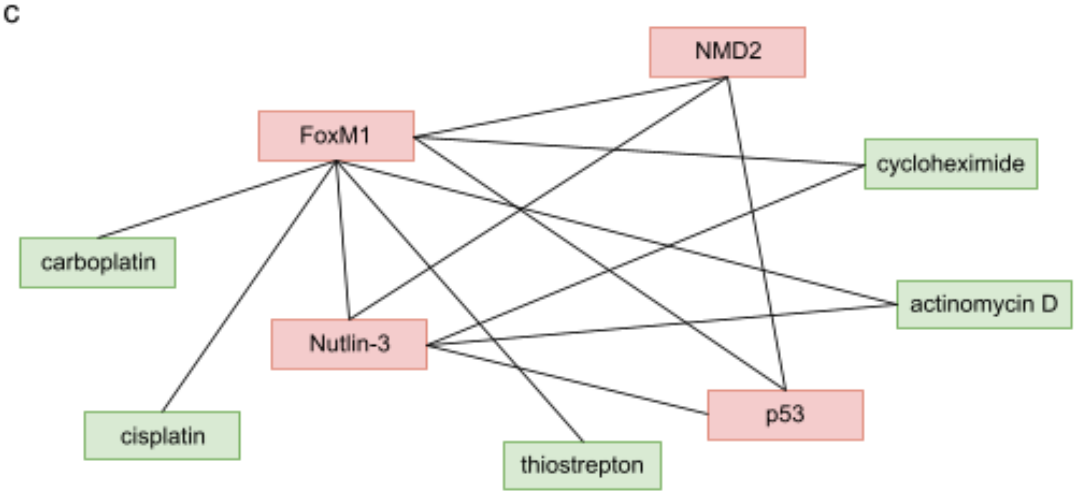


Figure 2: (a) Protein names are highlighted in red and drug names in green for abstract [25]. **(b)** sentence-based co-occurrence matrix presenting the interconnections between terms of the abstract. **(c)** Sentence-base co-occurrence network. All the entities that appear in the same sentence of the abstract are connected with undirected edges.

1.1.3 Term Frequency-Inverse Document Frequency (TF-IDF)

As previously mentioned, co-occurrence analysis and particularly the resulting matrices indicate the frequency of coexistence of terms within a specified text unit. However, the relation between identified entities can not be easily deduced merely by raw frequency measures, due to the lack of discrimination and partiality [26]. For instance, articles and pronouns are commonly used and could co-occur with almost any kind of word, hence the determination of word associations is doubtful. It is hardly surprising that the more frequent the coexistence of two entities, the more important the entities are. Yet words that are ubiquitous, such as articles, tend to be insignificant [26]. Thus, the problem of quantifying the significance of terms arises.

One commonly used measure of word importance that is often applied in Information Retrieval (IR), Text Mining and other NLP tasks to address the aforementioned issue is *Term Frequency-Inverse Document Frequency (TF-IDF)*. TF-IDF is a numerical statistic that evaluates the relevance of terms to a document in a collection or corpus [27]. By definition, TF-IDF multiples two different quantities: *Term Frequency* that weights the occurrence frequency of a term in a document or in a corpus and *Inverse Document Frequency* which can be interpreted as the amount of information a word provides [27]. The normalized TF value is obtained by the division of the raw frequency of a term *t* in a document *d* by the total number of terms in the document *d*, and then often the logarithm of the quotient is calculated (*Figure 3a*) [27]. IDF is the logarithm of the ratio of the total number of documents in collection, by the number of documents in which term *t* appears [26]. Consequently, the higher the occurrence of a term in documents, the lower the weight that is assigned to this term, while increased importance (IDF) is assigned to the rarely occurring terms (*Figure 3b*) [27].

a) TF(*t*,*d*) = Number of times a term *t* appears in a document *d* / Total number of terms in document *d*

b) IDF(*t*) = log_e (Total number of documents / Number of documents that contain term *t*)

Figure 3: (a) Term Frequency (TF) fraction. **(b)** Inverse Document Frequency (ITF) fraction

The multiplication of Term Frequency (TF) and Inverse Document Frequency (IDF) results in the TF-IDF score of a word in a document. Despite the numerous variations of calculation, the TF-IDF weighted value $w_{t,d}$ for the word *t* in the document *d* is usually defined as:

$$w_{t,d} = tf_{t,d}\, x\, log_{10}(\frac{N}{df_t})$$ [28].

According to the previous mathematical equation, TF-IDF score increases proportionally to the occurrence of a word in a document, indicating its importance in that particular document [28]. Generally, even though all terms in a document are assigned a TF-IDF value, the extraction of the most representative and significant ones is determined by a specified threshold, below which the words are regarded as irrelevant and discarded [28].

The TF-IDF weighting is considered among the most popular term-weighting schemes and is used to extract features across various NLP applications. The usefulness of this technique lies mainly in its computational efficiency and the fact that manual annotation is not required. Furthermore, TF-IDF assists the empowering of more complicated NLP algorithms and query retrieval systems, due to the simplicity of its encoding [29]. For example, in Information Retrieval systems, TF-IDF complements text mining and search algorithms, delivering the most relevant results according to a particular search query. However, a main shortage is its failure to detect the semantic-sensitive content, including the position of a term in the text and its co-occurrences with other words [29].

1.2 Literature review of computational tools

A variety of computational tools and web services that implement NLP and NER techniques have been proposed, including EXTRACT [30], PubTator [31], HunFlair [32], LitVar [33], Taxonfinder (https://github.com/pleary/node-taxonfinder) and Tesseract 4.0 (https://github.com/tesseract-ocr/tesseract). More specifically, EXTRACT 2.0, an easy-to-use, interactive annotation tool, offers the opportunity of versatile and browser-based annotation of texts as well as identification and extraction of bioentity terms, by pipelining Named Entity Recognition methods. EXTRACT is able to detect various genes/proteins, chemical compounds, organisms, environments, tissues, diseases, phenotypes and Gene Ontology terms mentioned in HTML pages exclusively, including PubMed abstracts, full-text journal articles and web pages. The main feature of this tool, the bookmarklet, is a browser bookmark containing a JavaScript script allowing both selected-text-based entity extraction and full-page tagging. In addition to annotation, EXTRACT enables the collection and mapping of identified terms to their corresponding ontology/taxonomy entries via another important component called popup.

PubTator Central (PTC) is a user-friendly, web-based tool for the automated annotation of bioentities, including genes/proteins, genetic variants, diseases, chemicals, species and cell lines, in 29 million PubMed abstracts and in 3 million PubMed Central full text biomedical articles. The annotation depends on integration of text-mining systems and disambiguation modules based on deep learning. Currently, the PTC web interface enables the generation of full text corpus and the visualization of each annotated document, while annotations are freely downloadable in various formats (XML, JSON and tab delimited) via the online interface, a RESTful web service and bulk FTP.

Moreover, HunFlair is a freely accessible and easily installed biomedical Named Entity Recognition (NER) tagger, incorporating 23 biomedical NER corpora. HunFlair accurately identifies five biomedical entities (cell lines, chemicals, diseases, genes and species) by implementing a character-level language model, pre-trained in a cross-corpus setting of approximately 24 million biomedical abstracts and 3 million full texts. Text parsing, document classification, hedge detection and the use of other language models are available, since HunFlair is integrated into the NLP framework Flair.

LitVar is another tool which provides a graphical web interface and is focused on the identification and extraction of standardized variant information obtained from more than 27 million PubMed abstracts, up to 1.8 million full-text articles from PubMed Central Open Access Subset, dbSNP [34], and ClinVar [35]. By performing Named Entity Recognition (NER) on both abstracts and full-text articles via tmVar [36], LitVar enables the normalization and standardization of multiple names of the same variant, while the implementation of text

mining techniques provides associations between variants and other bioentities, such as diseases and chemicals/drugs.

Besides the aforementioned biomedical research driven applications, a web tool that focuses exclusively on taxonomic entity extraction is Taxonfinder. By employing a dictionary-based approach, morphological analysis and Levenshtein Distance algorithm. Taxonfinder is able to identify various latin scientific organism names including Kingdom, Phylum, Class, Order, Family, Genus, Species, Subspecies, in the literature. The application can be installed and run locally or can be accessed programmatically via an API.

Finally, Tesseract 4.0 is an updated, freely accessible, and open-source command line Optical Character Recognition (OCR) engine, which has been sponsored by Google since 2006 and currently recognizes over 100 languages. It is a well-documented engine, written in C/C++ with Unicode (UTF-8) support and it has recently implemented a Long-Short Term Memory (LSTM) OCR module and a new neural net (LSTM) based OCR engine which is focused on line recognition. Notably, tesseract returns various output formats, including plain text, hOCR (HTML), PDF, invisible-text-only PDF and TSV, while ALTO (XML) output is experimentally supported in the master branch.

Altogether, these tools leverage unstructured knowledge to detect and extract named entity mentions, such as genes/proteins, genetic variants, diseases, chemical compounds, organisms, diseases and cell lines in biomedical documents.

CHAPTER 2: Functional enrichment analysis

The introduction of high-throughput technologies in biology has necessitated the development of novel data analysis techniques, in order for the interpretation and comprehensive understanding of the sheer volume of experimental data produced to be accomplished. The harnessing and integration of complex genomics, proteomics, and metabolomics datasets identified by these methods facilitates the acquisition of a holistic view and provides insight into the dynamics of biological systems at the organismal level. However, living cells can be characterized as intricate networks of molecular interactions, indicating the significance of perception of how each biomolecule influences a specific phenotype. Therefore, establishing the functional roles of individual molecules of interest in a particular experimental context is often invaluable in drawing conclusions from the experimental data [37]. To this end, a widely used downstream analytical application is the Functional Enrichment analysis, which is implemented for the identification of molecules of interest in a high-throughput dataset. In particular, functional enrichment determines classes (subset) from a long list of biological elements that are over-represented in a large collection of the corresponding element, each representing a biological relevant label (e.g. Gene Ontology term, molecular pathway, protein domain, disease, etc.) [37]. Statistically enriched biological elements are then identified by comparing their frequency against a reference background list.

Importantly, the hypergeometric test or its variants including the binomial and Fisher's exact tests, are widely applicable to many existing procedures that detect enrichment, especially due to its simplicity [38]. It utilizes the hypergeometric distribution to measure the statistical significance of having drawn exactly k items, classified as success, out of n random sample draws from a population of size N comprising of K successes [39]. For instance, in a high-throughput experiment, the classical hypergeometric P-value of enrichment is the probability of randomly observing n or more differentially expressed genes annotated to the GO term and is calculated by:

$$P - value = \sum_{k=n}^{min\,(g,d)} \frac{\binom{g}{k}\binom{f-g}{d-k}}{\binom{f}{d}},$$

where $\binom{m}{n} = \frac{m!}{n!(m-n)!}$ is defined as the binomial coefficient, g is the number of genes annotated to a certain GO term, f is the total number of genes evaluated and d is the total number of differentially expressed genes detected in the particular experiment [37]. Notably, the hypergeometric distribution is modeled based on a null hypothesis which indicates that

the functional term is irrelevant to the experiment, meaning that a gene being annotated by the GO term and this gene being differentially expressed are independent events [37].

2.1 Steps of functional enrichment

Generally, different enrichment strategies can be tailored by a plethora of tools aiming to perform functional enrichment analysis. Despite the minor but existing differences of each protocol, these tools share similar basic principles, which enable the assessment of the statistical significance of the observed functional patterns when transitioning from the genome to functional level [37]. Since functional enrichment analysis is the application of enrichment analysis to data generated by “omics” techniques, ranked or unranked, raw lists of biomolecules (e.g. genes, proteins or metabolites) are used as an input [40]. As might be expected, the desired outcome usually determines the type of the list selected (ranked or unranked) as well as the tool used for the analysis.

The first step in a standard enrichment analysis method involves the calculation of the enrichment score. The careful selection/generation of the appropriate background dataset of identified molecules, against which to test for over-representation, is of utmost importance, for the elimination of any biases and the improvement of the accuracy of the enrichment score [37,41]. Once the mapping of each molecule from the gene list to ontology terms is complete, the comparison of ontology terms for the query and background list follows in order to assess the enrichment of different terms and processes [37]. It is of note that the query molecule list must be a subset of the reference/background dataset. Enriched processes/pathways or functions are considered those that are over-represented in the query dataset relative to the background dataset. The resulting enrichment score is a qualitative value, determining the statistically significant functionalities of molecules, rather than the quantification of their alteration [42].

Subsequently, the enrichment score significance is evaluated by implementing the Binomial, Fisher's exact, Hypergeometric or Chi-square tests. However, alternative mathematical approaches can be used, according to the user’s preference. More concretely, when a relatively small background dataset is available, the Fisher’s exact, Chi-square and Hypergeometric distributions are more efficient, whereas the analysis of a larger list is better accomplished with use of the Binomial probability [43]. The majority of methods are based on the p-value for the quantification of the enrichment score’s statistical significance. All terms below a predefined p-value threshold are considered to be significantly enriched, while a more strict and enforced analysis may require at least 2-fold differences [43]. Interestingly, in contrast to p-value, sample frequency and semantics of each term are valuable measures

for the biological interpretation of results and not just for the identification of enriched terms [42].

Finally, the exclusion of random events that falsely appear significant in a list of hits originated from high-throughput experiments is vital in functional enrichment analysis. For this reason, the adjustment for multiple hypothesis testing is used in functional enrichment analysis, as multiple comparisons are being made simultaneously, resulting in the performance of numerous separate hypothesis tests. Multiple testing refers to any instance that involves the simultaneous testing of more than one hypothesis [44]. This implies that the probability of false positive results is high, if individual hypotheses are in accordance with the unadjusted marginal p-values [44]. Generally, upon the performance of n hypothesis tests, the probability of at least 1 false positive can be calculated as follows:

$$P\,(making\ at\ least\ 1\ error\ in\ n\ tests) = 1 - (1 - a)^n,$$

where $P\,(making\ an\ error) = a.$ It becomes evident that as the number of hypothesis tests increases, the probability of making at least one error increases as well [45]. Herein, most of the widely utilized multiple test corrections, including the family-wise error rate (FWER) and the false discovery rate (FDR), rely on controlling type I errors. FWER provides corrected p-values by controlling the probability of even a single erroneous rejection of the null hypothesis, whereas FDR controls the ratio of false discoveries, producing q-values which indicate the ratio of accepted false discoveries, when rejecting the null hypothesis [42].

2.2 Functional enrichment tools

Representative tools used for enrichment analysis include DAVID [46], PANTHER [47], Flame [48], WebGestalt [49], aGOtool [41] and g:Profiler [50,51], each performing different statistical analysis and supporting different enrichment options. Some of these tools are briefly described in this section.

In detail, the Database for Annotation, Visualization and Integration Discovery (DAVID) is a user-friendly and web-based program, which aims to facilitate the biological interpretation of large genome-scale datasets encoded by human, mouse, rat or fly. It integrates a broad spectrum of functional genomic annotation resources in combination with graphical displays, improving the quality of high-throughput functional annotation analysis. Visualization tools are also provided and assist the comprehensive annotation of lists of genes/proteins according to functional classification, biochemical pathway maps, and conserved protein domain architectures. All data is freely downloadable in text format.

In addition, g:Profiler consists of distinct tools that are part of computational analysis pipelines. g:GOSt is used for the functional enrichment analysis of individual or multiple gene lists provided by the user. The data are imported from Gene Ontology [52], pathways from KEGG [53], Reactome [54] and WikiPathways [55], protein complexes from CORUM [56], expression data from Human Protein Atlas [41], regulatory motifs from TRANSFAC [57] and miRTarBase [58], and phenotypes from the Human Phenotype Ontology [59]. g:Convert maps gene and protein identifiers between numerous namespaces and converts them based on information from the Ensemble database. g:Orth is a tool for mapping orthologous genes across multiple species based also in Ensemble data. Finally, g:SNPense maps human SNP identifiers to gene names, chromosomal locations and variant consequence terms from Sequence Ontology.

Similarly, aGOtool is an open-source, publicly accessible web tool for Gene Ontology (GO) enrichment analysis. It focuses on proteins and identifies terms from the UniProt keyword classification system [60], Kegg and Wiki pathways, Reactome, protein families and domains from Pfam [61] and InterPro [62], as well as human diseases and tissues from the DISEASES database [63] and Brenda tissue database [64] respectively.

CHAPTER 3: Fundamental concepts in Graph theory

3.1 Algebraic graph theory elements

Since the onset of graph theory in recreational math problems [65], it has been proven a really powerful and useful tool in numerous and seemingly diverse fields of science, ranging from mathematics, engineering and computer science to biosciences, chemistry, sociology and linguistics [66,67]. It is considered one of the main subjects of study in the domain of discrete mathematics, especially due to its interesting mathematical properties, and is used to depict the relations between objects or entities [67].

Conceptually, a graph or network is the pictorial representation of a set of vertices (points or nodes), connected by edges (arcs or lines) [67]. However, formally, a graph G can be defined as a pair of sets (V(G), E(G)), consisting of a set V(G) of vertices and a set E(G), disjoint from V(G), of edges, together with an incidence function Ψ_G that associates with each edge of G an unordered pair of (not necessarily distinct) vertices of G [68]. Even though there are multiple different ways for one graph to be depicted, two different graphs, whose sets of edges and vertices are of the same number and their connectivity is retained, are called isomorphic [67]. A subgraph G' = (V' , E') of the graph G = (V, E) is a graph where V' is a subset of V and E' a subset of E [69]. A finite graph G is a graph whose V(G) and E(G) are finite sets [70]. It is of note that the existence of loops (an edge which is drawn from a vertex to itself) and multiple edges is possible in any type of graph, apart from simple graphs. When any two vertices are joined by more than one edge, the graph is called a multigraph (multi-edge graph) [70]. Examples are presented in *Figure 4*.

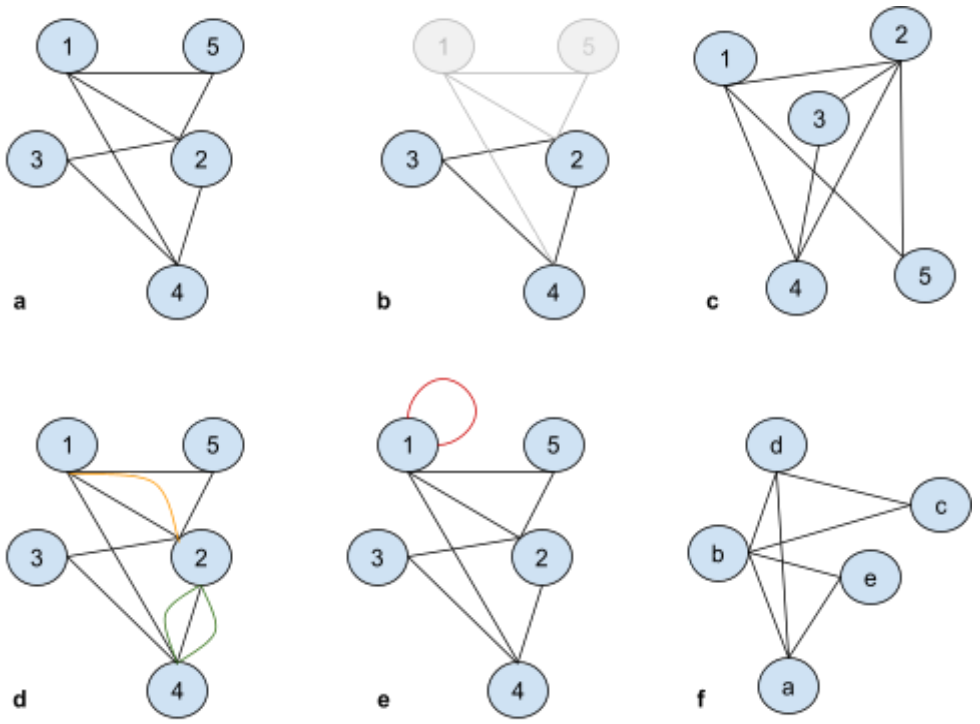


Figure 4: Simple graph and multigraph. (a) A graphical representation of a finite, simple graph *G = (V, E)*. **(b)** A graphical representation of a subgraph *G' = (V', E')*. **(c)** An alternative representation of the graph *G*. **(d)** A multigraph *G" = (V", E")*. **(e)** A simple graph *G* with a loop. **(f)** Graph *H = (V"", E"")* is isomorphic to graph *G = (V, E)*.

Besides simple and multi-edge graphs there are plenty of other graph types, including directed, undirected, connected, disconnected, weighted, bipartite and trees. A graph is considered directed (or digraph) if each edge bears an arrow mark, indicating its direction, while undirected (or non-directed) is the graph which has a single connection defined as $E = \{(i, j)|, i, j \in V\}$ between vertices i and j [69]. It is important to note that in a directed graph an edge is defined by an ordered pair of nodes, distinguishing it from an undirected graph. If between every pair of vertices in a graph exists a path, this graph is called connected and if two or more vertices are not connected, the graph is called disconnected [67]. A weighted graph is defined as a graph where E is a set of edges between the vertices i and j ($E = \{(i, j) \mid i, j \in V\}$) associated with a weight function $w: E \rightarrow R$, where R denotes the set of all real numbers [69]. A bipartite graph is a simple, undirected graph G = (V, E) with vertex partition into two disjoint sets V1 and V2 that $(i, j) \in E$ implies either ($i \in V'$ and $j \in V''$) or ($j \in V'$ and $i \in V''$). In general, in a bipartite graph an edge should connect any vertex in set V1 to any vertex in set V2, but not the vertices that belong to the same set [69]. Finally, a tree comprises a connected, undirected graph with no cycles [66]. Examples of the aforementioned graph types are shown in *Figure 5*.

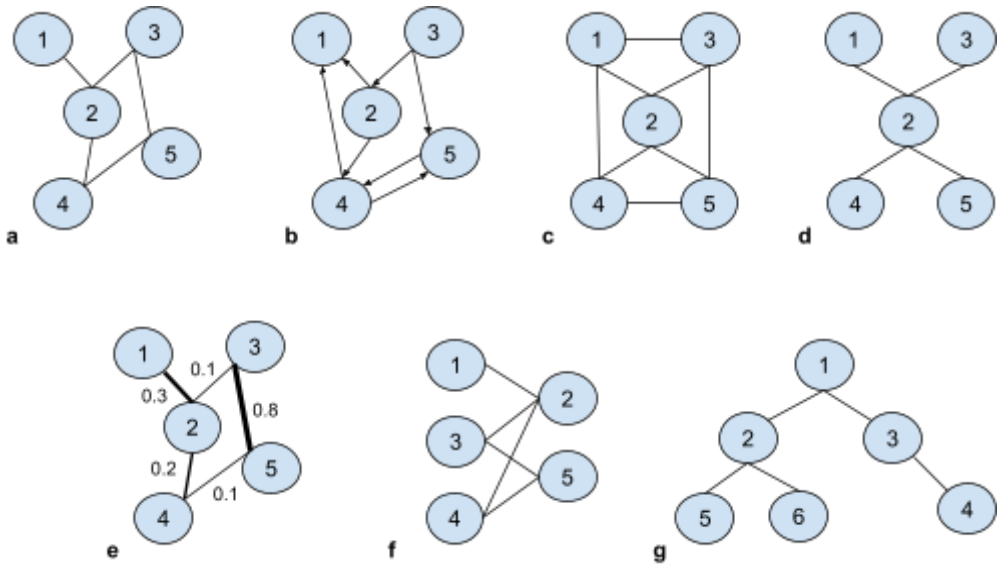


Figure 5: Different types of networks. (a) Undirected graph G_1= *(V_1,E_1)*. **(b)** Directed graph G_2= *(V_2,E_2)*. **(c)** Connected graph *H = (V, E).* ***(d)*** *Disconnected graph H' = (V', E').* **(e)** Weighted graph G_3= *(V_3,E_3)*. **(f)** Bipartite graph *G = ({V, V'}, E).* **(g)** A tree graph G_4= *(V_4,E_4)*.

3.2 Basic network properties

Graphs are characterized by various properties, depending on their structures. These network properties, and particularly topological features, are important to unravel the data included in a graph and analyse its complexity by extracting information from individual components. The topology of a network refers to the arrangement of nodes and edges and is applicable both to the whole network and to individual nodes and edges. Some of the most used topological properties and concepts are the *degree, degree distribution, density, clustering coefficient, distance* [69].

3.2.1 Degree

Firstly, the ***degree*** *deg(v)* (*Figure 6a*) of an undirected graph is defined as the number of vertices adjacent to a vertex V [69] and is a fundamental characteristic which influences other parameters. In a simple graph *G* with *n* number of vertices, the degree of any vertex is *deg(v) = n – 1* $\forall$ *v* $\in$ *G*, because as mentioned before there are no loops in simple graphs, thus the degree of vertex will be up to the number of vertices in the graph (n) minus 1, itself [66]. In case of a directed graph, the degree of a node can not be easily described by a single number, as the underlying information of an incoming and an outgoing edge may be of great importance and should be taken into consideration. Instead, by the calculation of all the

edges incident from a vertex and the edges incident to a vertex, one can obtain two numbers for the degree of a node, the ***indegree*** $deg(V)^{in}$ and the ***outdegree*** $deg(V)^{out}$ respectively, a terminology that reflects the direction of the edges. The sum of the indegree $deg(V)^{in}$ and the *outdegree* $deg(V)^{out}$ is defined as the total degree of the directed graph [69]. The ***average degree*** of a graph *G = (V, E)* measures the number of edges that are included in set *E* compared to the number of vertices in set *V*. Each edge is incident to two vertices and counts in the degree of both vertices, thus the average degree of an undirected graph is $2 * \frac{|E|}{|V|}$ [69].

3.2.1 Degree distribution

An arguably fundamental quantity associated with the network structure is the ***degree distribution*** $P_{deg}(k)$, a parameter which aims to capture the variance in the degree of nodes in a graph, while providing useful information about the structure of the network and how centralized or distributed it is [69]. It is a quantitative parameter, but its alteration can have a rather qualitative effect on the network. The degree distribution can be calculated by simply counting the number of randomly selected nodes with a degree equal to *k* and can be defined as the fraction of nodes in the graph with a degree *k*. In other words, it gives the observed frequency of a node of degree *k* [71]. Notably, a spectrum of network's degree distributions can be defined starting from homogeneous to heterogeneous. A degree distribution is homogeneous (*Figure 6b*), if all nodes have relatively similar degrees, like in Poisson and Gaussian distributions, whereas if there is great disparity between the node's degrees, the distribution is heterogeneous (*Figure 6c*). The latter can be described as a scale-free network if the distribution of the vertex degree (*k*) follows a power-law distribution of the form $P(k)k^{-\gamma}$ [72]. One of the main properties of this type of network is the formation of highly connected nodes, named ***hubs***, while other poorly connected nodes are also linked to these hubs [72].

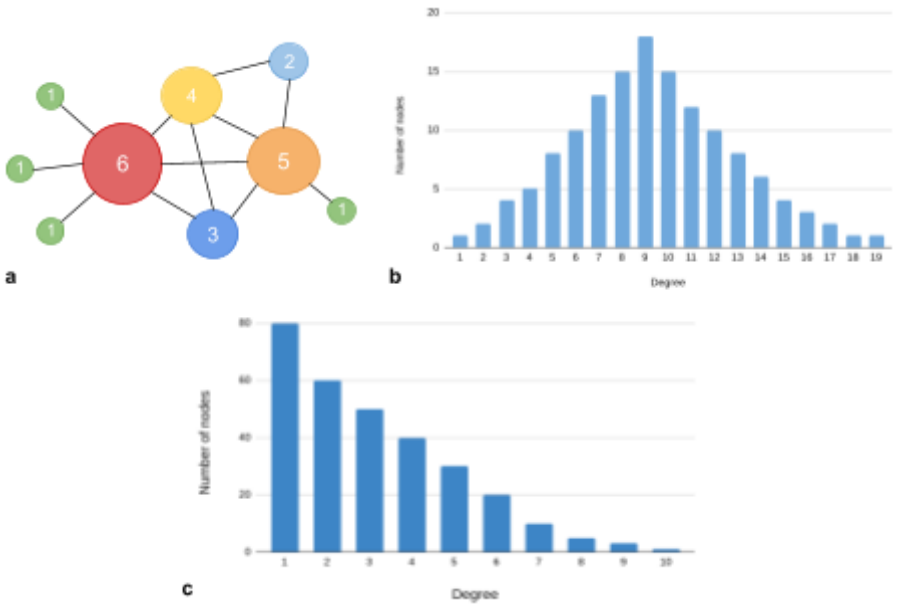


Figure 6: Degree and degree distribution. (a) Simple, undirected graph *G = (V, E)*, |V| = 9, |E| = 12. Each node's size has been adjusted according to its degree. **(b)** Column chart which depicts a homogeneous degree distribution of a graph (Poisson distribution). **(c)** Column chart which depicts a heterogeneous degree distribution of a graph.

3.2.1 Density

Another significant property of graphs is the ***density***, which is described as the ratio between the number of edges in a graph and the number of possible edges in the same graph. In particular, the maximum number of edges of a simple, undirected graph is |V| * $\frac{(|V|-1)}{2}$, implying that the density is *D* = 2 * $\frac{|E|}{(|V| * (|V|-1))}$, while for a simple, directed graph the density is *D* = $\frac{|E|}{(|V| * (|V|-1))}$, as this type of graph can have at most |V| * (|V| - 1) number of edges [69]. A graph is considered ***complete*** (*Figure 7)* when there's an edge between any two nodes, ***dense*** (*Figure 7)* if the number of edges is close to the maximal number of edges ($|E| \simeq |V|^k$, 2 > k > 1), whereas a graph with fewer edges ($|E| \simeq |V|$ or $|E| \simeq |V|^k$, k ≤ 1), it is considered as ***sparse*** (*Figure 7)* [69].

