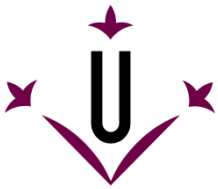
Development of computational tools to assist in the reconstruction of molecular networks

Ana Isabel Usié Chimenos

Dipòsit Legal: L.1715-2013
http://hdl.handle.net/10803/129848



Universitat de Lleida

Development of computational tools to assist in the reconstruction of molecular networks

ANA ISABEL USIÉ CHIMENOS

Supervisors

Rui Alves

Grup de Bioestadística i Biomatemàtica

Dept. de Ciències Mèdiques Bàsiques

Universitat de Lleida -IRBLleida -

Francesc Solsona

Grup de Computació Distribuïda

Dept. d'Informàtica i d'Enginyeria Industrial

Universitat de Lleida

Acknowledgments

I would like to start by thanking everyone who supported me during the course of this thesis.

First and foremost, I want to express my gratitude to my supervisors, Dr. Rui Alves and Dr. Francesc Solsona, for their continued encouragement and invaluable suggestions during this work. They have patiently supervised every little issue, always guiding me in the right direction. Without their help, this thesis dissertation would not have been possible.

A special thank you should also be given to all my more experienced colleagues, who helped me by sharing their research experience with me and providing useful advice. Some of these colleagues are from the Biostatistics and Mathematical Modeling in Biology Group (BBG): Albert Sorribas, Montserrat Rué, Ester Vilaprinyó, Montserrat Martinez and Joan Valls. The remaining are from the Group of Distributed Computing (GCD): Concepció Roig, Francesc Giné, Fernando Cores, Fernando Guirado, Josep Ll. Lérida, Josep Mª Solà and Valentí Pardo. I must say that working closely with such good people was a great experience.

During my Ph. D. research stay, I shared great moments with all the other students from BBG (Hiren Karathia, Carles Forné and Veronica Teixidó) and GCD (Ivan Teixidó, Josep Rius, Miquel Orobitg, Ignasi Barri, Damià Castellà, Hector Blanco, Alberto Montanyola, Jordi Vilaplana and Eloi Gabaldón). Without them it would have been lonely in the lab.

Furthermore, I want to thank all the members of the Structural Computational Biology Group (SCBG) from Centro Nacional de Investigaciones Oncologicas (CNIO). During six months, I had the chance to work in that group, with great people who made me feel at home. It was a great experience that marked me in many positive aspects. I would specially like to acknowledge Alfonso Valencia and Miguel Vázquez for all their support and for welcoming me as if I were one more.

Last but not least, I would like to dedicate this thesis to all my friends, especially to Llúcia who has endured my mood swings and given me good advice. Foremost, I dedicate this thesis to my closest family, especially to my boyfriend Frank for his emotional support, and for patience and understanding that he has shown during these long years. Thanks to my grandparents, Josep and Teresa, my parents, Fernando and Antonia, and my parents in law, Paco and Juani, for being there when I needed them. Finally, to my friend Georgina who unfortunately is no longer with us. I missed you a lot during these years.

Publications, Posters and Congress Presentations

Publications related with this thesis

1. A. Usié, H. Karathia, I. Teixidó, J. Valls, X. Faus, R. Alves and F. Solsona. **"Biblio-MetReS: A bibliometric network reconstruction application and server"**. *BMC Bioinformatics 2011, 12:387 doi:10.11.1186/1471-2105-12-387*

2. A. Usié, J. Cruz, J. Comas, F. Solsona and R. Alves. **"A tool for the identification of chemical entities (CheNER-BioC)"**. *Proceedings of the Fourth BioCreative Challenge Evaluation Workshop vol. 2* ,70-73

3. A. Usié, R. Alves, F. Solsona, M. Vázquez and A. Valencia. **"CheNER: Chemical Named Entity Recognizer"**. *Bioinformatics 2013, doi:10.1093/bioinformatics/btt639*

4. A. Usié, H. Karathia, I. Teixidó, R. Alves and F. Solsona. **"Biblio-MetReS for user-friendly mining of genes and biological processes in scientific documents"**. *[Submitted]*

Posters and oral talks in meetings related with this thesis

1. A. Usié, H. Karathia, F. Solsona and R. Alves. **"Biblio-MetReS: A Bibliometric Network Reconstruction Server"**. *The XII internation Congress on Molecular Systems Biology: Biological Desing Principle · ICMSB 8-12 May 2011. (Poster)*