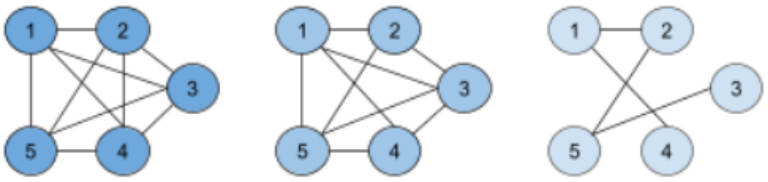


Figure 7: Density and Clustering coefficient. Representation of a simple, undirected complete graph (left), dense graph (middle) and sparse graph (right).

CHAPTER 4: Bioentity interaction databases

Traditionally, biological research has been dominated by an explanatory and methodological reductionism approach, providing a wealth of building explanations based on the dissection of biological systems into their constituent parts [73]. Even though the research at deep levels of organisation can potentially reveal hidden mechanisms, reductive research strategies neglect their context, limiting biological insights [74]. Therefore, considering that the specificity of even a discrete biological function is rarely attributed to individual molecules, biases and distortions of reduction method arise [73,74]. Instead, integrative strategies attempt to understand and further explain the structure and dynamics of the complex intercellular and intracellular systems at different levels, utilizing findings from multiple scientific fields [74].

This challenge of contemporary biology to embark on a systems-based approach is significantly supported by the technological advancements of high-throughput techniques of the last decade. For instance, microarray and RNAseq technologies provide insights about gene expression, while scRNAseq technology organizes cells into groups based on their gene expression profiles. As far as the proteins are concerned, mass spectrometry identifies proteins based on their molecular weights and mass-to-charge ratio, whereas Nuclear Magnetic Resonance (NMR) and X-Ray crystallography are used for the determination of 3D protein structures in space. In addition, whole genome and whole transcriptome analysis along with metabolomics are used to study small molecules and metabolites within cells, biofluids, tissues or organisms [75]. By harnessing these methods, researchers are enabled to catalogue vast numbers of information in the form of component molecules of biological networks at a genome-wide scale and under a large number of different experimental conditions [76].

As modern biomedical research evolves to address the inherent complexity of biological systems by combining multiple -omics approaches (e.g., genomics, proteomics, transcriptomics, metabolomics), specialized tools and repositories become a necessity in order for this complexity to be revealed, represented and interpreted. One advanced integration and visualization technique involves the use of biological interaction networks. In particular, Network Biology often attempts to illustrate a holistic picture of the interdependent relationships between biological entities and processes via the implementation of graph theory methods, statistics, mathematical modeling and visualization tools [77]. Graphs are mainly the means to model and portray compartments of whole systems and their biomolecular interactions. Commonly, a node represents a biomolecule (e.g., gene, protein, chemical, compound, disease) whereas an edge the relationship between them (e.g., co-expression, co-occurrence, sequence similarity, coevolution, orthology, homology, fusion, common function) [78]. Inarguably, biological interaction networks have been proven as an

invaluable tool in a wide range of diverse fields of biological analyses, including the organization of trophic webs, protein interactions, brain circuits and gene regulation. However, the most characteristic example is the Human Interactome Network [79], a proteome-scale analysis of protein-protein interactions for the entire human proteome, that has allowed the detection of previously unknown functional relationships and it is currently a reference map for the human proteome and its interactions.

Apparently, the increasing availability of interaction data originated from high throughput methods or generated through computational predictions often provide sufficient knowledge to acquire a dimensional view of many potential functional activities [80]. Numerous biomedical repositories, namely Pubmed [81], UniProt [60], GenBank [82] or Ensembl [83], store such evidence, offering organized datasets to be further investigated. Nonetheless, the intrinsic interrelatedness of biological phenomena as well as the successful generation and analysis of interaction networks indicates the necessity of a three-dimensional view and the majority of the aforementioned databases are not dedicated to the analysis of interactions. Therefore, biological interaction databases have emerged, appearing as specialized repositories for providing evidence on gene, protein, and small molecule interactions, as well as associations of these interactions with metabolic pathways, host-pathogen relationships, diseases, and even ecological data.

Typically, the type of interactions, the source of information and the data curation procedure define the classification of bioentity interaction databases. Particularly, the identity of a database primarily depends on the interaction type provided. For example, the physical or/and functional protein-protein or protein-small molecules interactions determine the protein interaction databases, whereas gene co-expression databases describe interactions based on similar expression patterns. According to the source of information and the data-acquisition policy, interaction databases can be further divided into three main groups: *i)* primary, *ii)* secondary and *iii)* predictive. More specifically, as primary are characterized the databases that independently compile the evidence collected from multiple primary sources (i.e., scientific publications or or from deposited interaction datasets, such as those derived from high-throughput experiments). Secondary or meta-databases however combine and annotate data curated by several primary databases in a single repository, rather than collect information directly from primary sources. Finally, predictive databases contain both experimentally verified and computationally predicted interaction evidence derived from various methods, such as sequence or structure analysis, or from automatic methods for parsing the literature (e.g., text mining). The categorization of databases is also specified by their data curation policy which involves the annotation, publication and presentation of integrated data originating from various sources. Data acquisition can be manual (i.e., handled by curators, or by the scientific community) or automated which is performed using

computational methods. Sometimes, high-throughput automation complements manual curation, resulting in a combination of the two previously mentioned methods [84]. In addition to the type of curation, level can also indicate two distinct categories of databases which include the lightly and the deeply curated. Lightly curated databases aim to publish the maximum amount of interaction information obtained from computational methods, without necessarily focusing on their detailed aspects. Therefore, errors, redundancy and overlapping information are common shortcomings of lightly curated databases. Instead, deeply curated databases offer more detailed information which is periodically manually annotated, validated through multiple sources and checked for redundancy. To date, various types of biological interactions have been characterized and provided by numerous different databases, some of which are discussed below.

4.1 Gene co-expression databases

Gene co-expression networks are transcript - transcript association networks, generally reported as undirected graphs, where each node corresponds to a gene and a pair of nodes is connected with an edge, if a significant co-expression association exists between them. Notably, functionally related genes generally share similar expression patterns on spatio-temporal states or environmental conditions. These genes are of great biological interest as they may participate in the same biological pathway or due to the fact that their regulation depends most probably on the same transcriptional regulatory program. Thus, they provide powerful information to estimate the functions of uncharacterized genes [85]. The modules or/and the highly connected subgraphs presented in a gene co-expression network depict the gene groups, which have analogous function or mediate the same biological processes, thus resulting in numerous interactions among themselves.

Gene co-expression networks are usually constructed using gene expression data from high-throughput gene expression profiling technologies (e.g. microarrays or more recently RNA-seq). The calculation of co-expression values and the selection of an appropriate significance threshold are prerequisites for building this type of network. Normally, the co-expression similarity score is calculated with the use of metrics like Pearson or Spearman. In this paragraph are mentioned several examples of co-expression network databases as well as information describing gene-gene relationships across various organisms.

COXPRESdb [86] (*Figure 8a)* is a database that deals with condition-independent co-expression data of protein-coding RNAs retrieved from 11 different model organisms. In order to improve the reliability of information and remove the biased relationships, COXPRESdb compares multiple coexpression data derived by different transcriptomics technologies and from various species. Specifically, the last update combines gene expression data from 23 different co-expression platforms, out of which, 123 experiments concern Human, 154 Mouse and 154 Rat, released by Gene Expression Omnibus (GEO). In total, COXPRESdb hosts 12 co-expression networks for various species created from approximately 157,000 microarray and 10,000 RNA-seq samples.

GeneMANIA [87] (*Figure 8b)* is an easy-to-use repository which identifies functionally similar genes to a set of input genes, utilizing a wealth of genomics and proteomics data originated from various sources, including GEO, the Biological General Repository for Interaction Datasets (BioGRID) [88], IRefIndex and Interologous Interaction Database (I2D) . In the current version, GeneMANIA has 2,300 networks consisting of approximately 600 million interactions between almost 164,000 genes of 9 different organisms (A. thaliana, C. elegans,

D. rerio, D. melanogaster, E. coli, H. sapiens, M. musculus, R. norvegicus and S. cerevisiae) are supported.

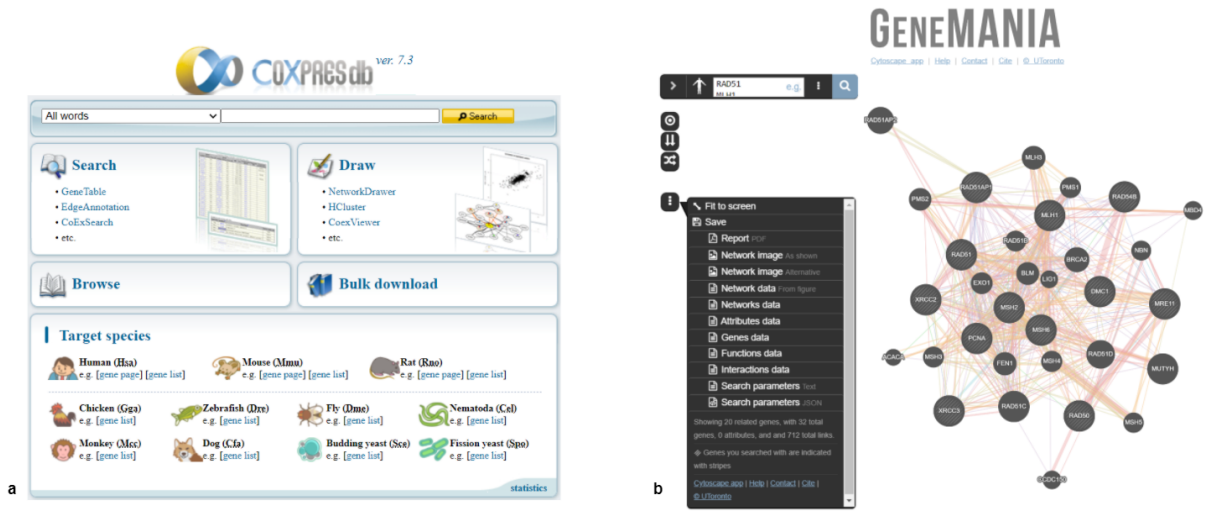


Figure 8: COXPRESdb and GeneMANIA databases. (a) User Interface and services of COXPRESdb. **(b)** Gene co-expression network example and visualization parameters of GeneMANIA.

Moreover, GeneFriends [89], Immuno-Navigator [90] and COEXPEDIA [91] are co-expression and gene expression databases for *H. sapiens* and *M. musculus*. In particular, the latest version of GeneFriends integrates updated gene and transcript networks based on RNA-seq data from 46,475 human and 34,322 mouse samples, whereas the Immuno-Navigator tool offers cell-type specific gene expression and correlation of expression data in cells of the immune system. Currently, it contains data from 4639 human samples, obtained from 19 cell types from 191 studies, as well as 3434 mouse samples, obtained from 24 cell types from 261 studies. In contrast to the databases mentioned above, COEXPEDIA provides 8 million co-functional co-expressions data, resulting from statistical assessment and derived from 384 and 248 GEO individual studies associated with biomedical information.

In addition to the biomedical-driven databases, various repositories are dedicated to gene co-expression networks for plant species, such as ATTED-II [92] (*Figure 9a)*, CoP [93] (*Figure 9a)* and PlaNet [94] (*Figure 9a)*, while the Arabidopsis Co-expression Tool (ACT) [95] and AraNet are *A. Thaliana*-specific. Currently, ATTED-II focuses on co-regulated gene relationships for nine plant species, supported by microarrays and RNA sequencing (RNAseq)-based co-expression data. Similarly, CoP provides condition-independent co-expression data associated with biological processes assembled from microarray datasets of 8 different plant species (*Figure 9b)*. PlaNet is an online platform that offers various tools enabling the visualization and analysis of co-expression networks for

photosynthetic organisms (*Figure 9b)*, while ACT and AraNet incorporate co-expressions between 21,273 *A. Thaliana* genes from microarrays and genome-scale functional networks.

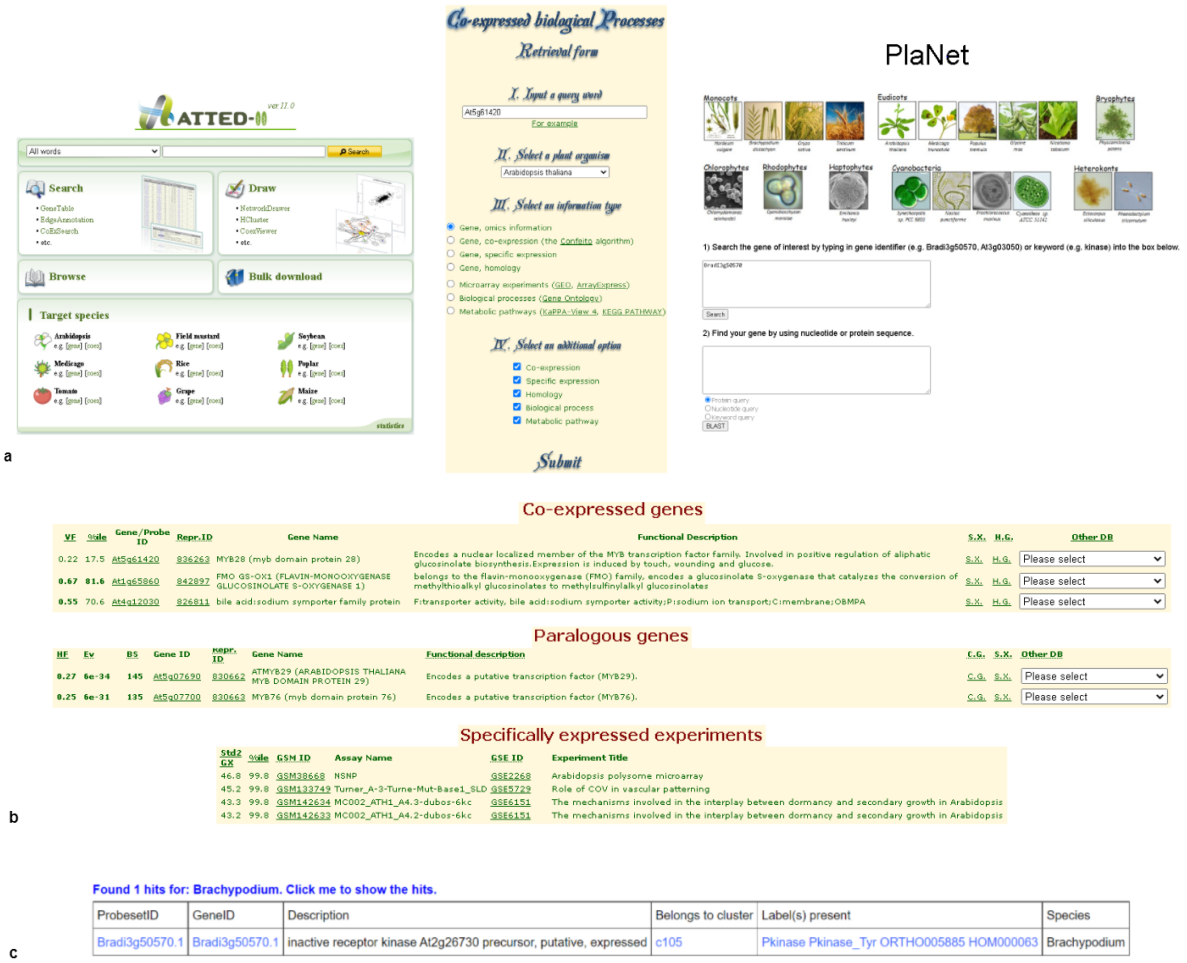


Figure 9: Gene co-expression databases for plant species. (a) User Interface and search options of ATTED II, CoP and PlaNet databases. **(b)** Search results based on an input query in the CoP database. **(c)** Search results based on a gene identifier input in the PlaNet database.

4.2 RNA interaction databases

RNA molecules are indispensable cellular components and participate in almost every essential biological process, such as transcriptional and post-transcriptional regulation, storage and flow of information and signal transduction through environmental sensing [96]. However, their functions and regulation are crucially dependent on the specificity and the efficacy of their intermolecular and intramolecular interactions. Diverse RNA types have the ability to interact with other RNAs, DNA, proteins, lipids, and metabolites, forming complex molecular networks [96]. Modification of these interactions is closely related to multiple different disease phenotypes. Therefore, the description of RNA interactions networks helps the unraveling of the underlying mechanisms of RNA functions which are mediated by the various different interactions.

4.2.1 RNA-protein interactions

The inherent instability of RNA molecules coupled with the diversity and versatility of their functions are largely responsible for their constant chaperoning by a plethora of different protein complexes. Besides the regulatory binding of proteins to RNA molecules, RNAs also interact with specific proteins to accomplish functions, reflecting both a protein-centric and an RNA-centric approach, respectively [97]. Notably, despite the significant contribution of the recently developed transcriptome-wide methods and integrative analyses, deciphering the intricate principles of these networks is undoubtedly challenging.

In order to facilitate the understanding of the complex, yet vital interactions for many biological processes, RNA-protein interaction databases integrate experimentally validated and computationally predicted data from published literature and high-throughput technologies, visualizing the RNA interactome in reference to the collected information [98]. Regarding the contents provided by each resource, RNA-protein interaction databases could be characterized either as comprehensive, incorporating data from multiple sources, specialized, emphasizing mainly on a category of interactions validated by various experimental methods or predictive, utilizing computational methods, apart from experimental data, to predict possible interactions.

Firstly, Protein–RNA interaction database (PRD) [99] is a comprehensive database which integrates literature-based physical RNA-protein interactions at the gene level. The current version of PRD contains 10,817 interactions among proteins and protein-coding RNAs, tRNAs, rRNAs, miRNAs and viral RNAs in 22 organisms, corresponding to 1539 unique gene pairs. Each interaction is enriched with further information curated from multiple other resources, concerning RNA and protein binding sites/motifs, Gene Ontology (GO) terms [100], detected methods and biological functions.

The RNA Interactome Database (RNAInter) [101], previously named RAID, is another comprehensive and manually curated database of RNA-associated interactions (RNA–Protein/RNA-RNA), integrating experimentally validated and computationally predicted data from published literature and 35 other resources. Apart from the fuzzy/batch search (*Figure 10)*, interaction network and RNA dynamic expression that are included in RNAInter, four RNA interactome tools are also embedded, namely, RIscoper [102], IntaRNA , PRIdictor [103] and DeepBind [104]. Currently, RNAInter contains 41,322,577 RNA-associated interactions of 22 different RNA types in 154 species, including 34,106,998 RPIs. Identifiers for external databases such as miRBase, NCBI, HGNC, Ensembl, Online Mendelian Inheritance in Man (OMIM) [105], Human Protein Reference Database (HPRD) [106] and UniProtKB are also provided. Data can be browsed by interaction type, detection method or species and are downloadable in text format, as well as obtainable through an API.

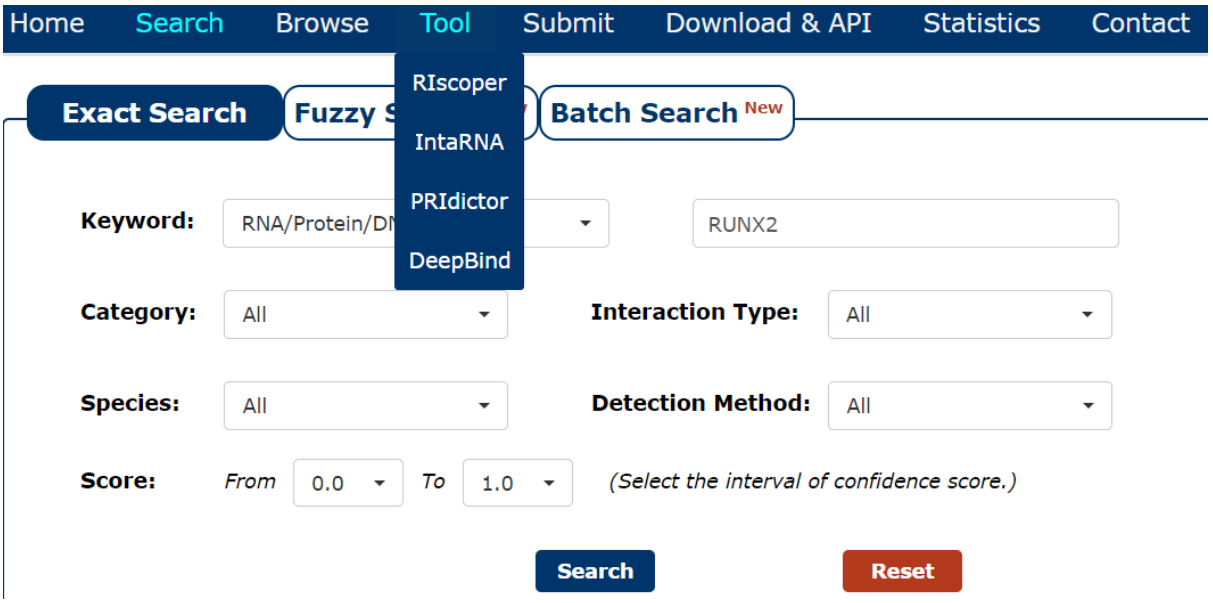


Figure 10: RNAInter database search options. Fuzzy/batch search as well as four RNA interactome tools available in RNAInter, namely, RIscoper, IntaRNA, PRIdictor and DeepBind.

Furthermore, POSTAR2 [107] and doRiNA [108] constitute more specialized repositories, concerning post-translational regulatory RNA–Protein interactions. Both databases provide functional association prediction and contain structural information about binding sites of RNA-binding proteins and RNAs originating from cutting-edge high-throughput sequencing techniques. In particular, POSTAR2 provides the largest collection of RNA-binding proteins (RBP) binding sites and functional annotations in 6 species, including human, mouse, fly, worm, A. thaliana and yeast. Three modules (RBP, RNA and translatome modules) and RBP–RNA interaction network in H. sapiens are supported, offering both functional and structural insights into translational and post-translational regulation. On the other hand,

doRiNA integrates experimentally validated RBPs and miRNA target sites data for H. sapiens, M. musculus and C. elegans, while computational methods for all species are also used for miRNA target sites prediction.

As far as predictive databases are concerned, Protein–RNA Interface Database (PRIDB) [109] contains a total of 30,056 RNA-Protein interactions (5694 unique RNA chains and 1702 unique protein chains) and incorporates structural information facilitating the analysis of RNA-protein complexes and their interface, by providing a user-friendly format. The RNA-Binding Protein DataBase (RBPDB) [110] is a manually curated resource of experimentally observed RNA-binding data for 1171 RBPs in humans, mice, flies and worms. Finally, RNA binding site DataBase (RsiteDB) [111] is another predictive database aiming to describe, classify and predict interactions between protein binding sites and single-stranded RNA bases.

4.2.2 RNA-DNA interactions

The recent advancement in technology and the development of high-throughput techniques has significantly altered the analysing methods of RNA structures and interactions. The combination of biochemical reactions and transcriptome-wide analysis has enabled the studying of not just one but multiple RNA interactions, including RNA-DNA [112]. The formation of RNA-DNA hybrids as well as RNA-DNA interactions possess key roles in diverse biological processes through both genetic and epigenetic regulations, such as dosage compensation, imprinting, development process and disease progression [96]. Therefore, in this paragraph, databases focusing on RNA interactions with DNA molecules are presented.

NPInter [113] (*Figure 11a*) is one example of a manually curated database which includes experimentally verified functional interactions between various types of ncRNAs and biomolecules such as proteins, RNAs and DNAs. Currently, NPInter incorporates a total of 1,100,658 interactions among 35 different organisms based on interaction data from the RISE [114]. NcRNA entries are annotated against NONCODE [115], miRBase [116] and circBase [117], while proteins from UniProt, Ensembl and RefSeq [118]. Additional metadata concerning the interaction class and the tissue/cell line of the experiment complement the interaction information provided.

Tarbase [119] (*Figure 11b*) is another manually curated database of experimentally supported miRNA-DNA interactions which gathers data primarily from the literature as well as from raw libraries like GEO and the DNA Data Bank of Japan (DDBJ) [120]. The latest version contains more than 1 million entries, corresponding to approximately 670,000 miRNA-target pairs. derived from more than 33 high-throughput techniques, applied to 516 cell types and 85 tissues, under 451 experimental conditions, across 18 species. Tarbase

also supports identifiers from Ensembl and miRBase and is interconnected with other DIANA-tools, like microT-CDS [121] for in silico identification of miRNA targets, LncBase v2.0 [122] for miRNA–lncRNA interactions identification and DIANA-miRPath v3.0 [123] for miRNA functional characterization.

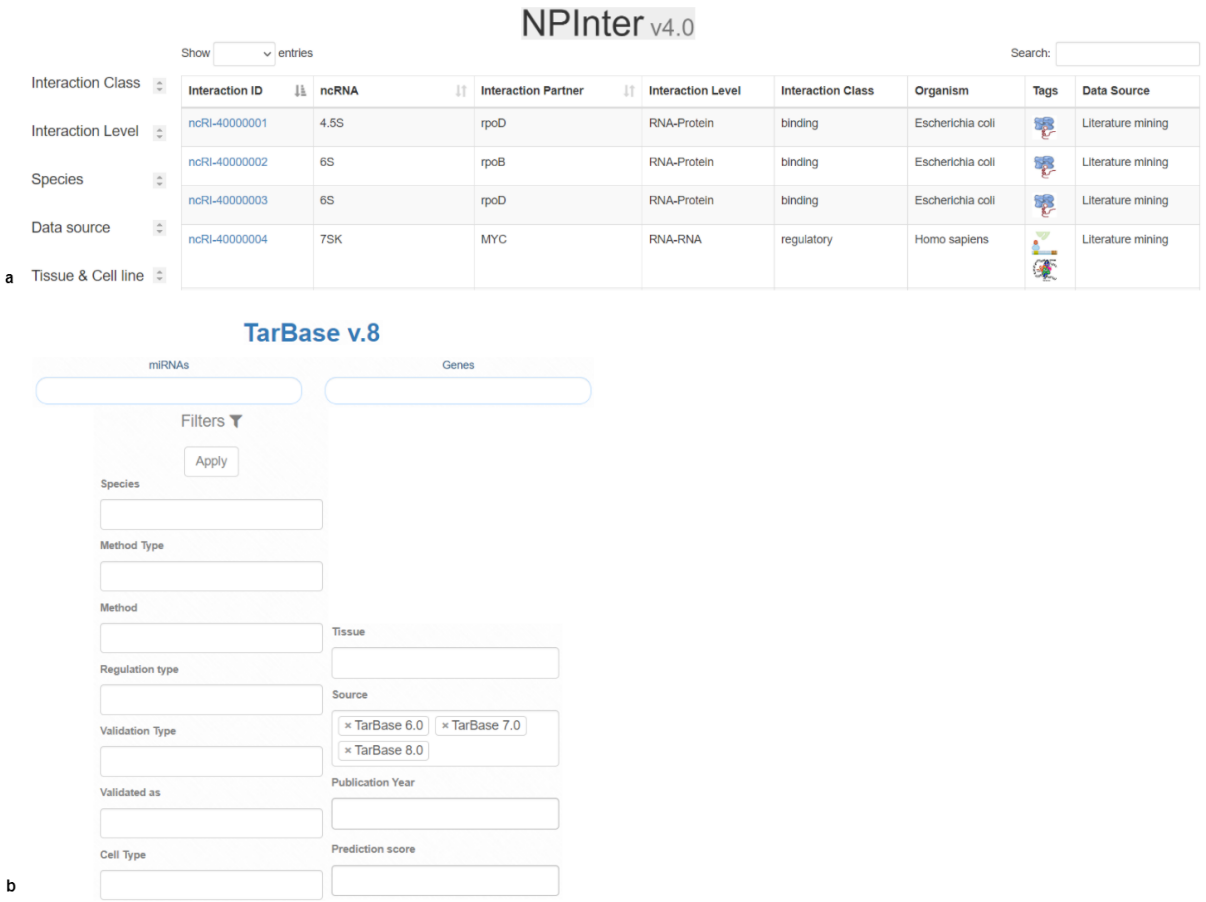


Figure 11: NPInter and TarBase v.8 database. (a) Browse interface of NPInter database. It provides various search options regarding the interaction class and level, species, data source, tissue and cell line. **(b)** Searching filters of TarBase v.8 database.

In addition to databases that incorporate interactions verified from high-throughput pipelines, data of EVLncRNAs [124] and LncRNA2Target [125] are validated mainly by low-throughput experiments. Particularly, EVLncRNAs is manually curated from the literature and is dedicated to lncRNA interactions with biomolecules such as DNA, RNA, proteins and TFs, supplemented by entries originated from other repositories (e.g., LncRNADisease [126] and Lnc2Cancer [127]). Its current version (v2.0, July 2020) covers 4,010 total lncRNAs and 6244 biomolecular interactions across 124 species, and 11,257 lncRNA-disease associations across 10,82 diseases. A network visualization of all available interactions is offered, as well as links to tools for lncRNA prediction. Similarly, LncRNA2Target is a comprehensive manually curated resource over 152 thousands of lncRNA-target gene

among H.sapiens and M.musculus associations inferred from lncRNA knockdown or overexpression experiments followed by high-throughput microarray/RNA-seq. All lncRNAs were annotated by NCBI Genbank, Ensembl, GENCODE [128] and Entrez ID/symbols and gene targets by Entrez ID/symbols [129].

Finally, as far as the plant-based resources are concerned, the Plant Non-coding RNA Database (PNRD) [130], which is an updated version of PMRD (plant microRNA database) [131], incorporates plant-related ncRNAs and is currently composed by 25,739 entries, from 11 different ncRNA types across 150 plant species. Notably, PNRD focuses on miRNA-target relationships, providing 178,138 target pairs across 46 plant species. The target information is enriched through psRNATarget [132] and the literature and concerns protein-coding genes, literature ncRNAs and NONCODE lncRNAs. Importantly, PNRD hosts a Cytoscape service for constructing miRNA-gene regulatory networks.

4.2.3 LncRNA-disease interactions

Undoubtedly, the interactions of RNA with other biomolecules unveils a plethora of regulatory functions that can be fulfilled, ranging from coding of proteins to catalysis. Consequently, as it becomes evident from the previous paragraphs, it is not surprising that multiple databases are dedicated to these relationships which are of great importance for cellular function and development. However, the normal function of cells depends largely on the accurate expression of both protein coding and non-coding RNAs. Especially, Long non-coding RNAs (lncRNAs) are transcripts that are longer than 200 nucleotides in length and significantly help to decipher the underlying mechanisms related to the pathogenesis of various diseases. Databases integrating information about LncRNA interactions with diseases are discussed in the paragraph below.

Firstly, LncRNADisease 2.0 [126] is a user-friendly, manually curated database which documents over 200,000 lncRNA-disease and circular RNA-disease associations across 4 different species, either experimentally or computationally validated. The computational prediction is based on LRLSLDA [133], LDAP [134], RWRlncD [135] and LncDisease [136], while the experimentally supported data is divided into strong and weak evidence. Importantly, each lncRNA-disease association entry contains detailed information, including gene symbol, gene category, disease information, regulatory relationship and PubMed information accompanied by a confidence score and each disease name is mapped to Disease Ontology (DO) [137] and Medical Subject Headings (MeSH) [138].

Another comprehensive repository of associations between lncRNAs and circular RNAs with diseases and specifically with various human cancer subtypes, is Lnc2Cancer [127]. This database comprises a collection of 10,303 experimentally supported interaction

evidence between 2,659 human lncRNAs, 743 circRNAs and 216 human cancer subtypes. All information is manually curated from the literature and additional metadata concerning regulatory mechanisms (miRNA, TF, genetic variant, methylation and enhancer), biological functions (cell growth, apoptosis, autophagy, EMT, immunity and coding ability) and clinical applications (metastasis, recurrence, circulation, drug-resistance, and prognosis) are provided.

Moreover, lncRNASNP2 [139] and LincSNP [140] are two SNP-centric databases, correlating diseases with information on functional SNPs and mutations in lncRNAs. To begin with, lncRNASNP2 (v2) is a comprehensive collection of SNPs in human and mouse lncRNAs, as well as their impact on lncRNA structure and function. 10,205,295 SNPs on 141,353 H. sapiens lncRNA transcripts and 5,104,701 SNPs on 117,405 M. musculus lncRNA transcripts are currently provided by lncRNASNP2, which are retrieved from 170,002 NONCODE lncRNA genes. In addition, it contains noncoding variants from COSMIC [141] cancer data as well as TCGA cancer mutations, whereas lncRNA-miRNA interactions and lncRNA-disease associations were predicted based on data originated from miRBase and the Human microRNA Disease Database (HMDD) [142] accordingly. Similarly, LincSNP stores and annotates experimentally supported disease or phenotype-associated variants, including SNPs, linkage disequilibrium SNPs (LD SNPs), somatic mutations and RNA editing in human lncRNAs and circRNAs or their regulatory elements.

4.3 Protein interactions databases

Protein molecules participate and often orchestrate a myriad of cellular activities, maintaining health or causing a number of pathological conditions. Herein, the investigation of these relationships is vital for the discovery of biological processes and next-generation therapeutics as well as the study of disease mechanisms. Protein interaction databases constitute an important source of knowledge of the intrinsic physical and functional protein associations with other proteins or with chemical compounds, such as ligands, drugs, and others. For this reason, a great number of the specific type of databases have emerged in the literature and a subset of them is described in the following paragraph.

4.3.1 Protein-protein interactions

Protein-Protein interactions are a key feature of the biological organization in all organisms as they are involved in the vast majority of cellular functions, mediating both physiological and pathological processes. These interactions are becoming one of the main objectives of system biology, assisting the understanding of the underlying functional relationships between proteins and the elucidation of various mechanisms. Therefore, their characterization has drawn attention in recent years, developing various experimental and predicting methods. As a result, a significant number of databases have emerged in order to catalog and annotate these interactions.

To this end, IntAct [143] (*Figure 12*) and MINT (Molecular INTeraction) [144] are manually curated resources that focus on experimental evidence derived from peer-reviewed publications. Specifically, IntAct constitutes one of the largest biomolecular interaction databases, offering over 11 million binary interactions, the majority of which refer to protein-protein complexes. It is a major participant in the International Molecular Exchange (IMEx) Consortium, a combined effort to provide an integrative, non-redundant dataset of biomolecular interactions [145]. The data provided are enriched by the integration of additional experimental evidence deposited to IntAct by curators, including MINT, UniProtKB/Swiss-Prot and PDB. Similarly to IntAct, MINT currently provides 4568 physical (direct) and functional (indirect) interactions evidence, accompanied by further information on promoter regions, mRNA transcripts and the functional annotation of its protein partners. In addition, each association can be displayed graphically via the MINT Viewer, while a numerical score (IntAct Mi-score) is applied for the evaluation of data[88] confidence.

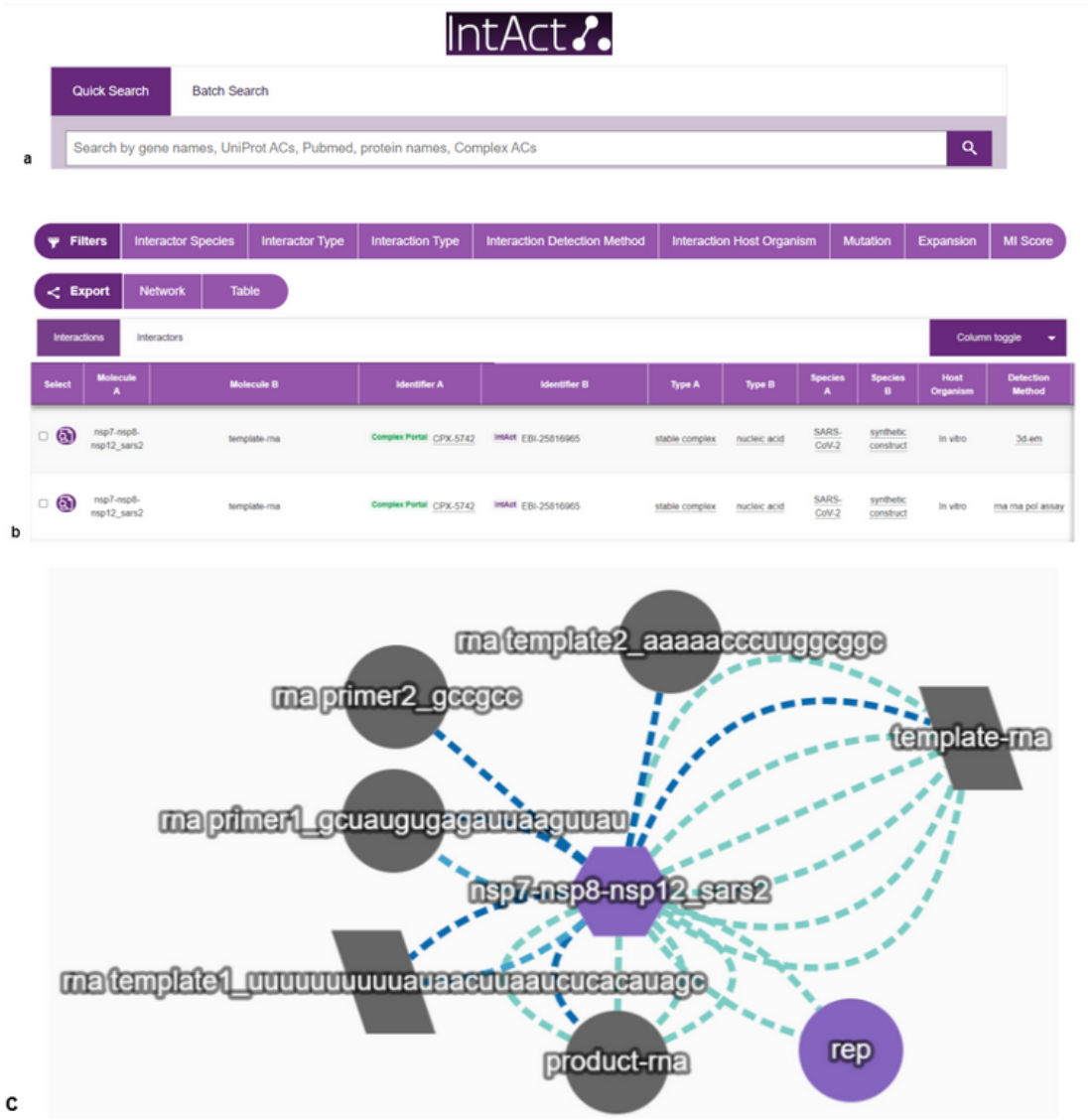


Figure 12: User interface of IntAct database. (a) Quick and Batch search of IntAct database. The user can query by gene names, Uniprot ACs, Pubmed, protein names, Complex ACs. **(b)** Search results depicted in table format. Filtering options enabling different views are provided as well. (c) Interaction network visualization of Complex ACs: CPX-5742.

In contrast to the MINT and IntAct, the Database of Interacting Proteins (DIP) [146] and the Integrated Interactions Database (IID) [147] are comprehensive repositories which incorporate mainly experimentally validated protein-protein interactions, manually as well as automatically curated, using computational approaches. The information provided in DIP is extracted from various sources and cross-references to major biological repositories (e.g. UniProt, RefSeq, and GO) is used for annotation of each association. IID comprises over 4.8 million in 18 species, including human, 5 model organisms and 12 domesticated species, while experimental evidence derived from other databases is combined based on several computational predictions. It also supports tools for topological and enrichment analyses of PPIs. Importantly, MINT, DIP, and IID are all active participants in the IMEx Consortium.

Two of the largest biological interaction repositories are BioGRID [88] (*Figure 13a*) and STRING [148]. The latter is thoroughly described in Chapter 6. BioGRID, in its current version, catalogs 1,740,000 protein-protein interactions curated from both high-throughput datasets and individual focused studies, derived from over 70,000 publications as well as genetic and chemical interactions and post-translational modifications. Both databases provide programmatic access through a REST API, as well as integration with Cytoscape. Even though BioGRID is not an active participant in the IMEx Consortium, it is a Prospective IMEx Consortium member, classified as IMEx Observer. Notably, I2D [149], formerly known as Online Predicted Human Interaction Database (OPHID), is another web-based collection of eukaryotic protein-protein interactions, which retrieves experimental evidence provided by BioGRID or IntAct, in addition to data obtained from high-throughput experiments. Predicted interactions are also offered and they are inferred by mapping experimental results between different species. In order to visualize and further analyze the PPI networks derived from its data, I2D implements NAViGaTOR [90], an online network analysis platform. I2D remains one of the most comprehensive sources of both known and predicted eukaryotic PPIs for model organisms, such as *S. cerevisiae*, *C. elegans*, *D. melonogaster*, *R. norvegicus*, *M. musculus*, and *H. sapiens*.

CORUM (Comprehensive Resource of Mammalian protein complexes) [56] (*Figure 13b*) is a resource of manually annotated mammalian protein complexes, whose information is retrieved exclusively from individual experiments published in literature. Despite the relatively few total number of interactions provided, data curation for each entry is significantly more detailed compared to other repositories. Particularly, CORUM 3.0 integrates 4274 mammalian protein complexes, while its annotation is based on the PSI-MI standard and includes protein complex function, localization, subunit composition, literature references, functional enrichment with GO terms, and associations with diseases.

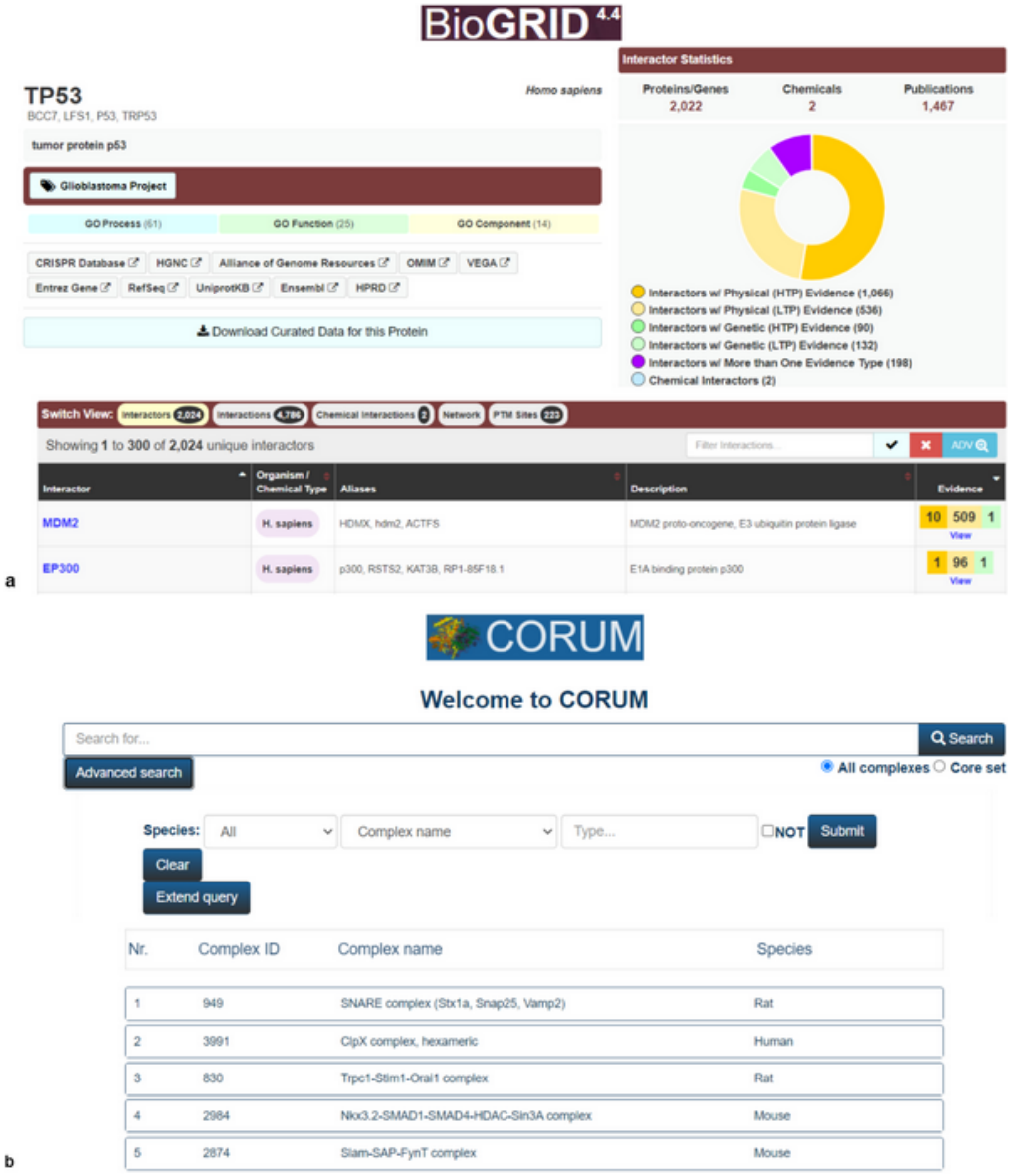


Figure 13: BioGRID and CORUM databases. (a) Different views of search results provided in the BioGRID database. **(b)** Advanced search options and table of query results in the CORUM database.

On the other hand, ComplexPortal [150] is a manually curated and annotated database of macromolecular complexes which even though covers a smaller number of mammalian protein complexes than CORUM, provides a broader species range. More specifically, ComplexPortal emphasized more on protein-protein associations, whereas it also provides protein-nucleic acid, and protein-small molecule complexes supported by experimental evidence, extracted and cross-referenced from the literature and deposited in IntAct. Another fundamental difference between CORUM and ComplexPortal is the definition of the term “macromolecular complex”. The latter is describing “complex” as an assembly of any two or more bioentities that are stable enough in vitro to be reconstituted and have been

demonstrated to have a specific molecular function. Thus, only constant protein-protein complexes are included in the database, while transient interactions are discarded.

Besides the previously mentioned comprehensive databases that cover a large spectrum of protein-protein interactions, a plethora of more specialized web services exist, focusing on specific associations and systems. For instance, databases such as the GPCRdb [151], the PrimesDB (Protein interaction machines in oncogenic EGF receptor signalling) [152] and the Channelpedia [153] are devoted to the interactions of particular groups of proteins with biomedical or pharmacological interest. Firstly, GPCRdb, as its name implies, catalogs structural and functional data on the interactions of G-protein coupled receptors (GPCRs) with ligands and heterotrimeric G-proteins. Moreover, PrimeDB focuses exclusively on the signaling mechanisms of Receptor-Tyrosine Kinases (RTKs) associations, including EGFR and ERBB, which are often implicated in diseases both as biomarkers and drug targets. This database also offers tools for the visualization of PPI networks and is a participant in the IMEx Consortium. Finally, Channelpedia is essentially a knowledge base providing both structured and unstructured evidence on ion channels and interactions between their subunits. Interestingly, the IUPHAR/BPS Guide to Pharmacology [154], which is another specialized online resource, concentrates and presents molecular interactions between all the protein classes mentioned above and their ligands in human, mouse and rat.

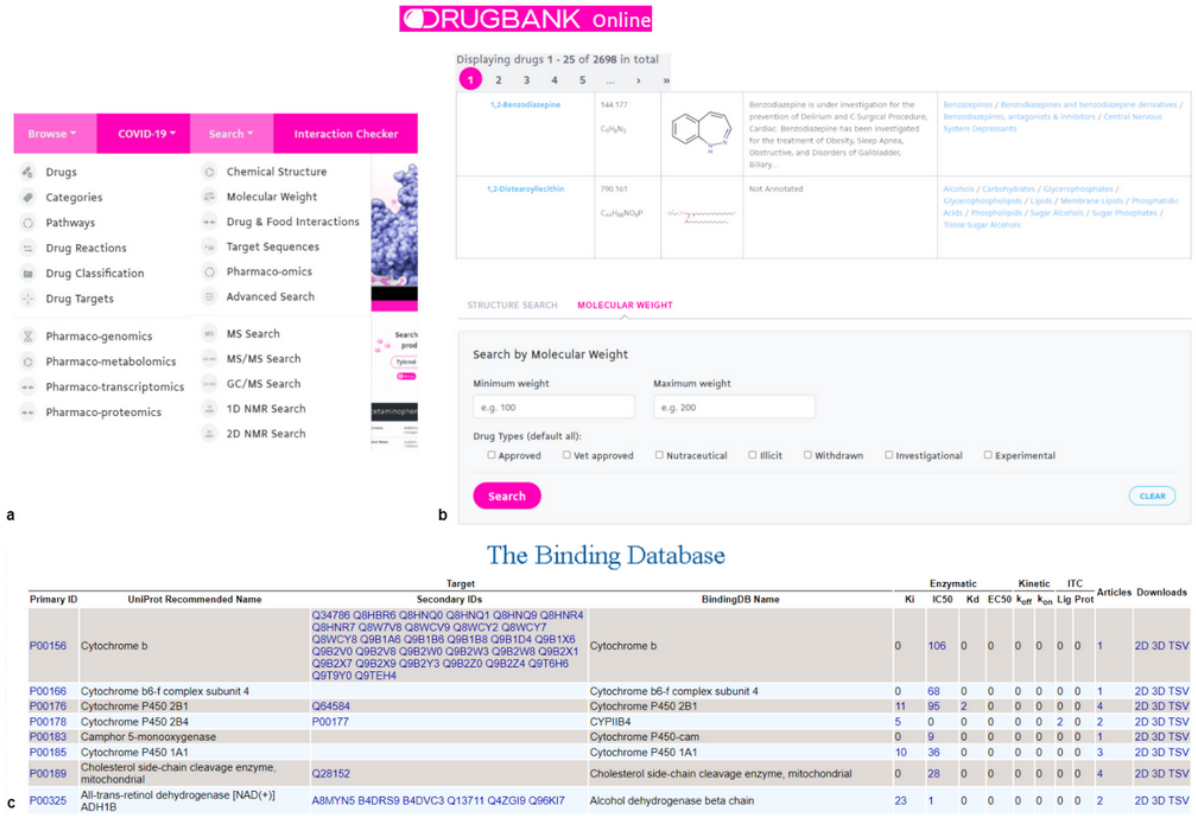
Lastly, in addition to protein-specific repositories, there is a plethora of databases which accommodate information on PPIs observed in specific subcellular locations. One example is MitoProteome [155] that provides interaction evidence on mitochondrial proteins, while PerMemDB [156] includes experimental data along with computationally predicted evidence on peripheral membrane proteins, including their interactions with transmembrane proteins. MatrixDB [157] is another manually curated repository of protein-protein interactions located at the extracellular matrix (ECM).

4.3.2 Protein-small molecule interactions

The interactions of proteins with small molecules are vital for a wide range of biological functions. Inside a cell, small molecules play a twofold role as substrates and products in various biochemical reactions and as ligands or hormones which regulate protein functions [158]. Additionally, bioactive small molecules are often used as probes to identify therapeutic protein targets in drug discovery. Information on the structures, calculated properties, and bioactivities for a large number of chemicals and drug-like compounds is integrated in specialized databases, including PubChem [159], ChEMBL [160] and SIDER [161], aiming to decipher their properties and facilitate the drug discovery process. Another essential data resource involves databases focused on protein-chemical interactions, which gather

information on the existence, stoichiometry and biological or biomedical relevance of protein-small molecule complexes [162].

The primary, and most often used source of information in protein-small molecule interactions comes from databases focusing on experimentally studied protein-chemical complexes. DrugBank [163] (*Figure 14a*) is currently one of the most popular databases in this category. It is a manually curated and publicly available resource that provides primarily experimental information about small molecules (i.e. chemical, pharmacological and pharmaceutical) and their protein targets (i.e. sequence, structure, metabolic pathways). In addition to drug-drug interactions, the database incorporates information for physical drug-target interactions and interactions with proteins known to metabolize a compound. Despite its name, however, the database does not focus solely on drugs, but also provides information on other compound types, such as metabolites. DrugBank is a frequently updated resource and the latest release (04/2021) integrates 14,524 drug entries, including 2,684 approved small molecule drugs, 1,464 approved biologics (proteins, peptides, vaccines, and allergenics), 131 nutraceuticals and over 6,654 experimental (discovery-phase) drugs. Finally, 5,249 non-redundant protein (i.e. drug target/enzyme/transporter/carrier) sequences are associated with the aforementioned drug entries. Another important, experimentally focused protein-small molecule interaction database is BindingDB [111]. BindingDB (*Figure 14b*) is a specialized repository of experimentally validated and measured binding affinities between drug-like compounds and therapeutically relevant protein targets. In particular, the latest version of BindingDB incorporates 41,328 Entries, each with a DOI, containing 2,259,122 binding data for 977,487 small molecules, which are mapped to 8,516 protein targets. The database is continuously curated, deriving data mainly from scientific articles as well as from US patents. The search interface is well-designed and enables combined query criteria, including target name, sequence, molecular weight, source organism, compound name, SMILES string, binding potency and article or patent information, while restrict search by data source (e.g., BindingDB, ChEMBL, PubChem, and patents) are also allowed.



Primary ID	UniProt Recommended Name	Target: Secondary IDs	Target: BindingDB Name	Enzymatic: Ki	Enzymatic: IC50	Enzymatic: Kd	Enzymatic: EC50	Kinetic: k_{off}	Kinetic: k_{on}	ITC: Lig	ITC: Prot	Articles	Downloads
P00156	Cytochrome b	Q34786 Q8HBR6 Q8HNQ0 Q8HNQ1 Q8HNQ9 Q8HNR4 Q8HNR7 Q8W7V8 Q8WCV9 Q8WCY2 Q8WCY7 Q8WCY8 Q9B1A6 Q9B1B6 Q9B1B8 Q9B1D4 Q9B1X6 Q9B2V0 Q9B2V8 Q9B2W0 Q9B2W3 Q9B2W8 Q9B2X1 Q9B2X7 Q9B2X9 Q9B2Y3 Q9B2Z0 Q9B2Z4 Q9T6H6 Q9T9Y0 Q9TEH4	Cytochrome b	0	106	0	0	0	0	0	0	1	2D 3D TSV
P00166	Cytochrome b6-f complex subunit 4		Cytochrome b6-f complex subunit 4	0	68	0	0	0	0	0	0	1	2D 3D TSV
P00176	Cytochrome P450 2B1	Q64584	Cytochrome P450 2B1	11	95	2	0	0	0	0	0	4	2D 3D TSV
P00178	Cytochrome P450 2B4	P00177	CYPIIB4	5	0	0	0	0	0	2	0	2	2D 3D TSV
P00183	Camphor 5-monooxygenase		Cytochrome P450-cam	0	9	0	0	0	0	0	0	1	2D 3D TSV
P00185	Cytochrome P450 1A1		Cytochrome P450 1A1	10	36	0	0	0	0	0	0	3	2D 3D TSV
P00189	Cholesterol side-chain cleavage enzyme, mitochondrial	Q28152	Cholesterol side-chain cleavage enzyme, mitochondrial	0	28	0	0	0	0	0	0	4	2D 3D TSV
P00325	All-trans-retinol dehydrogenase [NAD(+)] ADH1B	A8MYN5 B4DRS9 B4DVC3 Q13711 Q4ZGI9 Q96KI7	Alcohol dehydrogenase beta chain	23	1	0	0	0	0	0	0	2	2D 3D TSV

Figure 14: DrugBank and the BindingDB databases. (a) Browse and search options of the DrugBank database. **(b)** Example of drugs' display upon selecting the Drugs from Browse options (upper figure). Search interface based on molecular weight. **(c)** Table view of search results in the BindingDB.

Apart from the primary databases described above, several secondary repositories also exist, combining information from multiple sources. STITCH (Search Tool for Interactions of Chemicals) [164], the "sister" database of STRING, is a manually curated resource to explore both known and predicted interactions between 9,600,000 proteins from 2,031 eukaryotic and prokaryotic genomes and over 430,000 chemicals. Known interaction evidence is mainly derived from experimentally validated data as well as from manually curated datasets, including KEGG and Reactome. Protein-small molecule interactions are also accompanied by protein-protein interaction evidence, derived from STRING, to help illustrate the effect of chemicals on supramolecular assemblies. Text-mining based associations are compiled after parsing articles from PubMed Central (PMC) and PubMed. Like STRING, STITCH offers a REST API for programmatic access, as well as integration with Cytoscape.

A major field of interest in the study of protein-small molecule interactions involves the structural analysis of protein-ligand complexes. A number of specialized databases exist for this purpose. Some of these repositories are, essentially, subsets of PDB, containing analysis on the stoichiometry of protein - heteroatom interactions often found in the PDB

entries of experimental 3D structures. PLI (Protein-Ligand Interaction) [165] and PLIC (Protein-Ligand Interaction Clusters) [166] are two such databases that, as their names indicate, focus on protein-ligand associations. PLI database incorporates all the interactions between proteins and small molecules identified in the PDB with a Het_id code, while PLIC, by analyzing the similarities in binding sites and employing computational tools, provides clusters of similar binding sites from PDB. Notably, PLIC, unlike other protein-ligand specific databases, not only reports similarities in interactions but also hosts data on attributes like binding site shape, protein–ligand contacts and energetics among similar protein–ligand interactions.

In addition to the above, a number of structural databases also exist that complement crystallographic evidence with computational predictions, often derived from energy calculations, protein-ligand docking predictions or ab initio simulations. NLDB (Natural Ligand Database) [167] is a predictive database focusing on 3D protein-ligand interactions, specifically in enzymatic reactions of metabolic pathways registered in KEGG. Based on the latest update, NLDB offers data about known human genome polymorphisms on protein structures, as well as 87,400 experimentally validated protein-ligand complex structures in PDB, defined as natural complexes, while 31,672 analog complexes and 70,570 Ab initio complexes were predicted based on known protein structures in a complex with a similar ligand and by docking simulations accordingly. In case of unknown complex structures, 3D interactions are predicted by implementing state-of-the-art software programs and subsequently generating a database of the 3D protein-ligand interactions in various enzymatic reactions. NLDB also provides an enrichment analysis function based on a set of KEGG compound IDs. PoSSuM (Pocket Similarity Search using Multi-Sketches) [168] is another predictive database that aims to retrieve similar small-molecule binding pockets on proteins with both different and similar global folds, contributing to the structure-based drug discovery. It employs the SketchSort [169] algorithm for all-pair similarity searches, resulting in more than 163 million similar pairs of binding sites with annotations. Finally, PDID (Protein-Drug Interaction Database) [170] is a database of predicted protein-ligand interactions in the structural human proteome. PDID incorporates 9,652 structures from 3,746 proteins and provides a comprehensive set of 16,800 putative protein-drug interactions between 51 popular, FDA-approved drugs and over 10,000 protein structures, which were generated from approximately 1.1 million all-atom structure-based predictions.