2. A. Usié, I. Teixidó, H. Karathia, F. Solsona and R. Alves. **"Biblio-MetReS: A Bibliometric Network Reconstruction Server"**. *Student Symposium of the XIth Spanish Symposium on Bioinformatics · JBI2012 23-25 Jan 2012. (Poster)*

3. A.Usié, J. Cruz, J. Comas, F. Solsona and R. Alves. **"A tool for the identification of chemical entities (CheNER-BioC)"**. *Fourth BioCreative Challenge Evaluation Workshop. 2013. (Poster)*

Other publications

1. B. Salvado, H. Karathia, A. Usié, E. Vilaprinyó, S. Omholt, A. Sorribas and R. Alves. **"Methods for and results from the study of design principle in moleculars systems"**. *Mathematical Biosciences, Volume 231, Issue 1, Special issue on biological desing principles, May 2011, Pages 3-18, ISSN 0025-5564, DOI: 10.1016/j.mbs.2011.02.005.*

2. Abdelli, M. O., Usié, A., Karathia, H., Vilaplana, J., Solsona, F., and Alves, R. **"Parallelizing Biblio-MetReS, a Data Mining Tool"**. *Actas XXII Jornadas de Paralelismo (JP2011), Sep 2011.*

3. Teixidó, I., Usié, A., Lérida, J. L., Solsona, F., Comas, J., Torres, N., Karathia, H. and Alves, R. **"P-Biblio-MetReS, a parallel data mining tool for the reconstruction of molecular networks"**. *Proceedings of the 20th European MPI Users' Group Meeting*, Sep 2013, *Pages 247-252. ACM.*

4. Karathia, H., Usié, A., Teixidó, I., Vilaprinyó, E., Sorribas, A., Solsona, F., and Alves, R., **"A human centric comparison of eukaryotic proteomes: Implications for the study of human biology"**. *[Submitted]*

5. Karathia, H., Teixidó, I., Usié, A., Vilaprinyó, E., Solsona, F., Sorribas, A., and Alves, R. **"Homol-MetReS: An integrated framework tool to study evolutionary molecular systems biology"**. *[Under preparation].*

Other posters and oral talks in meetings

1. H. Blanco, J.L. Lérida, F. Guirado and A. Usié. **"Modelo de asignación de tareas en entornos multicluster heterogéneos y no dedicados"**. *XX Jornadas del paralelismo, A Coruña 16-18 Sep, 2009. (Poster)*

2. H. Karathia, A. Usié, E. Vilaprinyó, A. Sorribas, F. solsona and R. Alves. **"Proteome-MetReS: Network Reconstruction Based on Whole Proteome Comparisons"**. *The XII internation Congress on Molecular Systems Biology: Biological Desing Principle ·ICMSB 8-12 May 2011. (Poster)*

3. O. Abdelli, A. Usié, H. Karathia, J.Vilaplana, F. Solsona and R. Alves. **"Parallelizing Biblio-MetReS, a data mining tool"**. *XXII Jornadas del paralelismo, La Laguna 7-9 Sep, 2011 (Presentation)*

4. Karathia, H., Usié, A., Vilaprinyó, E., Solsona, F., Teixidó, I., Sorribas A., Alves, R. (2012) **"Homol-MetReS: A web application for integration between molecular systems biology and evolutionary biology"**. *Student Symposium of the XIth Spanish Symposium on Bioinformatics JBI2012 23-25 Jan 2012. (Presentation)*

5. Teixidó, I., Usié, A., Lérida, J. L., Solsona, F., Comas, J., Torres, N., Karathia, H. and Alves, R. **"P-Biblio-MetReS, a parallel data mining tool for the reconstruction of molecular networks"**. *Proceedings of the 20th European MPI Users' Group Meeting*, Sep 2013, *Pages 247-252. ACM. (Presentation)*

Abbreviations

AIM, *Ab Initio* Modeling

CEM, Chemical Entity Mention recognition

CDI, Chemical Document Indexing

CRF, Conditional Random Fields

FRM, Fold-Recognition Modeling

GO, Gene otology

GUI, Graphical User Interface

HM, Homology Modeling

IE, Information Extraction

IR, Information Retrieval

IUPAC, Union of Pure and Applied Chemistry

NEN, Named Entity Normalization

NER, Named Entity Recognition

NGS, Next Generation Sequencing

NLP, Natural Language Processing

PDB, Protein Data Bank

PMP ,Protein Model Portal

PPD, Protein-Protein Docking

PPI, Protein-Protein Interaction

SMR, SWISS-MODEL Repository

TM, Text-Mining

Summaries

1.1 English

The aim of this thesis is the development and implementation of a set of data mining tools to aid in the reconstruction of biological circuits through analysis and integration of large biological datasets. These circuits are important because they regulate all processes that maintain life and health in organisms.