The databases described above offer generalized information on the existence and properties of protein-small molecule complexes. However, specialized repositories also exist, focusing on the protein-chemical interactions associated with specific systems, phenotypes or diseases. One characteristic example involves cancer-specific databases, such as CancerDR [171] (*Figure 15a*), CAncerREsource 2 [172] (*Figure 15b*) and canSAR [173] (*Figure 15c*).

As their names indicate, these databases focus on protein-drug interactions related particularly to cancer. CancerDR incorporates 148 anticancer drugs which are mapped to 116 drug targets in 1000 cancer cell lines, offering also information about the function, structure, and gene sequences of each of these targets. In addition, CancerREsource 2 contains not only comprehensive data on 90,744 interactions between drugs and cancer-relevant protein targets, but also mRNA expression and non-synonymous mutation data from large-scale cancer genomics experiments. Similarly to the previously mentioned databases, canSAR is a comprehensive database which integrates protein-drug interactions between 564,407 proteins from all species and 3,312,866 compounds with unique chemical structures, as well as genomic and structural data.

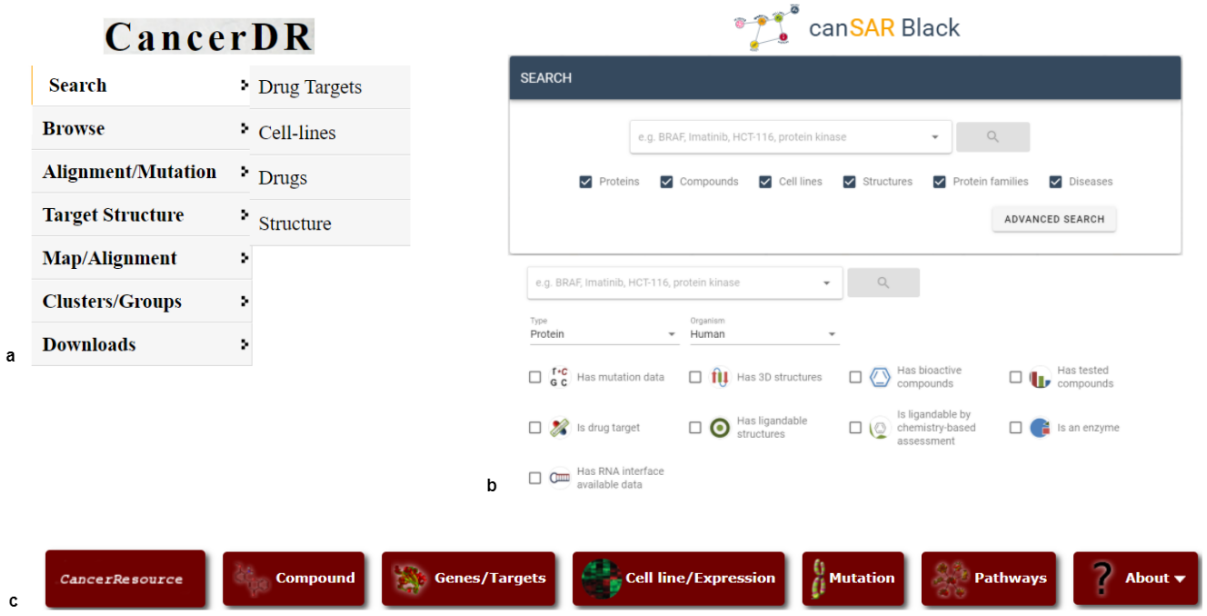


Figure 15: CancerDR, canSAR and CancerResource databases. (a) In the CancerDR database users are allowed to search by drug targets, cell lines, drugs and structure. Moreover, multiple sequence alignment of variants of drug targets, clustering mutants and predicting the tertiary and secondary structure of a drug target are also provided. **(b)** Simple (upper figure) and advanced search options of canSAR database. **(c)** CancerResource enables searching by compound, gene/target, cell line, mutation and pathways.

4.4 Signaling and metabolic pathways interactions databases

Signal transduction pathways, or else cell signaling, is the transmission of environmental and molecular signals from a cell's exterior to its interior (nucleus or target molecules), by cascades of modifications [54] . On the other hand, metabolic pathway networks describe the chemical reactions and/or the regulatory interactions between metabolites, that are mostly small biomolecules and enzymes. The simulation of these biological pathways has become almost a prerequisite in the majority of the fields in Biology, in order for a more efficient conceptualization to be accomplished.

First of all, one of the major biological pathway databases is Reactome [54] (*Figure 16*). The retrieval of information concerning cellular processes on a molecular level from 33,453 literature references enables the generation of numerous pathways and superpathways, which essentially constitute an extended metabolic map of *H. sapiens*. By corresponding human proteins to their molecular functions, Reactome plays a twofold role. Particularly, it is both a tool which facilitates the uncovering of functional relationships and a repository of biological processes, including transport and DNA replication, signal transduction as well as intricate metabolic functions. Currently, the database (version 76) contains 10,867 human genes, 415 drugs, 1,856 small molecules which serve as natural substrates, catalysts or regulators, 11,073 discrete proteins and 13,732 reactions incorporated into 2,516 human pathways grouped in 26 superpathways. Reactome data is downloadable in various formats and can be queried via an API.

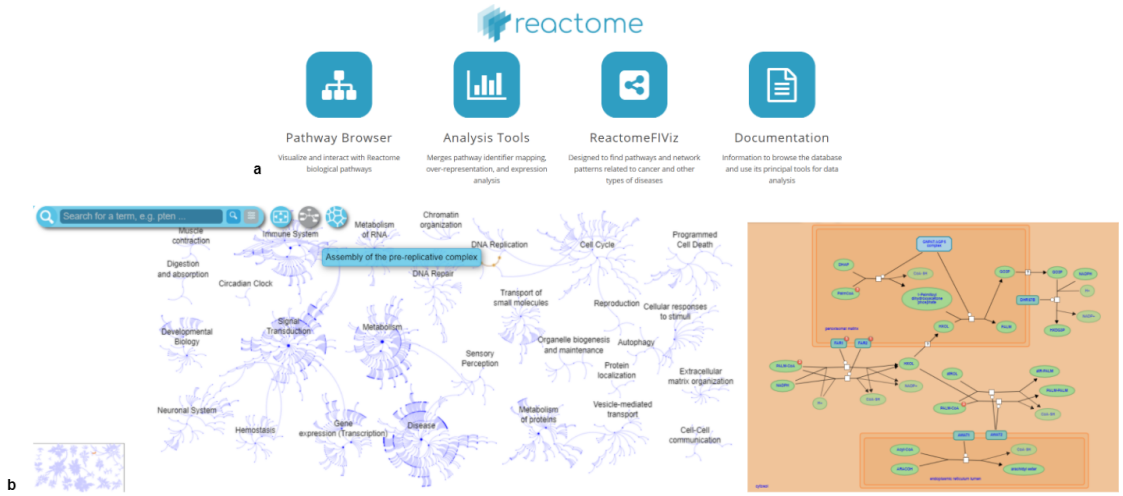


Figure 16: Reactome database. (a) The Reactome database enables the browsing and analysis of biological pathways as well as disease-related pathways and network patterns. **(b)** Interactive visualization of multiple different biological pathways (left). Graphical representation of a selected pathway.

In addition to Reactome, another important resource of signaling and metabolic pathways is KEGG [53]. The distinctiveness of KEGG lies in the integration of eighteen different databases which are manually curated and categorized into systems, genomic, chemical and health information. The central database of KEGG is KEGG PATHWAY [174] that consists of biological pathways represented graphically by manually drawn maps, similar to Reactome. Interestingly, within the pathway maps, fully sequenced genomes of cellular organisms are linked in a way to infer high-level functions. Such functions are depicted by a web of interactions and chemical reactions, drawn in the format of KEGG pathway maps, BRITE hierarchies, and KEGG modules. KEGG contains 34,042,792 genes, 781,759 pathways and 11,505 reactions pertaining to 545 eukaryotes, 6234 bacteria, and 343 Archaea (April 2021).

Another publicly accessible and user-friendly repository is WikiPathways [55]. It is a community-driven database, enabling its maintenance and curation by and for the scientific community. Herein, it facilitates the contribution of data on an almost daily basis, enhancing and complementing existing resources, such as KEGG and Reactome. Building based on the MediaWiki software, WikiPathways incorporates custom graphical tools for the representation of a plethora of biological pathways (*Figure 17*), while it also includes content originated from a large selection of databases covering major gene, protein, and small-molecule systems. A total of 2958 pathways (April 2021) and 46,105 interactions between proteins, genes, metabolites, and drugs are provided for 30 different species. The database offers an API for programmatic access as well as integration with PathVisio [175] and Cytoscape for further pathway analysis.

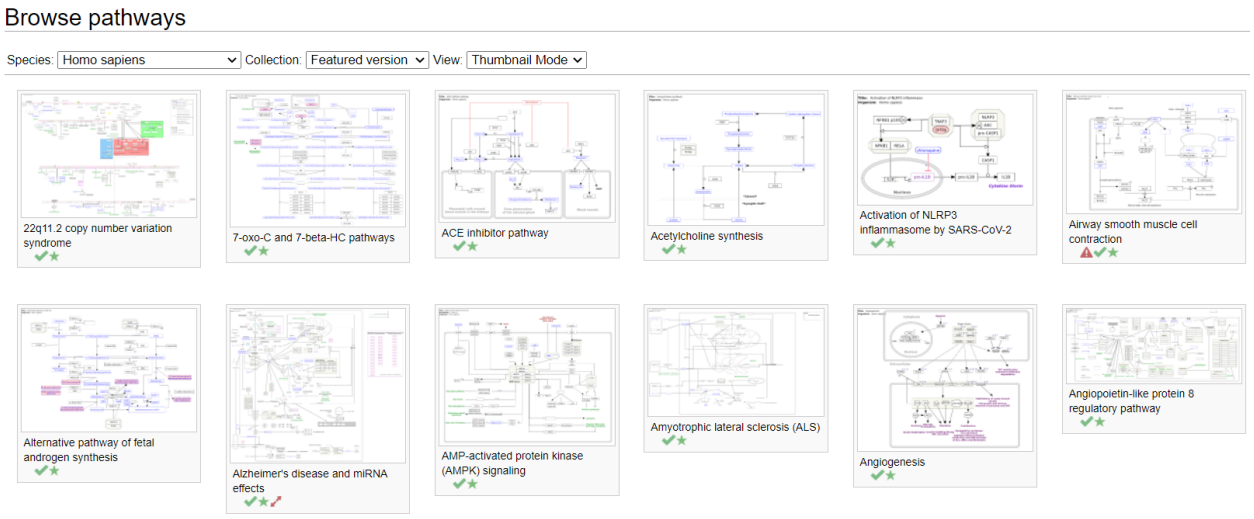


Figure 17: WikiPathways database. Overview of several available biological pathways of *Homo sapiens* in WikiPathways database.

Lastly, CBN (Causal Biological Network) [176] is a manually curated database which provides more than 120 network models using Biological Expression Language (BEL) [177] supported by over 80,000 unique, literature-based pieces of evidence. These causal networks represent the relationships in signaling pathways in 3 species (*H. sapiens*, *M. musculus* and *R. norvegicus*), covering a wide spectrum of biological processes, such as cell fate, cell stress, cell proliferation, inflammation, tissue repair and angiogenesis in the pulmonary and vascular systems (*Figure 18a*). Interactive data visualizations of proteins, DNA variants, coding and non-coding RNAs, chemicals, lipids, and processes (e.g., phosphorylation) is provided, allowing the user to model a network at will (*Figure 18b*). Notably, most of the pathway components and associations are annotated with a variety of metadata, regarding species, tissue and cell type.

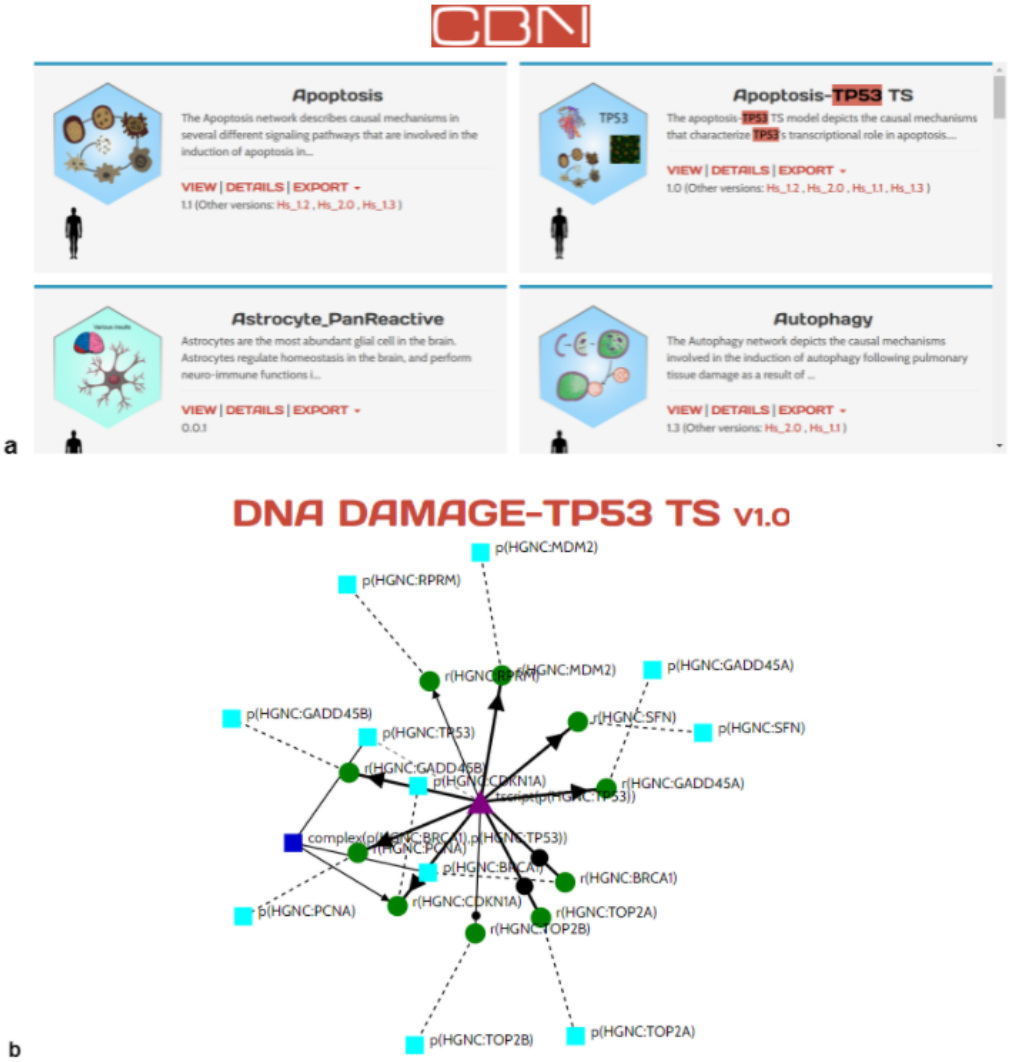


Figure 18: CBN database. (a) Several of the resulting biological pathways upon querying by a specific gene name (tp53). **(b)** Example of dna damage-tp53 pathway visualization.

4.5 Disease-related interactions databases

Disease-related elements networks have been proven a powerful way to elucidate the obvious as well as the hidden connections among the molecular origins of the disease and the resulting phenotypes [178]. In this type of networks, nodes could correspond to multiple different molecular entities, including RNAs, genes, proteins and diseases along with their phenotypes, while the links between them represent the underlying biochemical interactions. Notably, the untangling of these sometimes complex relationships has been beneficial, especially in the field of drug repurposing.

First of all, two specialized repositories focused on human disease-related intermolecular interactions are CIDeR [179] (*Figure 19*) and MiRNA SNP Disease Database (MSDD) [180]. Both databases are manually curated, providing data retrieved from the literature and accompanied by metadata. However, their main difference lies in the information each database is composed of. CIDeR is dedicated to metabolic and neurological disorders, containing 109,779 interactions between 12,406 biological entries (e.g. biomolecules, pathways, biological processes, phenotype), derived from 11,341 parsed articles. Furthermore, a plethora of interaction types is supported, such as expression patterns, co-occurrence, co-localization, processing, phosphorylation, transport, and folding. All entries are enriched with their corresponding PubMed ID and other related diseases, in addition to metadata which refer to the affected organism, tissue/cell line, and gender. Interconnectivity with Entrez Gene, KEGG, OMIM, miRBase, GO, CORUM, Mammalian Phenotype Ontology (MPO) [181] and BRENDA Tissue Ontology (BTO) [64] as well as 2D network visualization are also provided. On the other hand, MSDD contains 525 associations between 182 human miRNAs and 197 SNPs, regarding 153 genes and 164 human diseases, mined from 2,387 articles (last update: June 2017). Each interaction is accompanied by metadata, regarding tmiRNAs, SNPs, miRNA target genes and disease names, SNP locations and alleles, the miRNA dysfunctional pattern, experimental techniques, a brief functional description, the original reference and additional annotation, while data are freely downloadable in text format.

a

CIDeR

COVID-19: blood coagulation

Entry-ID		Interactor1 (Subject)	Interaction type	Interactor2 (Object)	additional information	PubMed-ID	Diseases
12.178	139750	COVID-19	increases_quantity of	C5A		32544563	COVID-19
13.337	125876	SARS-CoV-2 infection	affects_expression of	PLAT			COVID-19
	125877	SARS-CoV-2 infection	affects_expression of	EDN1			

b

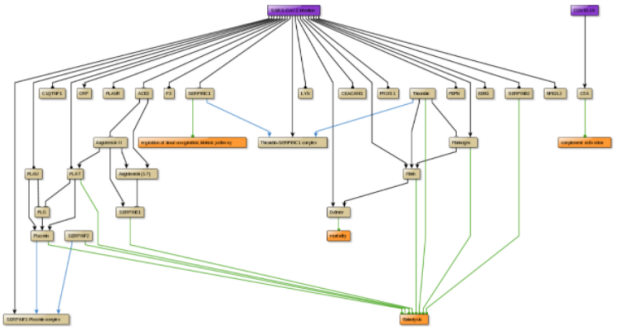


Figure 19: CIDeR database. (a) List of the interactions implicated in COVID-19 blood coagulation. **(b)** Graphical representation of the interactions between gene/proteins and processes in COVID-19 blood coagulation.

Besides CIDeR and MSDD, there are also other databases that document direct biomolecule-disease associations. DisGeNET [182] for instance is one of the largest versatile platforms of genes and variants involved in human diseases. It integrates data from curated repositories, GWAS catalogues, animal models and the scientific literature, regarding disease-gene, disease-variant, and disease-disease associations. It is regularly updated and it currently covers 1,134,942 gene-disease and 369,554 variant-disease associations, based on 30,170 disease entries (UMLS [183]), 21,671 genes (NCBI), and 194,515 variants (dbSNP). DisGeNET data is downloadable in tab-delimited and SQLite database formats. All interaction data are also accessible programmatically through a REST API, an RDF API, the disgenet2r R package and the Cytoscape application.

Additionally, EnDisease [184] (*Figure 20*) is a manually curated database specialised exclusively in enhancer-disease associations. It incorporates 535 experimentally verified associations between 133 diseases and 454 enhancers, extracted from 199 published articles in 11 species. The interaction data are freely downloadable in text format and represent the chromosomal position of the enhancer, the targeted gene and its UCSC identifier, and the related disease, with a respective entry link to the OMIM database.

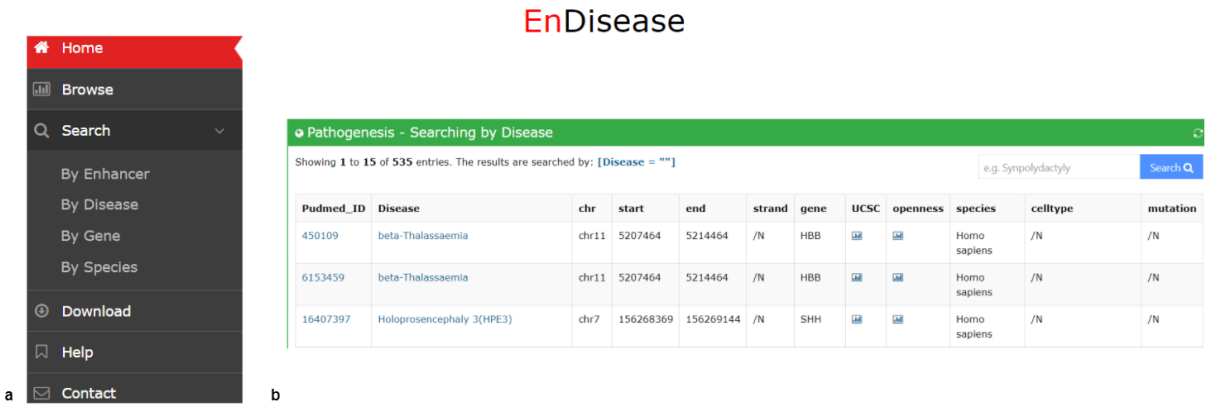


Figure 20: EnDisease database. (a) EnDisease database allows searching by enhancer, disease, gene and species. **(b)** Results displayed in a table view upon searching by disease.

A distinct category of disease-related associations includes host-pathogen interactions, which are available in various different repositories. Viruses.STRING [185], an extension of the STRING database mentioned in previous paragraph, provides exclusively intra-virus and virus-host protein-protein interactions. Particularly, 1,380,838,440 physical or functional interactions between 2,031 organisms and more than 9,5 million viral proteins are covered and supported by experimental and text-mining evidence. Importantly, the database enables interactive visualization via the generation of networks based on the queried interactions where all node entries are linked to Uniprot. Viruses.STRING is freely available and all data can be accessed and analyzed through a REST API and the Cytoscape STRING app, while they are also downloadable in text format and the whole database schema in SQL format.

Lastly, ViRBase [186] is a manually curated viral-host interactions online repository of virus-host ncRNA-associated interactions. The current version (v.21) consists of 781,476 viral and cellular ncRNA interactions between 93 viruses and 27 hosts, derived from 491 articles. In particular, microRNA entries were collected from miRBase, lncRNAs from lncRNAdb and the functional lncRNA database [187], snoRNAs from sno/scaRNAbase [188] and snoRNA-LBME-db, whereas ICTVdb (International Committee on Taxonomy of Viruses) [189] records provided virus names and abbreviations. Furthermore, users are allowed to query these interactions through an API and download them in XLSX and text formats.

4.6 Ecological interactions databases

The different species in a particular habitat are interlinked by symbiotic, mutualistic (bidirectional) or competitive (host-parasite) relationships, which lead to the formation of networks depicted mostly as food webs or interspecies interactions [190]. Therefore, databases that incorporate information on ecosystems and ecological network's structure are essential for the uncovering of the population dynamics, biodiversity and ecosystem function. For this reason, resources dedicated to species interactions and trophic webs are described in the following paragraph.

To begin with, Global Biotic Interactions (GloBI) [190] (*Figure 21*) is an extensive, online infrastructure of manually curated species interactions, retrieved from 284 open datasets (data journals and APIs) via an open source software. Besides the GloBI web interface, Encyclopedia of Life (EOL) [191] and Gulf of Mexico Species Interactions (GoMexSI) [192] projects incorporate structured species-interaction information from GloBI. Currently, it integrates 33 different interaction types (e.g. eats, kills, interacts with, parasite of) and a total of 8,148,483 interaction records, between more than 700,000 taxa, regarding predator-prey, pollinator-plant, pathogen-host, parasite-host relationships. Importantly, each record is attributed to a scientist, research institution, or other source, while entries that contain known taxa are additionally cross-referenced with entries in NCBI, World Register of Marine Species (WoRMS) [193], Integrated Taxonomic Information System (ITIS) (https://www.itis.gov), and Global Biodiversity Information Facility (GBIF) (https://www.gbif.org). GloBI enables programmatic access through a REST API, R (rglobi), JavaScript (eol-globi-data-js) libraries or SPARQL as well as Cypher queries.

Figure 21: GloBI database. (a) The user is enabled to search the interactions of interest by specifying the focal taxa and interaction type. **(b)** Hairball (left) and bundle diagram (right) representation of the interactions between different taxa. In the hairball diagram taxa are represented by nodes and their connections are indicated by the edges.

Similarly, the Web of life [194] (*Figure 22*) is another user-friendly, web-based service of ecological interactions which provides a graphical user interface, based on Google Maps for visualizing, searching and downloading ecological networks in a coordinate-based system. In contrast to GloBI, Web of Life is focused especially on relationships between animal-plants, plants-plants and host-plants, providing only “interacts with” type of association. At this moment, Web of Life contains 186 interaction networks, regarding 13,244 animal and plant species, which have been assembled by data from both published and unpublished projects. All data can be downloaded in various formats and a data transmission webservice in JavaScript Object Notation is also provided.

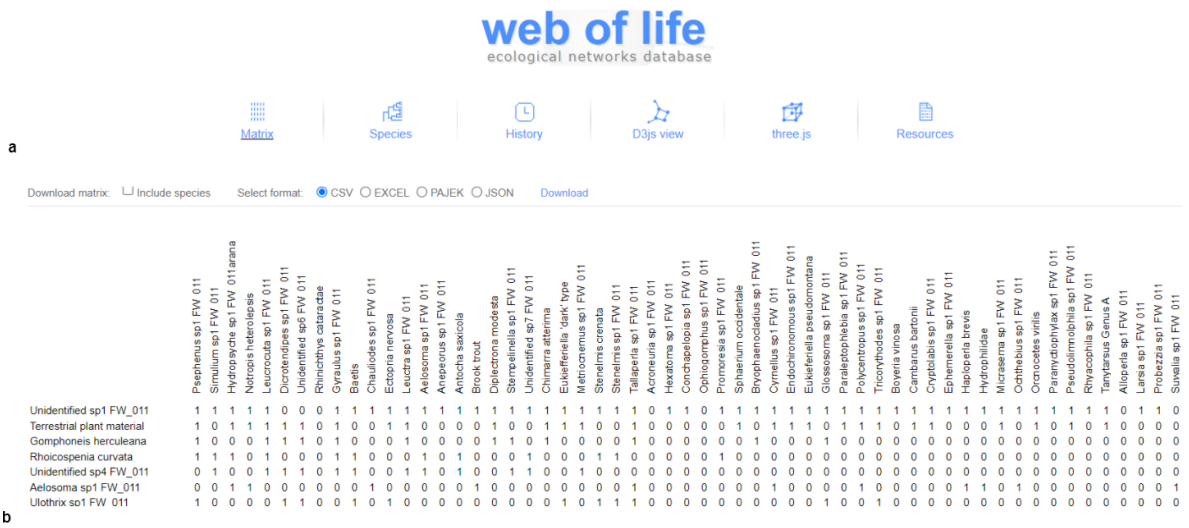


Figure 21: Web of life database. (a) Different available views of search results. **(b)** Species interactions depicted in matrix format.

Moreover, another freely accessible tool that provides an online collection of direct trophic interactions between approximately 7000 animals and plants is Food Web (GlobalWeb) [195]. The database currently hosts over 360 food webs, retrieving information manually from more than 120 reference papers. Each food web is enriched with data concerning the type of food web (e.g. marine/freshwater/terrestrial), the habitat (e.g. lake/river/estuary), the location, the size, the resolution of the web as well as additional characteristics.

Finally, Bat Eco-Interactions [196] is an online platform dedicated to bat interactions with plants and arthropods. More specifically, 13,383 interactions that occur between 479 bat species and 2,135 other organisms are provided, referring to several types of associations including consume, host, transport, cohabitate, roost, or be consumed. All information is gathered from 622 published and peer-reviewed articles and they are available in CSV format after registration. The database receives regular updates with bat-parasite and bat-mammal interactions, which include taxonomic and location metadata.

CHAPTER 5: Biological network analysis

5.1 Visualization tools for biological networks

Systematic understanding of genomic scale data is often required so as to better analyze and interpret complex biological concepts. Especially with the advent of “omics” science, established high-throughput technological advances have significantly contributed to the emergence of a networked perspective of contemporary biology, promoting a holistic, interconnected picture of cells with myriad intermolecular interactions. These associations reflect the inherent dynamics and heterogeneity of biological systems which is inevitably accompanied by great complexity. Therefore, the development of efficient, advanced and informative visualization tools becomes a necessity as flexible and comprehensible graphical representations enhance the ability to perceive and interpret the high dimensionality and interconnectivity of vast amount of data [197].

Biomolecular interactions are usually characterized as extensive networks of numerous vertices and edges which represent the components of a biological system (e.g. genes, proteins, metabolites, other small molecules etc.) and their relationship respectively[198]. Exploiting the versatility of graphs, edges express various different types of associations between bioentities, including physical interaction, functional annotation, evolutionary relationship, gene co-expression and literature co-occurrence, often resulting in multi-edge networks in order to capture the whole range of information [199]. Such variability in network types implies divergent properties and topological features, highlighting the importance of graph theory [200].

Currently, a variety of specialized and multifunctional visualization tools have been developed, enabling the data storage, retrieval, exploration, comparison and analysis. Certain representative examples are: Cytoscape [201], Gephi [52], Pajek [202], Ondex [203], Proviz [204], VisANT [205], Medusa [197], Osprey [206], Arena3D [207], and BioLayout Express [208]. Briefly, Proviz is a tool focused on protein-protein interactions, Ondex is a comprehensive database whose implementation allows graph-based analysis, Gephi and Pajek are mainly used to visualize large-scale generic networks, Medusa is specialized in network clustering and visualization of multi-edged graphs, while Arena3D in 3D multi-layer networks and BioLayout Express in advanced 3D visualizations. Finally, Osprey and Cytoscape have several plugins which enable the visualization and annotation of heterogeneous data [209].

A similar list of widely used tools for pathway analysis and visualization include: WikiPathways [55], a publicly accessible database for biological pathway editing maintained by the scientific community, KEGG [53] and KaPPA View [210] which focus on metabolic pathways from multiple organisms, as well as Interactive Pathways Explorer (iPath) [211] and PathVisio

[175], interactive tools for analysis, visualization and editing of biological pathways. MetaboAnalyst [212] is another important web-based tool suitable for statistical, functional and integrative analysis of metabolomic data [209].

5.2 The STRING database

5.2.1 Usage

STRING (v11.5, 11/2020) (Search Tool for the Retrieval of Interacting Genes and proteins) is a publicly accessible online metadatabase available at https://string-db.org/, which integrates experimentally validated and computationally inferred protein interaction networks for 5090 different genome-sequenced organisms [148]. All the interactions provided are annotated as direct (physical) or indirect (functional) and they are consolidated by data derived from various sources, including prior knowledge about experimentally determined associations, pathways and protein complexes from curated databases, as well as computationally predicted interactions from literature text mining of scientific texts, systematic co-expression analysis, genome-wide association studies and gene orthology [69,148]. In addition, STRING is programmatically accessible via a REST API, packages for the R and Python languages and direct integration with Cytoscape.

Users have the opportunity to query STRING via protein name(s) or identifier(s) of a specific organism of interest or via protein family by searching the clusters of orthologous groups (“COGs”). Alternatively, the raw amino acid sequence(s) can be supplied in any format as well as an entire experiment given as a list of proteins, optionally accompanied by a ranking value (e.g., fold-change or P-value). When at least one of the input forms have not been filled, a disambiguation page will be generated for this purpose [209].

Once querying criteria are applied, a fully interactive network visualization platform is provided in a results page for the analysis of the protein-protein interactions and the topological features (*Figure 8a*). A summary view of the network legend and the predicted functional links -ranked by estimated confidence- are also available (*Figure 8b*). The confidence score is scaled between zero and one and is calculated for all protein interactions by combining all the individually measured scores. In addition, a popup window is generated upon clicking a node or an edge of the graph containing concise information on the particular node or the associations respectively (*Figure 8c*). STRING also enables automated functional enrichment and network functional annotation with terms originated from various established repositories, including PubMed, OMIM [213], KEGG [53], Reactome

pathways [54], UniProt [214], Ensembl [83], GeneCards [215], RefSeq [118], Pfam [61], InterPro [216] and SMART [217].

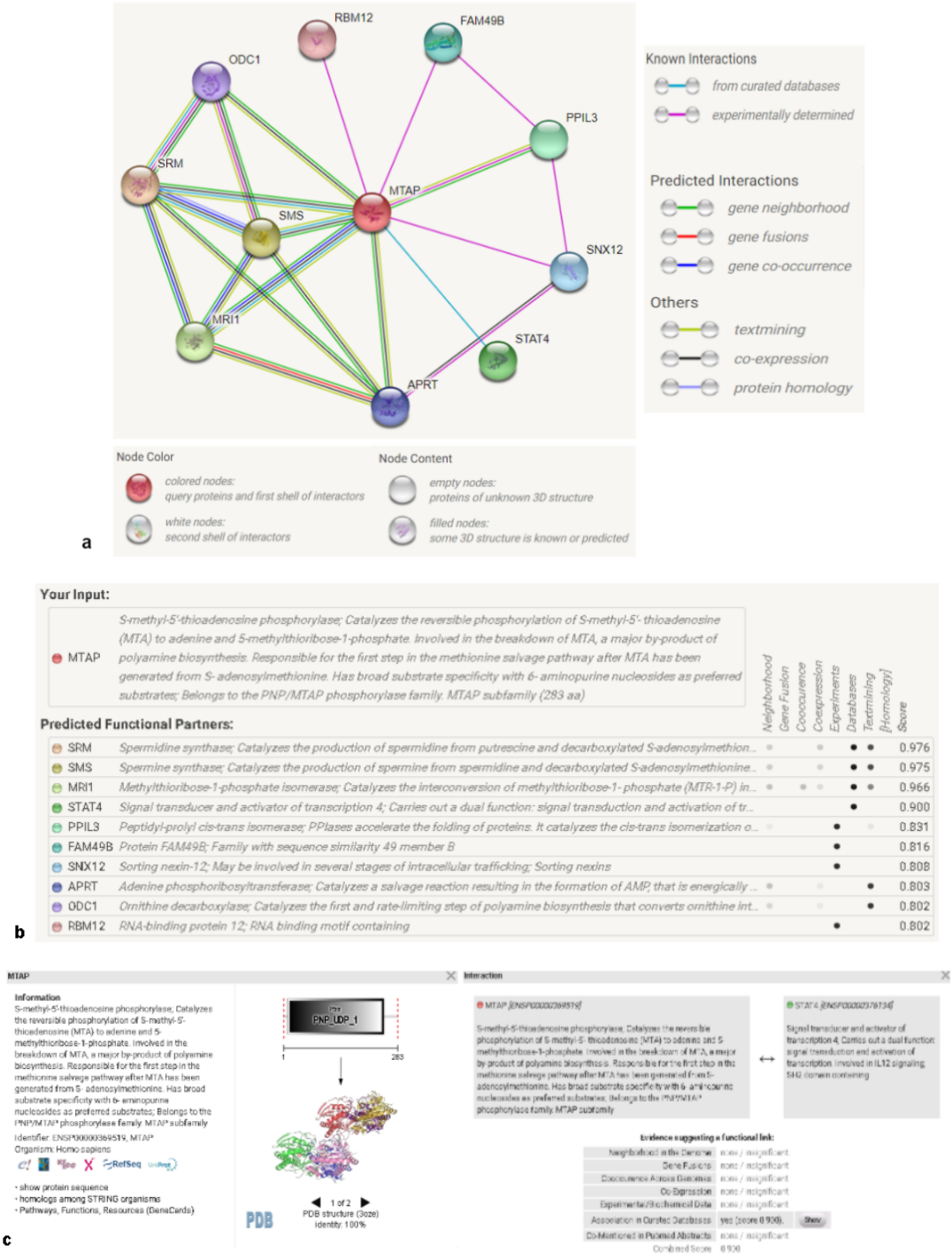


Figure 8: Network viewer in STRING. (a) A typical protein-protein association network as a multi-edge graph, where nodes represent proteins and edges represent their connection. As the legend indicates, each edge is colored based on the type of evidence that supports the particular interaction. **(b)** A table that contains predicted functional links ranked by estimated confidence scores. **(c)** Popup windows generated upon clicking on the MTAP node and the MTAP-STAT4 edge.

5.2.2 Database content

Each interaction type depicted as an edge of different color in the network visualization, stems from the combination of separately scored and specified association evidence, which is divided into one or more of the eight distinct “channels”, depending on the origin and type of the evidence (*Figure 9*). Notably, all channels provide interaction scores and viewers, while they can also be disabled individually or in combinations. Briefly, the prediction channels are:

Experiments -- This view focuses on experimentally validated protein-protein interaction data (e.g. biochemical, biophysical and genetic experiments), reported in primary, curated databases, including BIND [218], DIP [146], GRID, HPRD, IntAct [219], MINT [144], and PID [220]. The individual scores are calculated upon re-evaluation of the interaction records and their mapping against the KEGG database.

Databases -- In this channel, manually curated interaction evidence is imported from external pathway databases, such as KEGG [53], Reactome [54], BioCyc [221], Biocarta, PID and Gene Ontology [222]. Notably, due to data filtering which indicates the existence of only direct protein interactions, the confidence score of all associations -solely in the databases channel- is uniform (0.900).

Text-mining -- Statistical sentence or abstract based co-occurrence analysis across PubMed abstracts, articles from the PMC, OMIM (46) and SGD (47) is conducted for the text-mining channel. Updated dictionaries which contain a plethora of gene and protein names are implemented for accurate Name Entity Recognition (NER), while Natural Language Processing (NLP) is used to identify semantic links between proteins. Following the extraction of co-mentioned proteins from the literature, each pair is assigned an association score reflecting their frequency of co-occurrence [209].

Co-expression -- The output of this channel results from the collection, normalization, redundancy reduction and subsequent correlation of gene expression experiments (using both transcriptome and proteome measurements) deposited in the NCBI Gene Expression Omnibus (NCBI GEO) [223]. Similarly to the aforementioned channels, each protein pair is assigned an association score, calculated by a Pearson correlation against KEGG pathway maps. The greater the similarity in normalized expression patterns of a protein pair, the higher their co-expression score.

Fusion -- This view presents the individual gene fusion events per species and a dendogram depicts the clustering of these species in which a fusion event occurs. An association score is given to a pair of proteins that are likely the result of a fusion event. The score depends on the fusion of respective orthologs in at least one other organism/genome. A higher score is a better predictor of orthology of the participating genes.

Neighborhood -- Genes that are consistently reported to be located at short physical distance in the genome are shown in this channel. Specifically, the association score given to a pair of proteins is determined based on their proximity to each other on the chromosome (e.g. conserved, co-transcribed operons), as well as on their genomic similarity with other species, implying shared protein functions between co-expressed, neighboring genes. The functionality of this channel is mostly relevant for Bacteria and Archaea.

Co-occurrence -- In this channel, the similarity of occurrence patterns of a pair of genes throughout evolution is evaluated. Such similarities may be the consequence of transfer, loss or duplication of these genes at the same period during evolution. Orthologs that have a tendency to be present or absent in the same subsets of organisms, are assigned an association score.

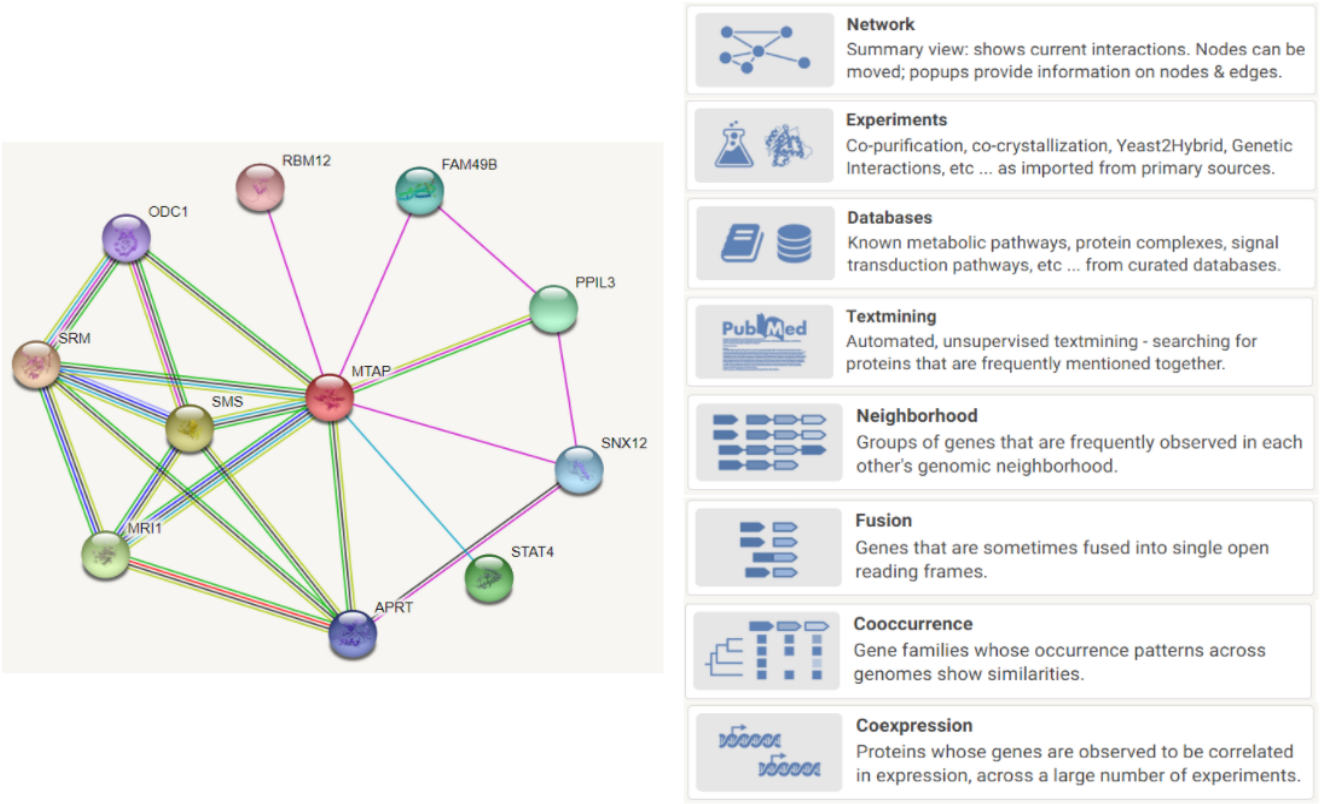


Figure 9: The eight prediction "channels". The prediction channels are: Network, Experiments, Databases, Text Mining, Neighbourhood, Fusion, Co-occurrence and Co-expression and they can be disabled individually or in combination. An interaction score and a unique viewer are provided in each channel.

5.3 The STITCH database

5.3.1 Sources of protein-compound interactions and database access

The significance of interactions between small molecules and proteins is evident in any biological system, as intracellularly they participate in various biochemical reactions both as substrates and products, regulating many protein functions. Particularly in diseases, which are often induced by alterations in the same biological pathway or protein complex, protein-small molecule interactions are essential to better understand the cellular impact of a drug. Additionally, numerous bioactive small molecules are often used as probes to identify therapeutic protein targets in drug development area. Therefore, the integration and combination of multiple sources of protein-chemical interactions from pathway databases, text mining and drug–target predictions facilitate the gaining of a holistic view.

Among the several online database focused on protein-chemical interaction networks, the STITCH 5 (Search Tool for Interactions of Chemicals) is a user-friendly and manually curated resource, which enables the investigation and analysis of both known and predicted interactions between 430,000 compounds and over 9,600,000 proteins across 2031 eukaryotic and prokaryotic genomes. The association evidence is mainly derived from manually curated datasets, including DrugBank [163], GPCR-ligand database (GLIDA) [224], Matador [225], the Therapeutic Targets Database (TTD) [226], the Comparative Toxicogenomics Database (CTD) [227] and several pathway databases such as the Kyoto Encyclopedia of Genes And Genomes (KEGG) [53], Reactome [54], and BioCyc [221]. This information is also combined with protein–protein interactions stored in the STRING database.

Experimentally validated data constitute another important source of interactions information, collected from ChEMBL [160], PDSP Ki Database [228], Protein Data Bank (PDB) [229] and two high-throughput kinase–ligand interactions studies [230,231]. Finally, automated text mining and a structure based prediction method [232] are implemented after parsing articles and abstracts from MEDLINE, PubMed Central and NIH RePORTER (https://projectreporter.nih.gov/). Both co-occurrence analysis and NLP methods are applied in the text-mining pipeline.

Besides the web interface which is available at http://stitch.embl.de/, STITCH also offers full programmatic access via an extensive API, allowing the alteration of all network parameters and the creation of images. Large-scale analysis is enabled via the freely downloadable precomputed network and the supplementary information [164].

5.3.2 Network channels and views

As an entry point, users can search by providing at least one identifier, chemical or protein name of interest accompanied by a selected organism. STITCH also allows the query of chemical structures as SMILES strings and proteins as amino acid sequences in any format. A disambiguation page will appear if neither an organism nor a chemical/protein has been added to input forms.

Notably, the results page of the STITCH and the STRING databases share many similarities. Firstly, in both resources the network viewer is a visualization of interactions, accompanied by a confidence score calculated for each association. Particularly, STITCH depicts both protein-protein and protein-chemical associations as edges, while protein and chemical structures are represented as nodes with a slightly different shape. More detailed information can be retrieved from a popup window generated once the user clicks on a node or an edge of the graph, as well as from external resources such as PubChem [159], PDB [229] and SMART [217].

Similarly to STRING, the types of the interaction evidence are scored separately and divided into five different channels, which can be disabled individually or in combinations by the user. The *co-expression* and the *experiments* channels import experimentally validated, functional genomics data mainly from primary databases. In the co-expression channel, all the collected evidence is normalized, compared and scored based on the similarity of gene expression profiles, whereas in the experiments channel, the interaction records are assigned a score following the benchmarking against KEGG [53]. Furthermore, both *databases* and *text-mining* channels process established knowledge on protein-protein and protein-chemical associations, parsed from curated pathway databases and abstracts or texts from the scientific literature accordingly [232].

Moreover, in order to facilitate the investigation and interpretation of interactions, STITCH provides the same settings for the meaning of edges as STRING which include the *confidence*, the *evidence*, the *molecular action* and the *binding affinity* view (*Figure 12*). In the confidence view, the thickness of the edges indicates the confidence score of the association, whereas in the binding affinity view the width of the line is equivalent to the binding affinity between proteins and chemicals. Additionally, in the evidence view, each interaction is illustrated with a color that corresponds to the type of evidence that supports it, while the different colors of edges in the molecular action view are used to visualize the type of interaction between nodes (e.g. activation, inhibition or metabolization).

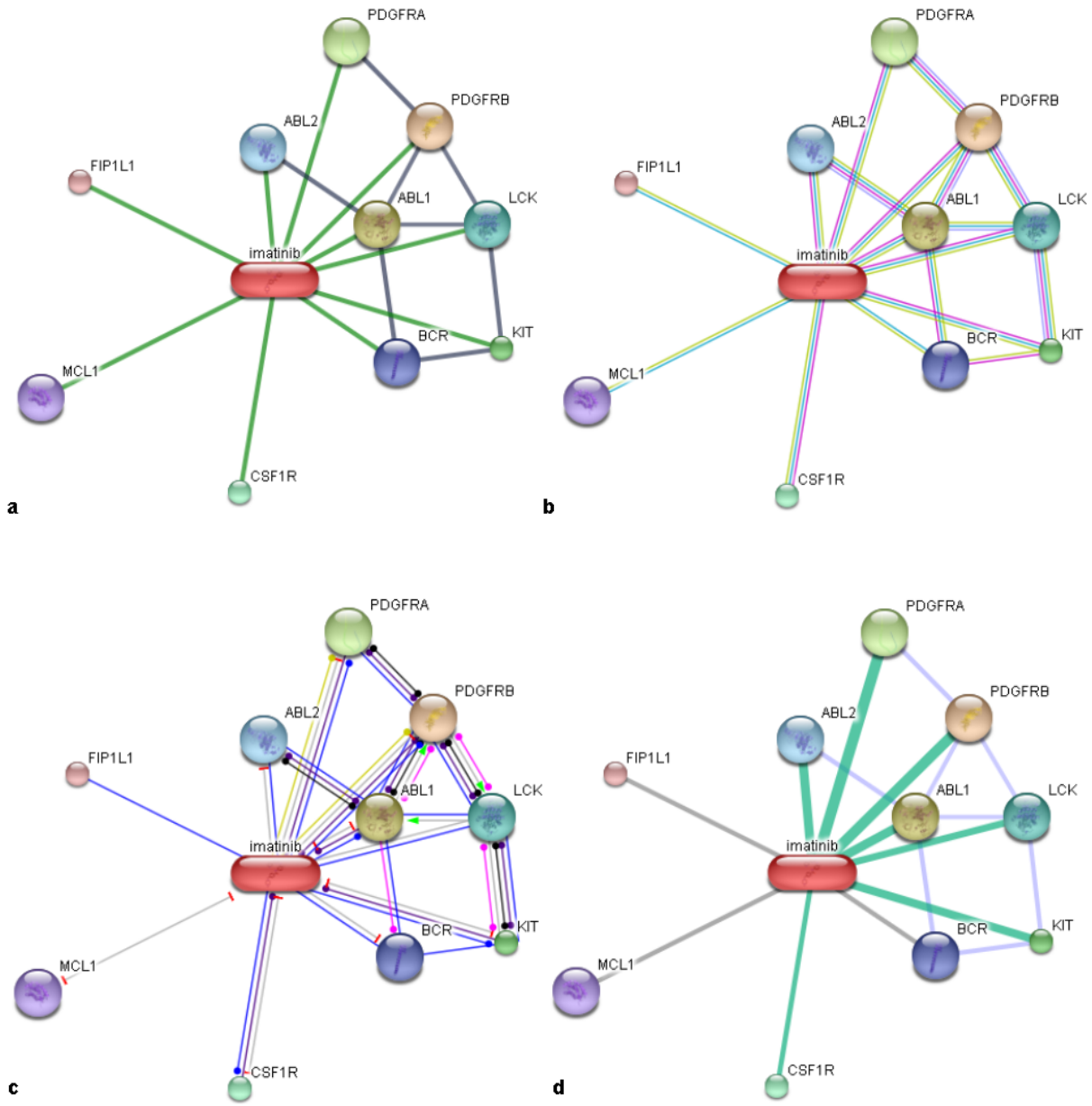


Figure 12: Meaning of network edges. (a) *confidence*: thickness of the edges is equivalent to the confidence score of the association. **(b)** *evidence*: the color of the edges indicates the type of the evidence. **(c)** *molecular action*: the shape of the edges illustrates the type of the molecular action between the connected nodes. **(d)** *binding affinity*: interactions between protein and chemicals that indicate their binding affinity.

CHAPTER 6: OnTheFly$^{2.0}$

6.1 What is OnTheFly$^{2.0}$

OnTheFly$^{2.0}$ is a web application aiming at providing a user-friendly and comprehensive environment to facilitate the knowledge integration, information extraction and visualization. It is the updated version of OnTheFly$^{1.0}$ [233] and it has been redeveloped to utilize state-of-art technologies and efficiently address various previously existing limitations. Firstly, the creation and designing of Graphical User Interface (GUI) was implemented with the use of R, Shiny, CSS, HTML and JavaScript technologies, instead of depending on a Java applet as its predecessor. Moreover, the replacement of commercial Windows-based converters with open-source, Unix-based ones, was another significant alteration, ameliorating the backend document format conversion and preservation of the original document layout. In addition to the aforementioned improvements, OnTheFly$^{2.0}$ is able to identify a broader spectrum of term types while it also supports 197 different organisms and OCR technology for processing multiple image files. Importantly, uploaded files are only stored temporarily in the OnTheFly$^{2.0}$ server just for parsing and no file backups, copies or personal data are kept. A more detailed comparison between OnTheFly$^{1.0}$ and OnTheFly$^{2.0}$ is presented in Table 1.

Functionality	OnTheFly$^{1.0}$	OnTheFly$^{2.0}$
Named Entity Recognition	genes, proteins, chemical compounds	genes/proteins, chemical compounds, organism names, environments, tissues, diseases, phenotypes, gene ontology terms
Supported files in textual formats	PDF (.pdf), Office texts (.doc), Flat text (.txt, .tsv, .csv)	PDF (.pdf) Microsoft Word (.doc and .docx) OpenOffice Writer(.odt) Microsoft Excel (.xls, .xlsm and .xlsx) OpenOffice Calc (.ods) Flat text (.txt, .tsv, .csv)
Image files with the use of OCR	N/A	Images (.bmp, .png, .jpg, .tif), PostScript (.ps, .eps) tesseract-ocr package
Interaction networks	STRING	STRING, STITCH
Infrastructure	Windows-based server	Unix-based (Linux) server and standalone package, compatible with Windows through WSL (Windows Subsystem for Linux)
File Converters	Commercial converters: ultra shareware, verypdf (PDF	Freely available, open-source converters: pdf2htmlEX, LibreOffice, ImageMagick (PDF, Text, spreadsheet & image layout preservation with minimal

	layout loss)	losses)
Tagging service	REFLECT	EXTRACT
GUI Implementation	Java Applet (obsolete)	R/Shiny, JavaScript
Functional Enrichment	Link to BioCompendium (outdated and Human/Mouse only)	In-house analysis based on g:Profiler and aGOtool. Enrichment terms include Gene Ontology, biological pathways, regulatory motifs, protein databases, human phenotype ontology etc.
Organisms	Human and Mouse	197 species
Combination of files and entity selection	N/A	Offered
Parameterization	N/A	Offered (e.g., functional enrichment options)
Browser Compatibility	Java Applets are not supported anymore	Safari, Tor, Firefox, Chrome, Edge, Opera

Table 1: Comparison of OnTheFly[1.0] and OnTheFly[2.0] .

The data pipeline of OnTheFly[2.0] consists of four interdependent steps: i) uploading of documents and simultaneous conversion from their original format to HTML, *ii)* implementation of EXTRACT services for the identification and extraction of bioentities from the input files, *iii)* functional and/or publication enrichment analysis on a created dataset of selected identifiers and *iv)* visualization and analysis of protein-protein and protein-chemical interaction networks. Each step is thoroughly described in the following subchapters.

6.2 Analysis pipelines

6.2.1 Text input and file conversion pipeline

Currently, OnTheFly[2.0] supports annotation for multiple different file formats, including PDF files (.pdf), Office-formatted documents (.doc, .docx, .ods, .odt), Spreadsheets (.xls, .xlsx), flat text files (.txt, .tsv, .csv), Rich Text Format (.rtf), images (.bmp, .jpg, .png, .tiff), as well as PostScript-compliant image file formats (.ps and .eps). In the online version, OnTheFly[2.0] provides the option to simultaneously upload multiple documents or paste/write a text in an input field and then process them separately or in combination. A maximum of 10 documents can be accommodated in each session with a file size that cannot exceed 10MBs, while image files are preferable to have a resolution / pixel density of at least 150 ppi/dpi. Notably, a GitHub repository (https://github.com/PavlopoulosLab/OnTheFly) is available enabling the download of the application in order to run locally.

The conversion of the uploaded files to HTML format is a prerequisite for the document annotation, thus OnTheFly[2.0] integrates a variety of tools and pipelines to efficiently cover a wide range of different file formats, while maintaining overall layout, text formatting, formulas and images to the extent possible. Hence, *pdf2htmlEX*, an open-source package [234] is used for the conversion of PDF files, whereas Office-formatted documents, Spreadsheets, flat text files, Rich Text Format and PostScript are converted with the *LibreOffice* universal converter (*u*noconv). Importantly, unoconv is able to separately convert and handle each sheet of a spreadsheet file. In addition, Optical Character Recognition (OCR) scan is implemented in case of image files with no text encoding. Both the open-source package *ImageMagick* and the *tesseract-ocr* package [235] are utilized for file format conversion and OCR scanning, respectively. After the OCR scanning has been completed, PDF files with parseable text are produced, which are then processed as previously described. The quality and resolution of the imported images can determine the overall outcome of the OCR scanning, therefore images containing text elements in rotated orientation or embedded in complex graphical shapes, or images with low resolution may result in poor OCR results.

Once the uploading and conversion of files have been completed, a checkbox list will appear, containing all the uploaded files and/or submitted texts, while deletion and renaming of one or multiple files are also allowed. Any additional uploaded/created documents are appended to the selection list. The resulting HTML version of the selected file(s) can be inspected in a reactive tab panel that displays each choice in a separate tab. By selecting or deselecting a file from the checkbox list, the corresponding tab will be dynamically inserted or removed accordingly. This allows the user to identify conversion or OCR problems before

continuing the analysis. An overview of the annotation file conversion pipeline is shown in *Figure 13*.

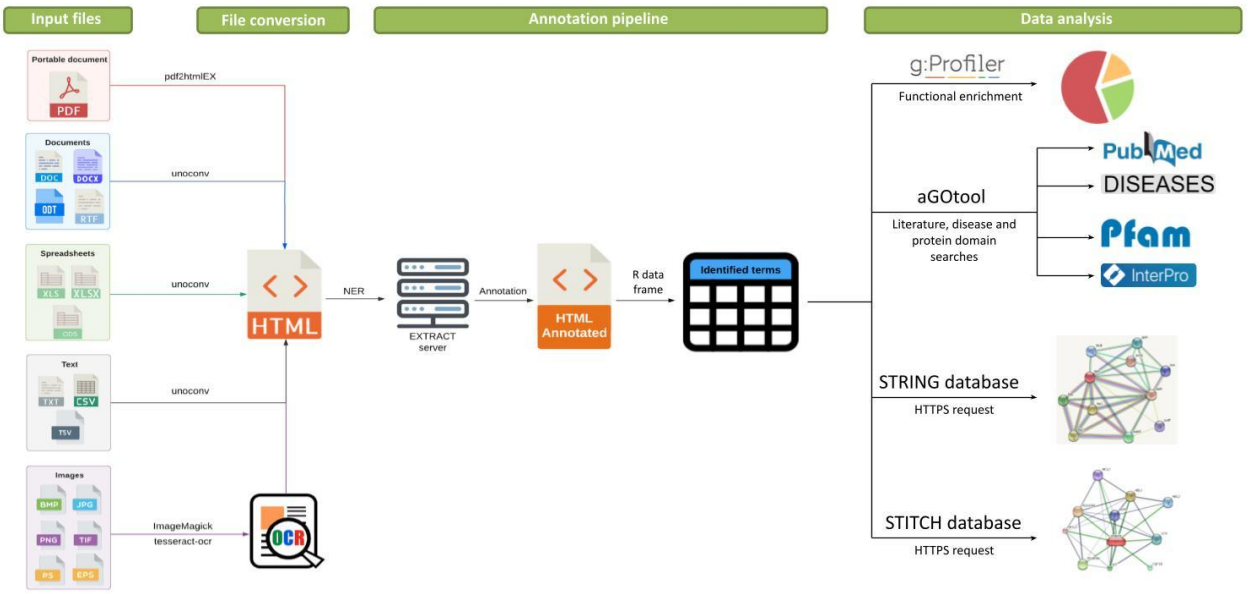


Figure 13: Flowchart of the OnTheFly[2.0] backend pipeline for file conversion, named entity recognition and data analysis.

6.2.2 Document annotation using Named Entity Recognition (NER)

6.2.2.1 Annotation parameters

OnTheFly[2.0] implements the EXTRACT tagging service [30] to perform dictionary-based Named Entity Recognition (NER) and biological annotation of the uploaded documents. Specifically, by integrating the tagger software [236], EXTRACT efficiently identifies 14 different entity types, including environment descriptive terms from Environment Ontology (e.g., desert, forest) [237], organism mentions from NCBI Taxonomy [238], tissue terms from BRENDA Tissue Ontology [64], disease mentions from Disease Ontology [137], phenotypes from Mammalian Phenotype Ontology [181], biological processes, cellular components, molecular functions from Gene Ontology [100,239], small chemical molecules from PubChem [159], non-coding RNAs from RAIN [240], and protein-coding genes from STRING [148]. Currently, OnTheFly[2.0] supports a set of 197 organisms, a more detailed list of which is available in a separate tab, containing the g:Profiler ID and KEGG code in addition to the species name, common name and the taxonomy ID. Notably, OnTheFly[2.0] gives the opportunity to select one or more entity types for which NER can be performed, whereas in case of protein identification, the user has to choose one organism, whose proteins will be detected in the text.

6.2.2.2 Annotation results

Upon setting the annotation parameters, based on which NER process is accomplished, the selected document will be tagged with all of the recognized terms linked and highlighted using different colors. A legend which assigns each entity term category to a specific color as well as a table with the parameters used during the annotation of the document are available. By hovering the mouse cursor over highlighted terms, OnTheFly[2.0] will generate a popup window that matches each word to the corresponding type, name and identifier and provides concise information about the particular biomedical entity, enriched with links to external databases. In case of term disambiguation (e.g., when a term comes from several organisms or corresponds to more than one entity type), all of the possible options are reported.

For further analysis and in-depth view of the identified bioentities, an interactive table is provided, containing the name and entity type of all the extracted terms accompanied by their database identifiers as hyperlinks. The identifiers for each term are retrieved from various databases such as ENSEMBL [83] for proteins and genes, NCBI PubChem [159] for chemical compounds, NCBI Taxonomy browser [238] for organisms, EMBL-EBI's QuickGO [241] browser for Gene Ontology terms, BRENDA [64] for tissues, Disease Ontology (DOID) [242] for diseases and finally EMBL-EBI Phenotype Ontology [242] for phenotypes. The table results can be filtered by entity type at any stage, while the entire table, as well as filtered results, can be exported as a CSV file. Similarly to the graphical view of the text, a table with the text-mining parameters used for the annotation is provided. The overall process of document annotation is presented in *Figure 14*.

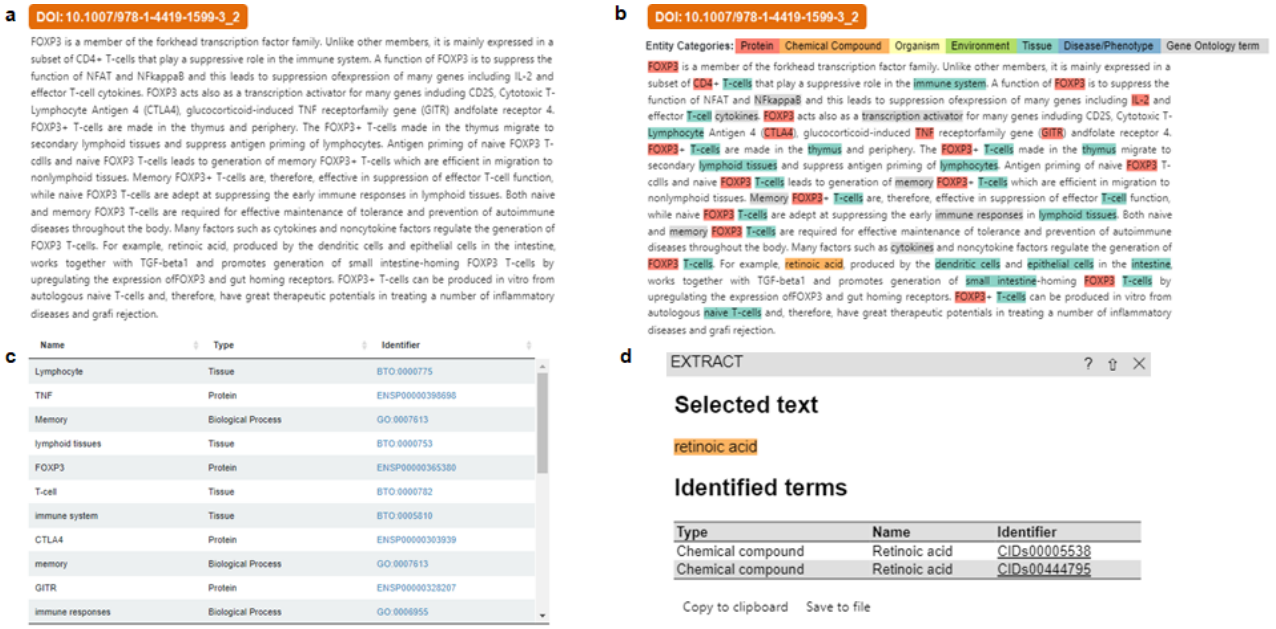


Figure 14: Document annotation using NER for article by Chang H Kim, 2009 [243]. **(a)** Abstract file in its initial form prior to annotation. **(b)** Annotated abstract using the *H. sapiens* as organism and all the available entity types. **(c)** A summary of the Interactive table containing the extracted entities identified in the abstract. **(d)** Example of the popup window with information about a specific identified term. The term is colored according to its type and original links to external databases are provided.

6.2.3 Dataset creation

OnTheFly[2.0] enables the creation of a customizable dataset consisting of extracted proteins and chemical compounds identified during the annotation process in one or multiple documents (*Figure 15a*). All the selected entities are collected and displayed in a reactive table with four different columns: the identifier, the type (e.g., gene, protein), the name of the term as well as the name of the *document* it originated from (*Figure 15b*). The table can be narrowed down after filtering by entity type and can also be downloaded in CSV, Excel or PDF format. In addition, the user is able to delete the entire dataset or single entities at will. Once submitted, further analysis can be performed in the resulting dataset, including Functional Enrichment Analysis using the g:Profiler, Literature Search with aGOtool, Protein Domain Search/Literature search which maps the selected proteins against the scientific literature and Protein-Protein Network or Protein-Chemical Network (*Figure 15b*). Each analysis method is thoroughly described in the next paragraphs.

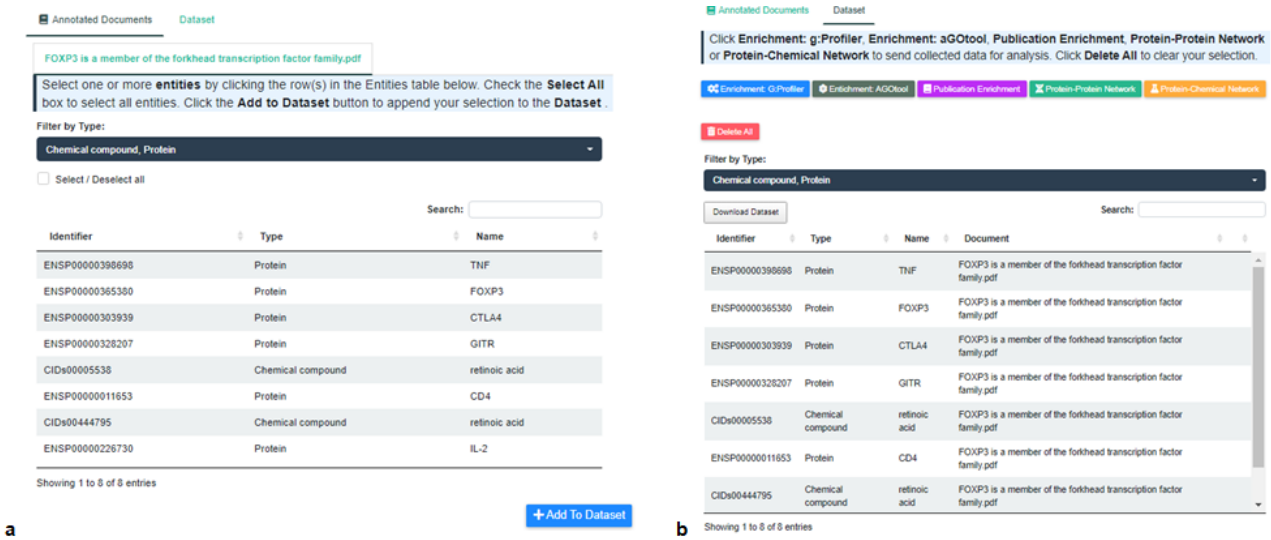


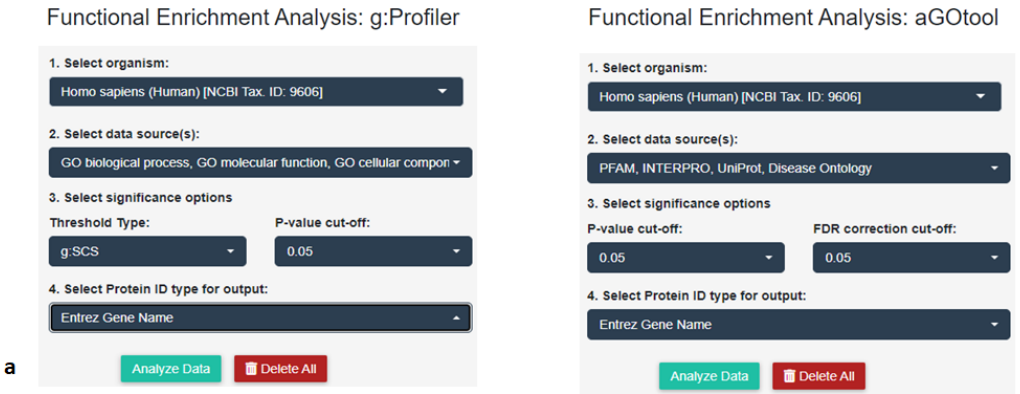
Figure 15: Creation of dataset for further analysis. **(a)** Summary of interactive datatable which displays all the proteins and chemical compounds identified in the uploaded files. **(b)** Resulting dataset depicted in the form of a table with four columns. The user is able to choose up to five different functionalities (functional enrichment analysis, literature search, protein domain search/literature search and Protein-Protein Network or Protein-Chemical Network) for further data analysis of the selected entities.

6.2.4 Functional enrichment analysis

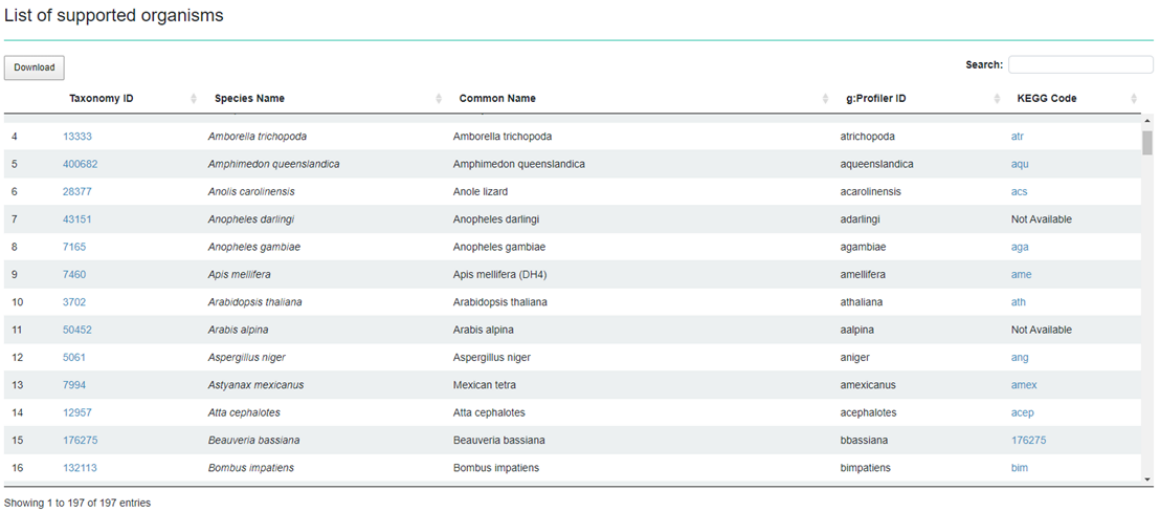
6.2.4.1 Input and functional enrichment parameters

OnTheFly[2.0] integrates two tools, *g:Profiler* [50,51] and *aGOtool* [41], to perform rich functional enrichment analysis for a selected dataset of genes/proteins orginited from one or multiple annotated documents. Several parameters can be customized in order for the enrichment analysis to be performed, including the selection of an organism among a list of 197 species (*Figure 16b*), the selection of data sources and protein ID as well as the setting of threshold type and cut-off value. More specifically, OnTheFly[2.0] exploits g:Profiler services to identify enriched functional terms from Gene Ontology [100,239], pathways from KEGG [53], Reactome [54] and WikiPathways [55], protein complexes from CORUM [56], expression data from Human Protein Atlas [244], regulatory motifs from TRANSFAC [57] and miRTarBase [58], and phenotypes from the Human Phenotype Ontology [245] (*Figure 16a*). Further enrichment analyses can be performed by the aGOtool, which is utilized for the identification of enriched terms from the UniProt keyword classification system, protein families and domains from Pfam [61] and InterPro [216], as well as human diseases from the DISEASES database [236] (*Figure 16a*).

g:Profiler and aGOtool test for statistically significant enrichment by using Fisher's exact test to compare the user's input dataset (foreground) to a background set from organism-specific genes annotated in the Ensembl database [83] and UniProt Reference Proteomes [60], respectively. As far as the g:Profiler is concerned, the resulting p-values are corrected for multiple testing using either g:SCS, Bonferroni correction or Benjamini-Hochberg false discovery rate (FDR), whereas in the case of aGOtool, p-values are corrected using Bonferroni correction or FDR. Enrichment analysis can also be performed as previously mentioned using ENSEMBL IDs as input, while results can be reported as Entrez, UniProt [60], EMBL [246], ENSEMBL [83] and RefSeq gene/protein names/identifiers, based on the user's selection criteria.



a



b

Figure 16: Customization of functional enrichment parameters. **(a)** Input parameters for functional enrichment analysis by g:Profiler and aGOtool. **(b)** A summary of the table containing the list of the 197 organisms available for analysis.

6.2.4.2 Functional enrichment results

Firstly, functional enrichment results are displayed in interactive searchable tables, organized by source (e.g., KEGG, Reactome, CORUM, etc.) and composed of seven distinct columns, providing comprehensive data about each functional term. The columns include the *term ID*, which is a hyperlink that points to the corresponding data source of the term, the *term name* that briefly describes the function, the hypergeometric *p-value*, resulting upon correction for multiple testing, the *term size* and *query size*, referring to the number of genes that are annotated to the term and included in the query respectively, the *no. of positive hits* which is the number of genes in the input query that are annotated to the corresponding term and the *positive hits* which constitute the last, hidden column of the result table and correspond to the identified genes/proteins from the query that found to be associated with the functional term. One can expand each row of the table to observe the positive hits.

In addition, an interactive bar blot is available for the visualization of enrichment results retrieved from both g:Profiler and aGOtool analysis. The x-axis in the plot denotes the enrichment metric function, while the y-axis represents the term name. Importantly, the user is able to adjust certain control parameters, including the *database*, the *enrichment metric* and *the number of terms*, in order to change the bar plot contents at will. Particularly, the database control option allows the selection of database(s) to plot, coloring each database type differently. The two options of enrichment metric control, which can change the bar lengths, are either -log10(P-value) or an enrichment score, defined as the % ratio of observed over expected terms, whereas the plot height can alter by adapting the number of terms that appear in the plot. An interactive and responsive datatable containing all the terms and information displayed in the plot, according to the control parameters, accompanies the graphical visualization and enables the download in various formats.

Lastly, only in case of g:Profiler, an interactive Manhattan plot is offered as well to supplement the visualization options. In this plot, the x-axis represents the color-coded, grouped functional terms, while the y-axis shows the significance (p-value) of each term. In order for the user to acquire a better overview, hovering over a data point reveals a tooltip with key information about the particular functional term. Finally, the most significant functional terms are shown as a bar chart, which the user can customize to show the desired number of terms. Notably, all of the aforementioned reports and visualization can be exported and saved in various file formats (CSV, XLS, PDF).

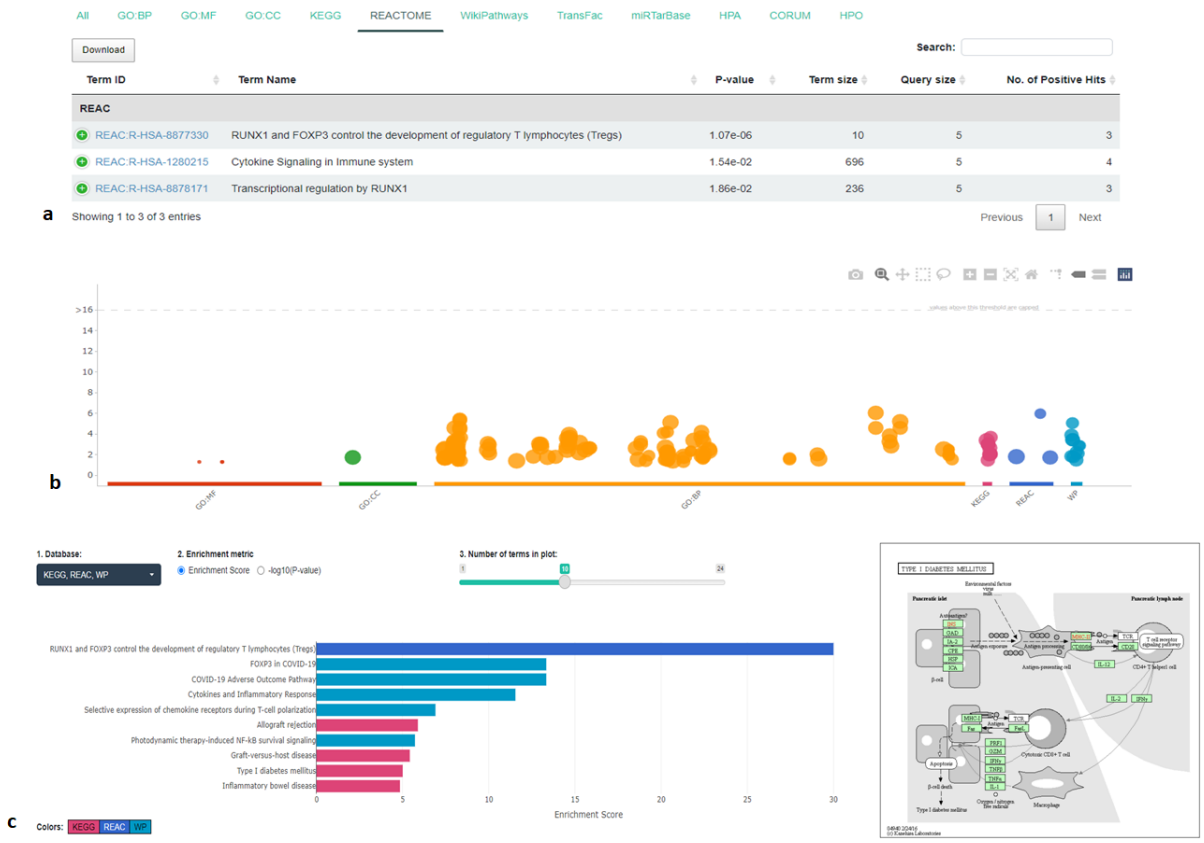


Figure 17: Functional enrichment analysis results. **(a)** Summary table with the functional terms and the corresponding identified entities. Results from REACTOME are shown. **(b)** Functional enrichment overview with the use of a Manhattan plot. **(c)** Bar plot representation of enriched genes distributed into metabolic pathways obtained from KEGG, Reactome and WikiPathways. The results of each database are color coded, while the bar length is proportional to the extent of enrichment for each term, as represented by the enrichment score value. Additionally, A KEGG pathway is shown which includes the genes identified in the document.

6.2.4 Literature enrichment analysis

In addition to functional enrichment analysis, OnTheFly$^{2.0}$ implements the aGOtool API to perform literature enrichment analysis for a gene/protein list extracted from the uploaded input files. As its name indicates, literature enrichment analysis is oriented towards the facilitation of retrieval of scientific publications that are strongly relevant to a given gene/protein list. Particularly, all PubMed abstracts and full-text articles from the PubMed Central Open Access subset are parsed by the same NER tagger used in EXTRACT and the results are updated with new documents on a weekly basis. Consequently, genes existing within the previously mentioned text corpora are automatically annotated and retrieved, generating millions of gene sets that are subsequently used by aGOtool in the same manner as all other gene sets.

The publication enrichment functionality in OnTheFly[2.0] can be performed on a list of 197 organisms from which the user is able to select up to 1,000 genes and proteins identified in the input files. The created list is subsequently submitted to aGOtool that tests each document from the precomputed corpus for statistically significant enrichment, again using Fisher's exact test. The resulting p-values as well as Bonferroni-corrected p-values and Benjamini-Hochberg FDR values can be used for filtering the results. In addition, the protein ID type (e.g., Entrez, ENSEMBL, Uniprot), that will be used in the analysis and the output is defined by the user.

Similarly to the functional enrichment analysis, results are reported in interactive searchable tables containing detailed information about each literature term (scientific publication or disease), such as the *term ID*, the *term name*, the *P-value*, the *FDR*, the *term* and *query size*, the *number of positive hits* as well as the *positive hits* themselves. Links are provided for publications and diseases to PubMed. Barchart plot visualization is also provided, enabling the ranking of publications based on their significance and the filtering of results by number of displayed reports. Both the table and the plot can be downloaded and saved in various file formats (CSV, XLS, PDF).

6.2.5 Interaction network analysis

So as to supplement the aforementioned enrichment functionalities and acquire a deeper understanding of the overall associations of the extracted bioentities, the visualization of interaction networks becomes almost a necessity. Thus, OnTheFly[2.0] by integrating the APIs of the STRING [148] and STITCH [164] databases, offers the capability to generate and depict protein-protein and protein-chemical interactions for a set of 197 organisms, respectively. Once the dataset creation is completed and submitted, the user is able to retrieve the associations between the extracted biomolecules and visualize the results as networks with the entities presented as nodes and their interactions as edges. For computational efficiency reasons, in its current version, OnTheFly[2.0] allows a maximum of 500 proteins per request for STRING and 100 proteins or small molecules per request for STITCH.

Following the same logic as the functionalities described previously, OnTheFly[2.0] requires the customization of certain setting options in order to construct the network as efficiently as possible, meeting the visualization needs of the user (*Figure 18*). Specifically, both STRING and STITCH classify the interactions as either *physical*, meaning that interacted entities constitute the same biomolecular complex, or *functional*, which refer to the entities involved in the same pathway/process. Consequently, the options include the *Full* set of interactions (both physical and functional) or the *Physical subnetwork* exclusively. In addition, users can adjust the *interaction score* and apply a cutoff on the edges. Regardless

of the type of interactions supported by the networks (functional or physical), the meaning of the edges can also alter, reflecting the evidence on which the interactions depend upon (*Evidence mode*), the interaction score (*Confidence mode*), the type and effect of each protein-chemical interaction (*Molecular Action*) as well as the the binding affinity between the proteins and bound chemicals (*Binding affinity*). In case of the confidence and binding affinity modes, the thickness of the edges is equivalent to a stronger interaction, whereas the different colors in the evidence and molecular action modes indicate the different types of interactions.

The resulting network, generated based on the visualization criteria customized by the user, is displayed in a separate Network Viewer panel. The interactivity as well as the characteristic STRING and STITCH network layout and style is preserved. An example of such networks is shown in *Figure 18*. One can export the displayed network as an image or as a TSV file, containing all the biomolecular interactions, while the redirection to the STRING or STITCH database for further analysis is also allowed.

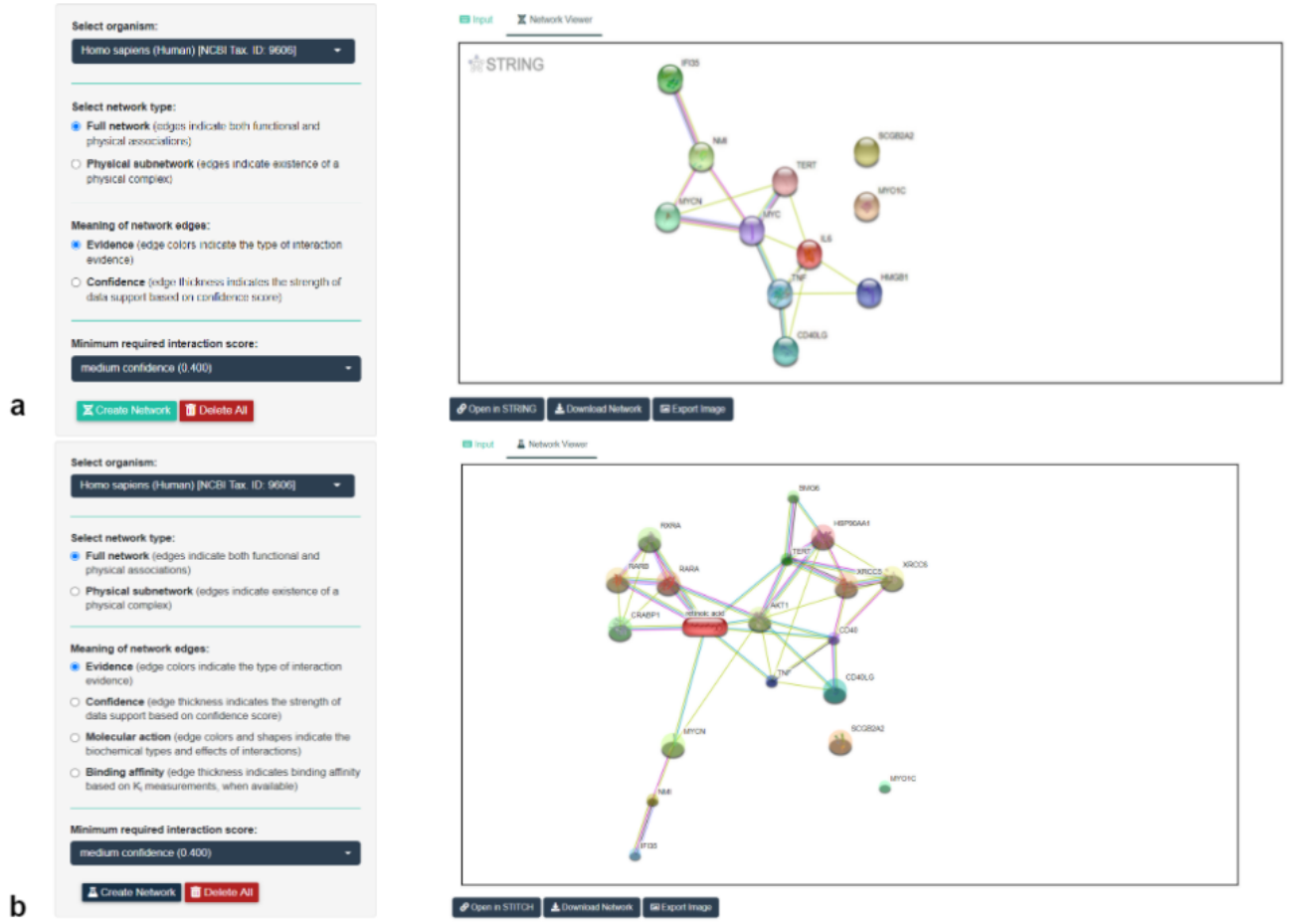


Figure 18: Visualization of interaction networks. **(a)** Setting options and visualization of a protein-protein interaction network with the STRING layout. **(b)** Setting options and visualization of a protein-chemical interaction network with the STITCH layout.

6.3 Implementation

OnTheFly[2.0] is a web application implemented in R and JavaScript, while the R/Shiny package, HTML and CSS are utilized for the creation of the Graphical User Interface (GUI). Both Shiny and ShinyJS serve as interoperable packages, establishing the connection between the R and JavaScript functions. In addition, several APIs are integrated to increase the versatility of this tool. Specifically, the API of the EXTRACT web service is used for the performance of NER via the *tagger* text mining utility, whereas the functional enrichment analyses are offered by R/gprofiler2 library and aGOtool API. Biological networks are constructed and visualized using the STRING API, as implemented in the STRING and STITCH databases. Various R libraries are also used for the creation of the interactive visualizations. Manhattan plots are generated with the implementation of plotly library [247], bar plots with ggplot [248] and the interactive datatables through the DT library.

OnTheFly[2.0] is available as a web service, and as a standalone package through a GitHub repository. The web service is fully functional in all major web browsers (Google Chrome, Mozilla Firefox, Microsoft Edge, Tor, Apple Safari, Opera). Linux and other Unix-based operating systems are the native environment OnTheFly[2.0] is designed to operate in. The existence of a Windows Subsystem for Linux (WSL) or other similar compatibility layers (e.g., Cygwin) is a prerequisite in order OnTheFly[2.0] run on Windows.

6.4 Case study

Six published meta-analysis reports on clinical biomarkers of severe COVID-19 were analysed [249–254], so as to demonstrate the capabilities of OnTheFly2.0 in a real case study and denote its functionality in extracting biological information. Briefly, COVID-19 is a newly emerged disease caused by the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2). Typically, pathogenesis of SARS-CoV-2 infection involves the entry of the virus into host cells through the ACE2 receptor and the release of its genome in order to accomplish viral replication [255]. This attack results in the upregulation of the immune system in an attempt to eliminate the virus from the body. However, the failure in downregulating this response ultimately induces a hyperinflammatory stage of COVID-19 called *cytokine storm* [255], which is characterized by an excessive production and secretion of pro-inflammatory cytokines and chemokines, promoting uncontrolled systemic inflammation. While the majority of coronaviruses affect the respiratory tract cells predominantly, SARS-CoV-2 virus can potentially lead to mutli-organ failure due to its ability to enter numerous mammalian body tissues [256]. Even though signs of COVID-19 vary according to the patient, the most common symptoms appearing at the earlier stage of the disease include fever, fatigue, dry cough, myalgia, anorexia and dyspnea [255].

Starting from the uploading and conversion of the articles from PDF to HTML format, the annotation pipeline was performed with the use of NER. Extracted bioentities were filtered in order to manually discard the false positives and were subsequently processed for both functional and publication enrichment analysis. As might be expected, the most significantly enriched diseases included “Respiratory failure”, “Pneumonia” and “COVID-19”. Specifically, several GO terms retrieved from GO enrichment for biological processes were found to participate in inflammation, cell activation and response to stress, indicating a potential relation to exaggerated lung inflammation and systemic immune dysfunction, frequently appearing symptoms of COVID-19 as previously mentioned. Similarly, the extracted terms were also significantly enriched for molecular functions involved in cytokine activity and cytokine receptor signaling. In addition, Uniprot keyword analysis was implemented, denoting “Cytokine”, “Inflammatory response”, “Host-virus interaction”, and “Host cell receptor for virus entry” terms to all be enriched, in line with the functional enrichment results.

Furthemore, based on established knowledge on coronaviruses, it was not surprising that extracellular space (GO:0005615, GO:0005576) and plasma membrane (GO:0009897, GO:0009986, GO:0098552) predicted to be extracellular components associated with biomarkers of severe COVID-19. Interestingly though, membrane microdomains (also called “membrane rafts”; GO:0098857, GO:0045121), known for mediating the initial binding of

SARS-CoV-2 to ACE2 receptor, virus internalization and cell-to-cell transmission [255], also appeared as essential cellular components in the pathogenesis of SARS-CoV-2, without prior reference in any of the six meta-analysis reports that were interrogated, unveiling the ability of OnTheFly$^{2.0}$ for rapid knowledge discovery.

The STRING option of OnTheFly$^{2.0}$ enabled the acquisition of a visual representation and discovery of both physical and functional protein-protein interactions. The analysis revealed a cluster of interacting cytokines and other immune components that is pertinent to the “cytokine storm” of severe COVID-19 (*Figure 19*). Notably, the results were further supported by a publication enrichment analysis, as the majority of studies about COVID-19 included cytokines references.

In order to acquire a more holistic view, identified terms were mapped onto biological pathways via the pathway enrichment analysis option of OnTheFly$^{2.0}$. Firstly, KEGG pathways were extracted, including “coronavirus disease - COVID-19” (KEGG: 05171), “viral protein interaction with cytokine and cytokine receptor” (KEGG: 04061) and “cytokine-cytokine receptor interaction” (KEGG: 04060). Intriguingly, “Yersinia infection” (KEGG: 05135) was also identified as a relevant KEGG pathway with high probability (p-value < 10-8). It is of note that Yersinia pestis, a Gram-negative bacterium which is the etiological pathogen of plague [257], activates the inflammasome-mediated IL-1β/IL-18 cytokine release [257] of pneumocytes and alveolar macrophages during pneumonic plague. Even though inflammasome activation has a protective role against a number of pathogens, in case of pneumonic plague, inflammasome contributes to neutrophil influx and exaggerated inflammation that ultimately results in lung tissue damage [257]. The similarity of immune responses and lung tissue reactions between Yersinia pestis and severe SARS-CoV-2 infection [258] warrant further insights into the immunological mechanisms of response to these unrelated pathogens. Of additional interest is the predicted involvement of the “IL-17 signaling pathway” (KEGG: 04657) in severe COVID-19 which is supported by a recent study reporting T cell skewing towards Th17, a specialized CD4+ effector T cell lineage characterized by secretion of IL-17 and IL-17F cytokines in patients with COVID-19 pneumonia [259].

In addition to KEGG, the REACTOME option of OnTheFly$^{2.0}$ was also explored in order to further detect the over-represented pathways, resulting in significant enrichment of several cytokine pathways associated with biomarkers of severe COVID-19. Similar to membrane microdomains mentioned previously, the “cellular senescence” pathway was predicted to be significantly enriched despite the absence of specific references to this biological term in any of the six annotated meta-analysis reports under study. In line with this prediction, COVID-19 pneumonia has recently been associated with immunosenescence [259] and accelerated aging of pneumocytes [260]. Overall, the aforementioned case study analysis

highlights the practical utility OnTheFly[2.0] and its capability to rapidly extract biological information from multiple documents and hence assisting knowledge discovery (*Figure 19*).

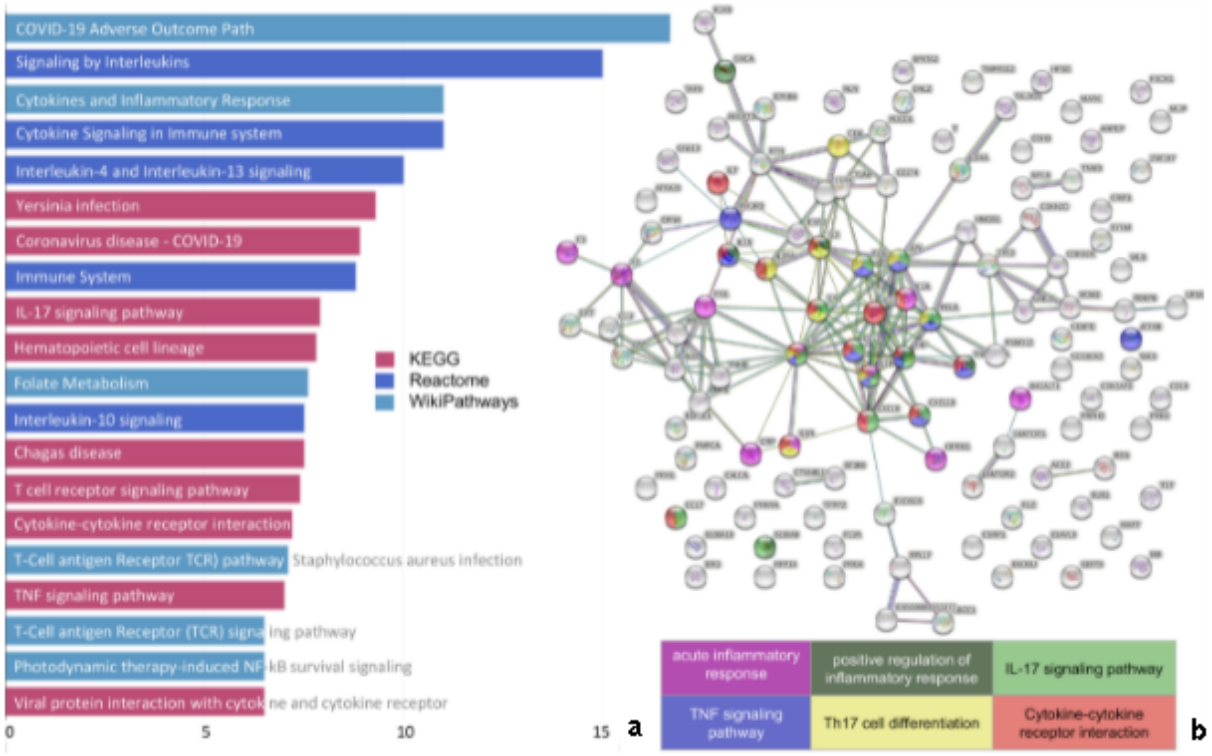


Figure 19: Analysis of clinical biomarkers of severe COVID-19 using OnTheFly[2.0]. **(a)** List of enriched pathways from the KEGG, Reactome and WikiPathways databases. **(b)** Analysis of putative protein-protein interactions through the STRING option of OnTheFly[2.0]. A cluster of interacting components of inflammatory/immune pathways, each represented by a different color.

Conclusions

OnTheFly$^{2.0}$ is a powerful and user-friendly web based tool which enables the identification of various biological terms in locally stored documents found in a plethora of different file formats, including PDFs, texts, Office and image files. The document annotation is accomplished according to the selection of entity type(s) by users such as proteins, genes, chemical compounds, organisms, tissues, environments, diseases, phenotypes and gene ontologies, while popup windows with informative summaries about a term and its links to external repositories are also generated. Furthermore, OnTheFly$^{2.0}$ efficiently combines multiple functionalities, covering a broad spectrum of analyses from functional and publication enrichment to protein-protein and protein-chemical interaction network visualization. These analyses are performed in a customizable dataset which is composed of selected bioentities extracted from the uploaded documents. Notably, OnTheFly$^{2.0}$ is designed to facilitate the annotation of locally stored documents and further exploration and analysis of their identified biomedical entities in a fully automated way. Considering the reliability, ease of use and accuracy of the offered capabilities, OnTheFly$^{2.0}$ can reach out to many users varying from experimentalists to highly specialized bioinformaticians.

Availability

OnTheFly[2.0] application:

http://onthefly.pavlopouloslab.info or http://bib.fleming.gr:3838/OnTheFly/

OnTheFly[2.0] source code and instructions:

https://github.com/PavlopoulosLab/OnTheFly

Publications

- Baltoumas, F.A., Zafeiropoulou, S., Karatzas, E., Koutrouli, M., Thanati, F., Voutsadaki, K., Gkonta, M., Hotova, J., Kasionis, I., Hatzis, P., Pavlopoulos, G.A., 2021. **Biomolecule and Bioentity Interaction Databases in Systems Biology: A Comprehensive Review**. Biomolecules 11, 1245. https://doi.org/10.3390/biom11081245

- Baltoumas, F.A., Zafeiropoulou, S., Karatzas, E., Paragkamian, S., Thanati, F., Iliopoulos, I., Eliopoulos, A.G., Schneider, R., Jensen, L.J., Pafilis, E., Pavlopoulos, G.A., 2021. **OnTheFly2.0: a text-mining web application for automated biomedical entity recognition, document annotation, network and functional enrichment analysis**. https://doi.org/10.1101/2021.05.14.444150

References

1. Simon, C. *et al.* BioReader: a text mining tool for performing classification of biomedical literature. *BMC Bioinformatics* **19**, 57 (2019).
2. Krallinger, M., Valencia, A. & Hirschman, L. Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biol.* **9**, S8 (2008).
3. BioTextQuest+: a knowledge integration platform for literature mining and concept discovery - PubMed. https://pubmed.ncbi.nlm.nih.gov/25673338/.
4. Application of Biomedical Text Mining | IntechOpen. https://www.intechopen.com/books/artificial-intelligence-emerging-trends-and-applications/application-of-biomedical-text-mining.
5. Encyclopedia of Bioinformatics and Computational Biology | ScienceDirect. https://www.sciencedirect.com/referencework/9780128114322/encyclopedia-of-bioinformatics-and-computational-biology.
6. Singh, S. Natural Language Processing for Information Extraction. *ArXiv180702383 Cs* (2018).
7. Ghoulam, A., Barigou, F. & Belalem, G. Information Extraction in the Medical Domain. *J. Inf. Technol. Res.* **8**, 1–15 (2015).
8. Adnan, K. & Akbar, R. Limitations of information extraction methods and techniques for heterogeneous unstructured big data. *Int. J. Eng. Bus. Manag.* **11**, 1847979019890771 (2019).
9. Perera, N., Dehmer, M. & Emmert-Streib, F. Named Entity Recognition and Relation Detection for Biomedical Information Extraction. *Front. Cell Dev. Biol.* **8**, (2020).
10. Eltyeb, S. & Salim, N. Chemical named entities recognition: a review on approaches and applications. *J. Cheminformatics* **6**, 17 (2014).
11. Song, H.-J., Jo, B.-C., Park, C.-Y., Kim, J.-D. & Kim, Y.-S. Comparison of named entity recognition methodologies in biomedical documents. *Biomed. Eng. OnLine* **17**, 158 (2018).
12. Mansouri, A., Affendey, L. & Mamat, A. Named Entity Recognition Approaches. *Int J Comp Sci Netw Sec* **8**, (2008).
13. Song, M., Yu, H. & Han, W.-S. Developing a hybrid dictionary-based bio-entity recognition technique. *BMC Med. Inform. Decis. Mak.* **15 Suppl 1**, S9 (2015).
14. Goecks, J., Jalili, V., Heiser, L. M. & Gray, J. W. How Machine Learning Will Transform Biomedicine. *Cell* **181**, 92–101 (2020).
15. Goyal, A., Gupta, V. & Kumar, M. Recent Named Entity Recognition and Classification techniques: A systematic review. *Comput. Sci. Rev.* **29**, 21–43 (2018).
16. Wang, Y. *et al.* Supervised methods for symptom name recognition in free-text clinical records of traditional Chinese medicine: An empirical study. *J. Biomed. Inform.* **47**, 91–104 (2014).
17. Saha, S. K., Narayan, S., Sarkar, S. & Mitra, P. A composite kernel for named entity recognition. *Pattern Recognit. Lett.* **31**, 1591–1597 (2010).
18. Introduction to Bootstrapping in Statistics with an Example. *Statistics By Jim* http://statisticsbyjim.com/hypothesis-testing/bootstrapping/ (2018).
19. Heo, G. E., Xie, Q., Song, M. & Lee, J.-H. Combining entity co-occurrence with specialized word embeddings to measure entity relation in Alzheimer's disease. *BMC Med. Inform. Decis. Mak.* **19**, (2019).
20. Buzydlowski, J. W. Co-occurrence analysis as a framework for data mining. **6**, 19 (2015).
21. Bourgeois, N., Cottrell, M., Lamasse, S. & Olteanu, M. Search for Meaning Through the Study of Co-occurrences in Texts. in *International Work-Conference on Artificial Neural Networks* (2015).
22. Lou, W. & Qiu, J. Semantic information retrieval research based on co-occurrence analysis. *Online Inf. Rev.* **38**, (2014).
23. Freilich, S. *et al.* The large-scale organization of the bacterial network of ecological co-occurrence interactions. *Nucleic Acids Res.* **38**, 3857–3868 (2010).
24. Pavlopoulos, G. A., Promponas, V. J., Ouzounis, C. A. & Iliopoulos, I. Biological Information Extraction and Co-occurrence Analysis. in *Biomedical Literature Mining* (eds. Kumar, V. D. & Tipney, H. J.) vol. 1159 77–92 (Springer New York, 2014).
25. Zhang, X. *et al.* Targeting of mutant p53-induced FoxM1 with thiostrepton induces cytotoxicity and

enhances carboplatin sensitivity in cancer cells. *Oncotarget* **5**, 11365–11380 (2014).
26. Speech and Language Processing. https://web.stanford.edu/~jurafsky/slp3/.
27. Qaiser, S. & Ali, R. Text Mining: Use of TF-IDF to Examine the Relevance of Words to Documents. *Int. J. Comput. Appl.* **181**, (2018).
28. Zhang, Y., Zhou, Y. & Yao, J. Feature Extraction with TF-IDF and Game-Theoretic Shadowed Sets. in *Information Processing and Management of Uncertainty in Knowledge-Based Systems* (eds. Lesot, M.-J. et al.) 722–733 (Springer International Publishing, 2020). doi:10.1007/978-3-030-50146-4_53.
29. Jalilifard, A., Caridá, V. F., Mansano, A. F., Cristo, R. S. & da Fonseca, F. P. C. Semantic Sensitive TF-IDF to Determine Word Relevance in Documents. *ArXiv200109896 Cs Stat* **735**, (2021).
30. EXTRACT 2.0: text-mining-assisted interactive annotation of biomedical named entities and ontology terms | bioRxiv. https://www.biorxiv.org/content/10.1101/111088v1.
31. Wei, C.-H., Allot, A., Leaman, R. & Lu, Z. PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Res.* **47**, W587–W593 (2019).
32. Weber, L. *et al.* HunFlair: An Easy-to-Use Tool for State-of-the-Art Biomedical Named Entity Recognition. *Bioinforma. Oxf. Engl.* (2021) doi:10.1093/bioinformatics/btab042.
33. Allot, A. *et al.* LitVar: a semantic search engine for linking genomic variant data in PubMed and PMC. *Nucleic Acids Res.* **46**, W530–W536 (2018).
34. Sherry, S. T., Ward, M. & Sirotkin, K. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. *Genome Res.* **9**, 677–679 (1999).
35. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res.* **46**, D1062–D1067 (2018).
36. Wei, C.-H. *et al.* tmVar 2.0: integrating genomic variant information from literature with dbSNP and ClinVar for precision medicine. *Bioinforma. Oxf. Engl.* **34**, 80–87 (2018).
37. Wang, J. Functional Enrichment Analysis. in *Encyclopedia of Systems Biology* (eds. Dubitzky, W., Wolkenhauer, O., Cho, K.-H. & Yokota, H.) 772–772 (Springer, 2013). doi:10.1007/978-1-4419-9863-7_491.
38. Cao, J. & Zhang, S. A Bayesian Extension of the Hypergeometric Test for Functional Enrichment Analysis. *Biometrics* **70**, 84–94 (2014).
39. Rivals, I., Personnaz, L., Taing, L. & Potier, M.-C. Enrichment or depletion of a GO category within a class of genes: which test? *Bioinformatics* **23**, 401–407 (2007).
40. Yang, D. *et al.* Gaining confidence in biological interpretation of the microarray data: the functional consistence of the significant GO categories. *Bioinformatics* **24**, 265–271 (2008).
41. Schölz, C. *et al.* Avoiding abundance bias in the functional annotation of post-translationally modified proteins. *Nat. Methods* **12**, 1003–1004 (2015).
42. 9 - Perform simple functional enrichment analysis and understand the concepts involved. https://gtpb.github.io/ADER18F/pages/L09.
43. Manda, S., Michael, D., Jadhao, S. & Nagaraj, S. H. Functional Enrichment Analysis. in *Encyclopedia of Bioinformatics and Computational Biology* 218–229 (Elsevier, 2019). doi:10.1016/B978-0-12-809633-8.20097-6.
44. Romano, J. P., Shaikh, A. M. & Wolf, M. multiple testing. in *The New Palgrave Dictionary of Economics* vol. 4 (Palgrave Macmillan, 2010).
45. Multiple Testing · Pathway Guide. https://www.pathwaycommons.org/guide/primers/statistics/multiple_testing/.
46. X, J. *et al.* DAVID-WS: a stateful web service to facilitate gene/protein list analysis. *Bioinforma. Oxf. Engl.* **28**, 1805–1806 (2012).
47. Mi, H. *et al.* PANTHER version 16: a revised family classification, tree-based classification tool, enhancer regions and extensive API. *Nucleic Acids Res.* **49**, D394–D403 (2021).
48. Thanati, F. *et al. FLAME: a web tool for functional and literature enrichment analysis of multiple gene lists*. 2021.06.02.446692 https://www.biorxiv.org/content/10.1101/2021.06.02.446692v1 (2021) doi:10.1101/2021.06.02.446692.
49. Liao, Y., Wang, J., Jaehnig, E. J., Shi, Z. & Zhang, B. WebGestalt 2019: gene set analysis toolkit with revamped UIs and APIs. *Nucleic Acids Res.* **47**, W199–W205 (2019).

50. Kolberg, L., Raudvere, U., Kuzmin, I., Vilo, J. & Peterson, H. gprofiler2 -- an R package for gene list functional enrichment analysis and namespace conversion toolset g:Profiler. *F1000Research* **9**, 709 (2020).
51. Raudvere, U. *et al.* g:Profiler: a web server for functional enrichment analysis and conversions of gene lists (2019 update). *Nucleic Acids Res.* **47**, W191–W198 (2019).
52. Bastian, M., Heymann, S. & Jacomy, M. Gephi : An Open Source Software for Exploring and Manipulating Networks. 2.
53. Kanehisa, M., Furumichi, M., Sato, Y., Ishiguro-Watanabe, M. & Tanabe, M. KEGG: integrating viruses and cellular organisms. *Nucleic Acids Res.* **49**, D545–D551 (2021).
54. Jassal, B. *et al.* The reactome pathway knowledgebase. *Nucleic Acids Res.* **48**, D498–D503 (2020).
55. Slenter, D. N. *et al.* WikiPathways: a multifaceted pathway database bridging metabolomics to other omics research. *Nucleic Acids Res.* **46**, D661–D667 (2018).
56. Giurgiu, M. *et al.* CORUM: the comprehensive resource of mammalian protein complexes-2019. *Nucleic Acids Res.* **47**, D559–D563 (2019).
57. Wingender, E. The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. *Brief. Bioinform.* **9**, 326–332 (2008).
58. Huang, H.-Y. *et al.* miRTarBase 2020: updates to the experimentally validated microRNA-target interaction database. *Nucleic Acids Res.* **48**, D148–D154 (2020).
59. Maleki, F., Ovens, K., Hogan, D. J. & Kusalik, A. J. Gene Set Analysis: Challenges, Opportunities, and Future Research. *Front. Genet.* **11**, (2020).
60. UniProt Consortium. UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **49**, D480–D489 (2021).
61. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
62. Apweiler, R. *et al.* The InterPro database, an integrated documentation resource for protein families, domains and functional sites. *Nucleic Acids Res.* **29**, 37–40 (2001).
63. Pletscher-Frankild, S., Pallejà, A., Tsafou, K., Binder, J. X. & Jensen, L. J. DISEASES: text mining and data integration of disease-gene associations. *Methods San Diego Calif* **74**, 83–89 (2015).
64. Gremse, M. *et al.* The BRENDA Tissue Ontology (BTO): the first all-integrating ontology of all organisms for enzyme sources. *Nucleic Acids Res.* **39**, D507–D513 (2011).
65. Number game. *Encyclopedia Britannica* https://www.britannica.com/topic/number-game.
66. Graph Theory. http://discrete.openmathbooks.org/dmoi2/ch_graphtheory.html.
67. Wilson, R. J. *Introduction to graph theory*. (Prentice Hall, 20).
68. What's the difference between a graph and a network? - Nothing. https://bence.ferdinandy.com/2018/05/27/whats-the-difference-between-a-graph-and-a-network/.
69. Koutrouli, M., Karatzas, E., Paez-Espino, D. & Pavlopoulos, G. A. A Guide to Conquer the Biological Network Era Using Graph Theory. *Front. Bioeng. Biotechnol.* **8**, (2020).
70. Dickson, A. Introduction to Graph Theory. 5.
71. Node degree statistics. http://rsat.sb-roscoff.fr/tutorials/neat_tutorial/Node_degree_statistics.html.
72. Colchester, J. Network Degree Distribution. *Systems Innovation* https://systemsinnovation.io/degree-distribution-articles/ (2016).
73. Barabási, A.-L. & Oltvai, Z. N. Network biology: understanding the cell's functional organization. *Nat. Rev. Genet.* **5**, 101–113 (2004).
74. Darden, L. Reductionism in Biology. in *eLS* (ed. John Wiley & Sons Ltd) 1–7 (John Wiley & Sons, Ltd, 2016). doi:10.1002/9780470015902.a0003356.pub2.
75. Lightbody, G. *et al.* Review of applications of high-throughput sequencing in personalized medicine: barriers and facilitators of future progress in research and clinical application. *Brief. Bioinform.* **20**, 1795–1811 (2019).
76. Charitou, T., Bryan, K. & Lynn, D. J. Using biological networks to integrate, visualize and analyze genomics data. *Genet. Sel. Evol.* **48**, 27 (2016).
77. Zhang, P. & Itan, Y. Biological Network Approaches and Applications in Rare Disease Studies. *Genes* **10**, 797 (2019).

78. Pavlopoulos, G. A., Soldatos, T. G., Barbosa-Silva, A. & Schneider, R. A reference guide for tree analysis and visualization. *BioData Min.* **3**, 1 (2010).
79. Luck, K. *et al.* A reference map of the human binary protein interactome. *Nature* **580**, 402–408 (2020).
80. Kaushik, A. C., Mehmood, A., Dai, X. & Wei, D.-Q. WeiBI (web-based platform): Enriching integrated interaction network with increased coverage and functional proteins from genome-wide experimental OMICS data. *Sci. Rep.* **10**, 5618 (2020).
81. Lindberg, D. A. Internet access to the National Library of Medicine. *Eff. Clin. Pract. ECP* **3**, 256–260 (2000).
82. Benson, D. A. *et al.* GenBank. *Nucleic Acids Res.* **43**, D30-35 (2015).
83. Howe, K. L. *et al.* Ensembl 2021. *Nucleic Acids Res.* **49**, D884–D891 (2021).
84. Odell, S. G., Lazo, G. R., Woodhouse, M. R., Hane, D. L. & Sen, T. Z. The art of curation at a biological database: Principles and application. *Curr. Plant Biol.* **11–12**, 2–11 (2017).
85. Yu, H., Huang, J., Zhang, W. & Han, J.-D. J. Network Analysis to Interpret Complex Phenotypes. in *Applied Statistics for Network Biology* 1–12 (John Wiley & Sons, Ltd, 2011). doi:10.1002/9783527638079.ch1.
86. Obayashi, T., Kagaya, Y., Aoki, Y., Tadaka, S. & Kinoshita, K. COXPRESdb v7: a gene coexpression database for 11 animal species supported by 23 coexpression platforms for technical evaluation and evolutionary inference. *Nucleic Acids Res.* **47**, D55–D62 (2019).
87. Franz, M. *et al.* GeneMANIA update 2018. *Nucleic Acids Res.* **46**, W60–W64 (2018).
88. Oughtred, R. *et al.* The BioGRID interaction database: 2019 update. *Nucleic Acids Res.* **47**, D529–D541 (2019).
89. Raina, P., Lopes, I., Chatsirisupachai, K., Farooq, Z. & Magalhães, J. P. de. GeneFriends 2021: Updated co-expression databases and tools for human and mouse genes and transcripts. *bioRxiv* 2021.01.10.426125 (2021) doi:10.1101/2021.01.10.426125.
90. Vandenbon, A. *et al.* Immuno-Navigator, a batch-corrected coexpression database, reveals cell type-specific gene networks in the immune system. *Proc. Natl. Acad. Sci. U. S. A.* **113**, E2393-2402 (2016).
91. Yang, S. *et al.* COEXPEDIA: exploring biomedical hypotheses via co-expressions associated with medical subject headings (MeSH). *Nucleic Acids Res.* **45**, D389–D396 (2017).
92. Obayashi, T., Aoki, Y., Tadaka, S., Kagaya, Y. & Kinoshita, K. ATTED-II in 2018: A Plant Coexpression Database Based on Investigation of the Statistical Property of the Mutual Rank Index. *Plant Cell Physiol.* **59**, e3 (2018).
93. Ogata, Y., Suzuki, H., Sakurai, N. & Shibata, D. CoP: a database for characterizing co-expressed gene modules with biological information in plants. *Bioinformatics* **26**, 1267–1268 (2010).
94. *Plant Genomics Databases: Methods and Protocols*. (Humana Press, 2017). doi:10.1007/978-1-4939-6658-5.
95. Manfield, I. W. *et al.* Arabidopsis Co-expression Tool (ACT): web server tools for microarray-based gene expression analysis. *Nucleic Acids Res.* **34**, W504–W509 (2006).
96. Dai, X., Zhang, S. & Zaleta-Rivera, K. RNA: interactions drive functionalities. *Mol. Biol. Rep.* **47**, 1413–1434 (2020).
97. Ramanathan, M., Porter, D. F. & Khavari, P. A. Methods to study RNA-protein interactions. *Nat. Methods* **16**, 225–234 (2019).
98. Yi, Y., Zhao, Y., Huang, Y. & Wang, D. A Brief Review of RNA–Protein Interaction Database Resources. *Non-Coding RNA* **3**, 6 (2017).
99. Fujimori, S., Hino, K., Saito, A., Miyano, S. & Miyamoto-Sato, E. PRD: A protein-RNA interaction database. *Bioinformation* **8**, 729–730 (2012).
100. Gene Ontology Consortium. The Gene Ontology resource: enriching a GOld mine. *Nucleic Acids Res.* **49**, D325–D334 (2021).
101. Lin, Y. *et al.* RNAInter in 2020: RNA interactome repository with increased coverage and annotation. *Nucleic Acids Res.* **48**, D189–D197 (2020).
102. Mann, M., Wright, P. R. & Backofen, R. IntaRNA 2.0: enhanced and customizable prediction of RNA-RNA interactions. *Nucleic Acids Res.* **45**, W435–W439 (2017).

103. Tuvshinjargal, N., Lee, W., Park, B. & Han, K. PRIdictor: Protein-RNA Interaction predictor. *Biosystems* **139**, 17–22 (2016).
104. Alipanahi, B., Delong, A., Weirauch, M. T. & Frey, B. J. Predicting the sequence specificities of DNA- and RNA-binding proteins by deep learning. *Nat. Biotechnol.* **33**, 831–838 (2015).
105. Amberger, J. S. & Hamosh, A. Searching Online Mendelian Inheritance in Man (OMIM): A Knowledgebase of Human Genes and Genetic Phenotypes. *Curr. Protoc. Bioinforma.* **58**, 1.2.1-1.2.12 (2017).
106. Keshava Prasad, T. S. *et al.* Human Protein Reference Database--2009 update. *Nucleic Acids Res.* **37**, D767-772 (2009).
107. Zhu, Y. *et al.* POSTAR2: deciphering the post-transcriptional regulatory logics. *Nucleic Acids Res.* **47**, D203–D211 (2019).
108. Blin, K. *et al.* DoRiNA 2.0--upgrading the doRiNA database of RNA interactions in post-transcriptional regulation. *Nucleic Acids Res.* **43**, D160-167 (2015).
109. Lewis, B. A. *et al.* PRIDB: a Protein-RNA interface database. *Nucleic Acids Res.* **39**, D277-282 (2011).
110. Cook, K. B., Kazan, H., Zuberi, K., Morris, Q. & Hughes, T. R. RBPDB: a database of RNA-binding specificities. *Nucleic Acids Res.* **39**, D301-308 (2011).
111. Shulman-Peleg, A., Nussinov, R. & Wolfson, H. J. RsiteDB: a database of protein binding pockets that interact with RNA nucleotide bases. *Nucleic Acids Res.* **37**, D369-373 (2009).
112. Nguyen, T. C., Zaleta-Rivera, K., Huang, X., Dai, X. & Zhong, S. RNA, action through interactions. *Trends Genet. TIG* **34**, 867–882 (2018).
113. Teng, X. *et al.* NPInter v4.0: an integrated database of ncRNA interactions. *Nucleic Acids Res.* **48**, D160–D165 (2020).
114. Gong, J. *et al.* RISE: a database of RNA interactome from sequencing experiments. *Nucleic Acids Res.* **46**, D194–D201 (2018).
115. Fang, S. *et al.* NONCODEV5: a comprehensive annotation database for long non-coding RNAs. *Nucleic Acids Res.* **46**, D308–D314 (2018).
116. Kozomara, A. & Griffiths-Jones, S. miRBase: annotating high confidence microRNAs using deep sequencing data. *Nucleic Acids Res.* **42**, D68-73 (2014).
117. Glažar, P., Papavasileiou, P. & Rajewsky, N. circBase: a database for circular RNAs. *RNA N. Y. N* **20**, 1666–1670 (2014).
118. O'Leary, N. A. *et al.* Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res.* **44**, D733–D745 (2016).
119. Karagkouni, D. *et al.* DIANA-TarBase v8: a decade-long collection of experimentally supported miRNA–gene interactions. *Nucleic Acids Res.* **46**, D239–D245 (2018).
120. Kodama, Y. *et al.* DNA Data Bank of Japan: 30th anniversary. *Nucleic Acids Res.* **46**, D30–D35 (2018).
121. Paraskevopoulou, M. D. *et al.* DIANA-microT web server v5.0: service integration into miRNA functional analysis workflows. *Nucleic Acids Res.* **41**, W169-173 (2013).
122. Paraskevopoulou, M. D. *et al.* DIANA-LncBase v2: indexing microRNA targets on non-coding transcripts. *Nucleic Acids Res.* **44**, D231–D238 (2016).
123. Vlachos, I. S. *et al.* DIANA-miRPath v3.0: deciphering microRNA function with experimental support. *Nucleic Acids Res.* **43**, W460–W466 (2015).
124. Zhou, B. *et al.* EVLncRNAs: a manually curated database for long non-coding RNAs validated by low-throughput experiments. *Nucleic Acids Res.* **46**, D100–D105 (2018).
125. Cheng, L. *et al.* LncRNA2Target v2.0: a comprehensive database for target genes of lncRNAs in human and mouse. *Nucleic Acids Res.* **47**, D140–D144 (2019).
126. Bao, Z. *et al.* LncRNADisease 2.0: an updated database of long non-coding RNA-associated diseases. *Nucleic Acids Res.* **47**, D1034–D1037 (2019).
127. Gao, Y. *et al.* Lnc2Cancer 3.0: an updated resource for experimentally supported lncRNA/circRNA cancer associations and web tools based on RNA-seq and scRNA-seq data. *Nucleic Acids Res.* **49**, D1251–D1258 (2020).
128. Harrow, J. *et al.* GENCODE: the reference human genome annotation for The ENCODE Project.

Genome Res. **22**, 1760–1774 (2012).
129. Maglott, D., Ostell, J., Pruitt, K. D. & Tatusova, T. Entrez Gene: gene-centered information at NCBI. *Nucleic Acids Res.* **39**, D52-57 (2011).
130. Yi, X., Zhang, Z., Ling, Y., Xu, W. & Su, Z. PNRD: a plant non-coding RNA database. *Nucleic Acids Res.* **43**, D982–D989 (2015).
131. Zhang, Z. *et al.* PMRD: plant microRNA database. *Nucleic Acids Res.* **38**, D806–D813 (2010).
132. Dai, X., Zhuang, Z. & Zhao, P. X. psRNATarget: a plant small RNA target analysis server (2017 release). *Nucleic Acids Res.* **46**, W49–W54 (2018).
133. Novel human lncRNA–disease association inference based on lncRNA expression profiles | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/29/20/2617/276977.
134. Lan, W. *et al.* LDAP: a web server for lncRNA-disease association prediction. *Bioinforma. Oxf. Engl.* **33**, 458–460 (2017).
135. Sun, J. *et al.* Inferring novel lncRNA–disease associations based on a random walk model of a lncRNA functional similarity network. *Mol. Biosyst.* **10**, 2074–2081 (2014).
136. Wang, J. *et al.* LncDisease: a sequence based bioinformatics tool for predicting lncRNA-disease associations. *Nucleic Acids Res.* **44**, e90 (2016).
137. Schriml, L. M. *et al.* Human Disease Ontology 2018 update: classification, content and workflow expansion. *Nucleic Acids Res.* **47**, D955–D962 (2019).
138. Lipscomb, C. E. Medical Subject Headings (MeSH). *Bull. Med. Libr. Assoc.* **88**, 265–266 (2000).
139. Miao, Y.-R., Liu, W., Zhang, Q. & Guo, A.-Y. lncRNASNP2: an updated database of functional SNPs and mutations in human and mouse lncRNAs. *Nucleic Acids Res.* **46**, D276 (2018).
140. Gao, Y. *et al.* LincSNP 3.0: an updated database for linking functional variants to human long non-coding RNAs, circular RNAs and their regulatory elements. *Nucleic Acids Res.* **49**, D1244–D1250 (2021).
141. Forbes, S. A. *et al.* COSMIC: mining complete cancer genomes in the Catalogue of Somatic Mutations in Cancer. *Nucleic Acids Res.* **39**, D945-950 (2011).
142. Huang, Z. *et al.* HMDD v3.0: a database for experimentally supported human microRNA–disease associations. *Nucleic Acids Res.* **47**, D1013–D1017 (2019).
143. Orchard, S. *et al.* The MIntAct project--IntAct as a common curation platform for 11 molecular interaction databases. *Nucleic Acids Res.* **42**, D358-363 (2014).
144. Chatr-aryamontri, A. *et al.* MINT: the Molecular INTeraction database. *Nucleic Acids Res.* **35**, D572–D574 (2007).
145. Orchard, S. *et al.* Protein interaction data curation: the International Molecular Exchange (IMEx) consortium. *Nat. Methods* **9**, 345–350 (2012).
146. Xenarios, I. *et al.* DIP: the database of interacting proteins. *Nucleic Acids Res.* **28**, 289–291 (2000).
147. Kotlyar, M., Pastrello, C., Malik, Z. & Jurisica, I. IID 2018 update: context-specific physical protein–protein interactions in human, model organisms and domesticated species. *Nucleic Acids Res.* **47**, D581–D589 (2019).
148. Szklarczyk, D. *et al.* The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Res.* **49**, D605–D612 (2021).
149. Brown, K. R. & Jurisica, I. Unequal evolutionary conservation of human protein interactions in interologous networks. *Genome Biol.* **8**, R95 (2007).
150. Meldal, B. H. M. *et al.* Complex Portal 2018: extended content and enhanced visualization tools for macromolecular complexes. *Nucleic Acids Res.* **47**, D550–D558 (2019).
151. Kooistra, A. J. *et al.* GPCRdb in 2021: integrating GPCR sequence, structure and function. *Nucleic Acids Res.* **49**, D335–D343 (2020).
152. PRIMES: Protein interaction machines in oncogenic EGF receptor signalling. *University of Edinburgh Research Explorer* https://www.research.ed.ac.uk/en/projects/primes-protein-interaction-machines-in-oncogenic-egf-receptor-sig.

153. Ranjan, R. *et al.* Channelpedia: An Integrative and Interactive Database for Ion Channels. *Front. Neuroinformatics* **0**, (2011).
154. Armstrong, J. F. *et al.* The IUPHAR/BPS Guide to PHARMACOLOGY in 2020: extending immunopharmacology content and introducing the IUPHAR/MMV Guide to MALARIA PHARMACOLOGY. *Nucleic Acids Res.* **48**, D1006–D1021 (2020).
155. Cotter, D., Guda, P., Fahy, E. & Subramaniam, S. MitoProteome: mitochondrial protein sequence database and annotation system. *Nucleic Acids Res.* **32**, D463-467 (2004).
156. Nastou, K. C., Tsaousis, G. N. & Iconomidou, V. A. PerMemDB: A database for eukaryotic peripheral membrane proteins. *Biochim. Biophys. Acta Biomembr.* **1862**, 183076 (2020).
157. Clerc, O. *et al.* MatrixDB: integration of new data with a focus on glycosaminoglycan interactions. *Nucleic Acids Res.* **47**, D376–D381 (2019).
158. Li, X., Wang, X. & Snyder, M. Systematic investigation of protein-small molecule interactions. *IUBMB Life* **65**, 2–8 (2013).
159. Kim, S. *et al.* PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res.* **49**, D1388–D1395 (2021).
160. Mendez, D. *et al.* ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res.* **47**, D930–D940 (2019).
161. Kuhn, M., Letunic, I., Jensen, L. J. & Bork, P. The SIDER database of drugs and side effects. *Nucleic Acids Res.* **44**, D1075-1079 (2016).
162. McFedries, A., Schwaid, A. & Saghatelian, A. Methods for the elucidation of protein-small molecule interactions. *Chem. Biol.* **20**, 667–673 (2013).
163. Wishart, D. S. *et al.* DrugBank 5.0: a major update to the DrugBank database for 2018. *Nucleic Acids Res.* **46**, D1074–D1082 (2018).
164. Szklarczyk, D. *et al.* STITCH 5: augmenting protein–chemical interaction networks with tissue and affinity data. *Nucleic Acids Res.* **44**, D380–D384 (2016).
165. Gallina, A. M., Bisignano, P., Bergamino, M. & Bordo, D. PLI: a web-based tool for the comparison of protein-ligand interactions observed on PDB structures. *Bioinforma. Oxf. Engl.* **29**, 395–397 (2013).
166. Anand, P., Nagarajan, D., Mukherjee, S. & Chandra, N. PLIC: protein–ligand interaction clusters. *Database* **2014**, (2014).
167. Murakami, Y., Omori, S. & Kinoshita, K. NLDB: a database for 3D protein–ligand interactions in enzymatic reactions. *J. Struct. Funct. Genomics* **17**, 101–110 (2016).
168. Ito, J., Ikeda, K., Yamada, K., Mizuguchi, K. & Tomii, K. PoSSuM v.2.0: data update and a new function for investigating ligand analogs and target proteins of small-molecule drugs. *Nucleic Acids Res.* **43**, D392–D398 (2015).
169. Tabei, Y. & Tsuda, K. SketchSort: Fast All Pairs Similarity Search for Large Databases of Molecular Fingerprints. *Mol. Inform.* **30**, 801–807 (2011).
170. Wang, C. *et al.* PDID: database of molecular-level putative protein-drug interactions in the structural human proteome. *Bioinforma. Oxf. Engl.* **32**, 579–586 (2016).
171. Kumar, R. *et al.* CancerDR: cancer drug resistance database. *Sci. Rep.* **3**, 1445 (2013).
172. Gohlke, B.-O., Nickel, J., Otto, R., Dunkel, M. & Preissner, R. CancerResource--updated database of cancer-relevant proteins, mutations and interacting drugs. *Nucleic Acids Res.* **44**, D932-937 (2016).
173. Coker, E. A. *et al.* canSAR: update to the cancer translational research and drug discovery knowledgebase. *Nucleic Acids Res.* **47**, D917–D922 (2019).
174. Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y. & Morishima, K. KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res.* **45**, D353–D361 (2017).
175. Kutmon, M. *et al.* PathVisio 3: an extendable pathway analysis toolbox. *PLoS Comput. Biol.* **11**, e1004085 (2015).
176. Boué, S. *et al.* Causal biological network database: a comprehensive platform of causal biological network models focused on the pulmonary and vascular systems. *Database J. Biol. Databases Curation* **2015**, bav030 (2015).
177. Slater, T. Recent advances in modeling languages for pathway maps and computable biological

networks. *Drug Discov. Today* **19**, 193–198 (2014).
178. Zhou, X., Menche, J., Barabási, A.-L. & Sharma, A. Human symptoms–disease network. *Nat. Commun.* **5**, 4212 (2014).
179. Lechner, M. *et al.* CIDeR: multifactorial interaction networks in human diseases. *Genome Biol.* **13**, R62 (2012).
180. Yue, M. *et al.* MSDD: a manually curated database of experimentally supported associations among miRNAs, SNPs and human diseases. *Nucleic Acids Res.* **46**, D181–D185 (2018).
181. Smith, C. L. & Eppig, J. T. The mammalian phenotype ontology: enabling robust annotation and comparative analysis. *Wiley Interdiscip. Rev. Syst. Biol. Med.* **1**, 390–399 (2009).
182. Piñero, J. *et al.* DisGeNET: a comprehensive platform integrating information on human disease-associated genes and variants. *Nucleic Acids Res.* **45**, D833–D839 (2017).
183. Bodenreider, O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res.* **32**, D267–D270 (2004).
184. Zeng, W., Min, X. & Jiang, R. EnDisease: a manually curated database for enhancer-disease associations. *Database* **2019**, (2019).
185. Cook, H. V., Doncheva, N. T., Szklarczyk, D., von Mering, C. & Jensen, L. J. Viruses.STRING: A Virus-Host Protein-Protein Interaction Database. *Viruses* **10**, 519 (2018).
186. Li, Y. *et al.* ViRBase: a resource for virus–host ncRNA-associated interactions. *Nucleic Acids Res.* **43**, D578–D582 (2015).
187. Niazi, F. & Valadkhan, S. Computational analysis of functional long noncoding RNAs reveals lack of peptide-coding capacity and parallels with 3' UTRs. *RNA N. Y. N* **18**, 825–843 (2012).
188. Xie, J. *et al.* Sno/scaRNAbase: a curated database for small nucleolar RNAs and cajal body-specific RNAs. *Nucleic Acids Res.* **35**, D183-187 (2007).
189. Fauquet, C. & Fargette, D. International Committee on Taxonomy of Viruses and the 3,142 unassigned species. *Virol. J.* **2**, 64 (2005).
190. Poelen, J. H., Simons, J. D. & Mungall, C. J. Global biotic interactions: An open infrastructure to share and analyze species-interaction datasets. *Ecol. Inform.* **24**, 148–159 (2014).
191. Parr, C. S. *et al.* The Encyclopedia of Life v2: Providing Global Access to Knowledge About Life on Earth. *Biodivers. Data J.* **2**, e1079 (2014).
192. Proceedings of the Environmental Information Management Conference 2011. https://www.lulu.com/content/paperback-book/shop/corinna-gries-and-matthew-jones/proceedings-of-the-environmental-information-management-conference-2011/paperback/product-1dn2n29g.html?page=1&pageSize=4.
193. Vandepitte, L. *et al.* A decade of the World Register of Marine Species – General insights and experiences from the Data Management Team: Where are we, what have we learned and how can we continue? *PLOS ONE* **13**, e0194599 (2018).
194. Fortuna, M. A., Ortega, R. & Bascompte, J. The Web of Life. *ArXiv14032575 Q-Bio* (2014).
195. Thompson, R. M. *et al.* Food webs: reconciling the structure and function of biodiversity. *Trends Ecol. Evol.* **27**, 689–697 (2012).
196. Bat Eco-Interactions. https://www.batbase.org/.
197. Pavlopoulos, G. A., Hooper, S. D., Sifrim, A., Schneider, R. & Aerts, J. Medusa: A tool for exploring and clustering biological networks. *BMC Res. Notes* **4**, 384 (2011).
198. Heberle, H., Carazzolle, M. F., Telles, G. P., Meirelles, G. V. & Minghim, R. CellNetVis: a web tool for visualization of biological networks using force-directed layout constrained by cellular components. *BMC Bioinformatics* **18**, 395 (2017).
199. Pavlopoulos, G., Wegener, A.-L. & Schneider, R. A survey of visualization tools for biological network analysis. *BioData Min.* **1**, 12 (2008).
200. Pavlopoulos, G. A., Paez-Espino, D., Kyrpides, N. C. & Iliopoulos, I. Empirical Comparison of Visualization Tools for Larger-Scale Network Analysis. *Adv. Bioinforma.* **2017**, e1278932 (2017).
201. Shannon, P. *et al.* Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res.* **13**, 2498–2504 (2003).
202. Mrvar, A. & Batagelj, V. Analysis and visualization of large networks with program package Pajek. *Complex Adapt. Syst. Model.* **4**, 6 (2016).

203. Köhler, J. *et al.* Graph-based analysis and visualization of experimental results with ONDEX. *Bioinformatics* **22**, 1383–1390 (2006).
204. Iragne, F., Nikolski, M., Mathieu, B., Auber, D. & Sherman, D. ProViz: protein interaction visualization and exploration. *Bioinformatics* **21**, 272–274 (2005).
205. Hu, Z. *et al.* VisANT 3.5: multi-scale network visualization, analysis and inference based on the gene ontology. *Nucleic Acids Res.* **37**, W115–W121 (2009).
206. Breitkreutz, B.-J., Stark, C. & Tyers, M. Osprey: a network visualization system. *Genome Biol.* **4**, R22 (2003).
207. Karatzas, E., Baltoumas, F. A., Panayiotou, N. A., Schneider, R. & Pavlopoulos, G. A. Arena3Dweb: Interactive 3D visualization of multilayered networks. *bioRxiv* 2020.11.20.391318 (2020) doi:10.1101/2020.11.20.391318.
208. Theocharidis, A., Dongen, S. van, Enright, A. J. & Freeman, T. C. Network visualization and analysis of gene expression data using BioLayout Express 3D. *Nat. Protoc.* **4**, 1535–1550 (2009).
209. Koutrouli, M., Hatzis, P. & Pavlopoulos, G. Exploring Networks in the STRING and Reactome Database. in (2020). doi:10.1016/B978-0-12-801238-3.11516-8.
210. Tokimatsu, T. *et al.* KaPPA-View. A Web-Based Analysis Tool for Integration of Transcript and Metabolite Data on Plant Metabolic Pathway Maps. *Plant Physiol.* **138**, 1289–1300 (2005).
211. Darzi, Y., Letunic, I., Bork, P. & Yamada, T. iPath3.0: interactive pathways explorer v3. *Nucleic Acids Res.* **46**, W510–W513 (2018).
212. Chong, J., Wishart, D. S. & Xia, J. Using MetaboAnalyst 4.0 for Comprehensive and Integrative Metabolomics Data Analysis. *Curr. Protoc. Bioinforma.* **68**, e86 (2019).
213. Amberger, J., Bocchini, C. A., Scott, A. F. & Hamosh, A. McKusick's Online Mendelian Inheritance in Man (OMIM®). *Nucleic Acids Res.* **37**, D793–D796 (2009).
214. The UniProt Consortium. UniProt: a worldwide hub of protein knowledge. *Nucleic Acids Res.* **47**, D506–D515 (2019).
215. Stelzer, G. *et al.* The GeneCards Suite: From Gene Data Mining to Disease Genome Sequence Analyses. *Curr. Protoc. Bioinforma.* **54**, 1.30.1-1.30.33 (2016).
216. Mitchell, A. L. *et al.* InterPro in 2019: improving coverage, classification and access to protein sequence annotations. *Nucleic Acids Res.* **47**, D351–D360 (2019).
217. Letunic, I., Khedkar, S. & Bork, P. SMART: recent updates, new developments and status in 2020. *Nucleic Acids Res.* **49**, D458–D460 (2021).
218. Bader, G. D. & Hogue, C. W. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **4**, 2 (2003).
219. Hermjakob, H. *et al.* IntAct: an open source molecular interaction database. *Nucleic Acids Res.* **32**, D452-455 (2004).
220. Schaefer, C. F. *et al.* PID: the Pathway Interaction Database. *Nucleic Acids Res.* **37**, D674–D679 (2009).
221. Caspi, R. *et al.* BioCyc: A Genomic and Metabolic Web Portal with Multiple Omics Analytical Tools. *FASEB J.* **33**, 473.2-473.2 (2019).
222. Expansion of the Gene Ontology knowledgebase and resources. *Nucleic Acids Res.* **45**, D331–D338 (2017).
223. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets—update. *Nucleic Acids Res.* **41**, D991–D995 (2013).
224. Okuno, Y. *et al.* GLIDA: GPCR - Ligand database for chemical genomics drug discovery - Database and tools update. *Nucleic Acids Res.* **36**, D907-12 (2008).
225. Günther, S. *et al.* SuperTarget and Matador: resources for exploring drug-target relationships. *Nucleic Acids Res.* **36**, D919-922 (2008).
226. Wang, Y. *et al.* Therapeutic target database 2020: enriched resource for facilitating research and early development of targeted therapeutics. *Nucleic Acids Res.* **48**, D1031–D1041 (2020).
227. Davis, A. P. *et al.* Comparative Toxicogenomics Database (CTD): update 2021. *Nucleic Acids Res.* **49**, D1138–D1143 (2021).
228. Roth, B. L., Lopez, E., Patel, S. & Kroeze, W. K. The Multiplicity of Serotonin Receptors: Uselessly Diverse Molecules or an Embarrassment of Riches? *The Neuroscientist* **6**, 252–262

(2000).
229. Burley, S. K. *et al.* Protein Data Bank (PDB): The Single Global Macromolecular Structure Archive. *Methods Mol. Biol. Clifton NJ* **1607**, 627–641 (2017).
230. Anastassiadis, T., Deacon, S. W., Devarajan, K., Ma, H. & Peterson, J. R. Comprehensive assay of kinase catalytic activity reveals features of kinase inhibitor selectivity. *Nat. Biotechnol.* **29**, 1039–1045 (2011).
231. Miljković, F. & Bajorath, J. Reconciling Selectivity Trends from a Comprehensive Kinase Inhibitor Profiling Campaign with Known Activity Data. *ACS Omega* **3**, 3113–3119 (2018).
232. Kuhn, M. *et al.* STITCH 4: integration of protein-chemical interactions with user data. *Nucleic Acids Res.* **42**, D401-407 (2014).
233. Pavlopoulos, G. A., Pafilis, E., Kuhn, M., Hooper, S. D. & Schneider, R. OnTheFly: a tool for automated document-based text annotation, data linking and network generation. *Bioinforma. Oxf. Engl.* **25**, 977–978 (2009).
234. Online publishing via pdf2htmlEX - Lu Wang and Wanmin Liu. https://wanminliu.github.io/doc/pdf2htmlEX/tb108wang.html.
235. Smith, R. An Overview of the Tesseract OCR Engine. in *Ninth International Conference on Document Analysis and Recognition (ICDAR 2007)* vol. 2 629–633 (2007).
236. Pafilis, E. & Jensen, L. J. Real-time tagging of biomedical entities. *bioRxiv* 078469 (2016) doi:10.1101/078469.
237. Buttigieg, P. L. *et al.* The environment ontology: contextualising biological and biomedical entities. *J. Biomed. Semant.* **4**, 43 (2013).
238. Schoch, C. L. *et al.* NCBI Taxonomy: a comprehensive update on curation, resources and tools. *Database J. Biol. Databases Curation* **2020**, baaa062 (2020).
239. Ashburner, M. *et al.* Gene Ontology: tool for the unification of biology. *Nat. Genet.* **25**, 25–29 (2000).
240. Junge, A. *et al.* RAIN: RNA-protein Association and Interaction Networks. *Database J. Biol. Databases Curation* **2017**, (2017).
241. Binns, D. *et al.* QuickGO: a web-based tool for Gene Ontology searching. *Bioinformatics* **25**, 3045–3046 (2009).
242. Jupp, S. *et al.* A New Ontology Lookup Service at EMBL-EBI. 2.
243. Xiahou, Z. *et al.* NMI and IFP35 serve as proinflammatory DAMPs during cellular infection and injury. *Nat. Commun.* **8**, 950 (2017).
244. Uhlén, M. *et al.* Proteomics. Tissue-based map of the human proteome. *Science* **347**, 1260419 (2015).
245. Köhler, S. *et al.* The Human Phenotype Ontology in 2021. *Nucleic Acids Res.* **49**, D1207–D1217 (2021).
246. Kanz, C. *et al.* The EMBL Nucleotide Sequence Database. *Nucleic Acids Res.* **33**, D29-33 (2005).
247. Sievert, C. *Interactive web-based data visualization with R, plotly, and shiny*.
248. Wickham, H. *ggplot2: Elegant Graphics for Data Analysis*. (Springer International Publishing, 2016). doi:10.1007/978-3-319-24277-4.
249. Henry, B. M., de Oliveira, M. H. S., Benoit, S., Plebani, M. & Lippi, G. Hematologic, biochemical and immune biomarker abnormalities associated with severe illness and mortality in coronavirus disease 2019 (COVID-19): a meta-analysis. *Clin. Chem. Lab. Med.* **58**, 1021–1028 (2020).
250. Danwang, C. *et al.* A meta-analysis of potential biomarkers associated with severity of coronavirus disease 2019 (COVID-19). *Biomark. Res.* **8**, 37 (2020).
251. Leisman, D. E. *et al.* Cytokine elevation in severe and critical COVID-19: a rapid systematic review, meta-analysis, and comparison with other inflammatory syndromes. *Lancet Respir. Med.* **8**, 1233–1244 (2020).
252. Elshazli, R. M. *et al.* Diagnostic and prognostic value of hematological and imunological markers in COVID-19 infection: A meta-analysis of 6320 patients. *PLOS ONE* **15**, e0238160 (2020).
253. Figliozzi, S. *et al.* Predictors of adverse prognosis in COVID-19: A systematic review and meta-analysis. *Eur. J. Clin. Invest.* **50**, e13362 (2020).

254. Tian, W. *et al.* Predictors of mortality in hospitalized COVID-19 patients: A systematic review and meta-analysis. *J. Med. Virol.* **92**, 1875–1883 (2020).
255. Choudhary, S., Sharma, K. & Silakari, O. The interplay between inflammatory pathways and COVID-19: A critical review on pathogenesis and therapeutic options. *Microb. Pathog.* **150**, 104673 (2021).
256. Mohamadian, M. *et al.* COVID‐19: Virology, biology and novel laboratory diagnosis. *J. Gene Med.* **23**, e3303 (2021).
257. Demeure, C. E. *et al.* Yersinia pestis and plague: an updated view on evolution, virulence determinants, immune subversion, vaccination, and diagnostics. *Genes Immun.* **20**, 357–370 (2019).
258. Gkouskou, K. *et al.* COVID-19 enters the expanding network of apolipoprotein E4-related pathologies. *Redox Biol.* **41**, 101938 (2021).
259. De Biasi, S. *et al.* Marked T cell activation, senescence, exhaustion and skewing towards TH17 in patients with COVID-19 pneumonia. *Nat. Commun.* **11**, 3434 (2020).
260. Evangelou, K. *et al.* Alveolar type II cells harbouring SARS-CoV-2 show senescence with a proinflammatory phenotype. *bioRxiv* 2021.01.02.424917 (2021) doi:10.1101/2021.01.02.424917.