The main part of the thesis is focused on analyzing bibliomic data, for which I developed two tools. One of the tools, Biblio-MetReS, extracts information about co-occurrence of proteins and/or biological processes from scientific documents, relying on the idea that if genes or proteins co-occur in the same document(s) they are likely to be functionally related. The detection of such co-occurrence can be used for the reconstruction of Protein-Protein Interaction network and to identify the processes in which the networks are involved. The other text-mining tool, CheNER, identifies different types of chemical compound names in scientific documents. The identification of such names is the starting step to implement subsequent tools that identify how the chemical compounds regulate proteins and biological processes, in both health and disease.

The final tool I developed focuses on the integration of methods for structural analysis and modeling of proteins with docking methods for the prediction of native protein-protein physical complexes. This integration might in the future facilitate using these methods to further assist in biological network reconstruction.

1.2 Català

L'objectiu d'aquesta tesi és desenvolupar i implementar un conjunt d'eines de mineria de dades per ajudar en la reconstrucció de circuits biològics a través de l'anàlisi i la integració de grans conjunts de dades biològiques. Aquests circuits són importants perquè regulen tots els processos que controlen la vida i la salut dels organismes .

El treball principal de la tesi es centra en l'anàlisi de les dades bibliòmiques, desenvolupant-se amb aquest fi dues eines diferents. Una de les eines, Biblio-MetReS, extreu informació sobre co-ocurrència de proteïnes i/o processos biològics dels documents científics, basant-se en la idea que si els gens o proteïnes co-ocorren en el mateix document(s) és probable que estiguin funcionalment relacionats. La detecció d'aquestes co-ocurrències es pot utilitzar per a la reconstrucció de la xarxa d'interacció proteïna-proteïna i per la identificació dels processos en què intervenen aquestes xarxes. L'altra eina de mineria de text, CheNER, identifica els diferents tipus de noms dels compostos químics en documents científics. La identificació d'aquests noms és el primer pas per començar a posar en pràctica les eines subsegüents que identifiquen els compostos químics regulen les proteïnes i els processos biològics, tant en la salut com en la malaltia

L'eina final desenvolupada es centra en la integració de mètodes per a l'anàlisi estructural i modelització de proteïnes amb mètodes d'acoblament per a la predicció de complexos físics de proteïna-proteïna. Aquesta integració en el futur podria facilitar l'ús d'aquests mètodes per ajudar en la reconstrucció de la xarxa biològica.

1.3 Castellano

El objetivo de esta tesis es desarrollar e implementar un conjunto de herramientas de minería de datos para ayudar en la reconstrucción de circuitos biológicos a través del análisis y la integración de grandes conjuntos de datos biológicos. Estos circuitos son importantes porque regulan todos los procesos que controlan la vida y la salud de los organismos.

El trabajo principal de la tesis se centra en el análisis de los datos bibliómicos, desarrollándose con este fin dos herramientas diferentes. Una de las herramientas, Biblio-MetReS, extrae información sobre co-ocurrencia de proteínas y/o procesos biológicos de los documentos científicos, basándose en la idea de que si los genes o proteínas co-ocurren en el mismo documento (s) es probable que estén funcionalmente relacionados. La detección de estas co-ocurrencias se puede utilizar para la reconstrucción de la red de interacción proteína-proteína y para identificar los procesos en los que intervienen estas redes. La otra herramienta de minería de texto, CheNER, identifica los diferentes tipos de nombres de compuestos químicos en documentos científicos. La identificación de estos nombres es el primer paso para empezar a poner en práctica las herramientas subsiguientes que identifican los compuestos químicos que regulan las proteínas y los procesos biológicos, tanto en la salud como en la enfermedad.

La herramienta final que he desarrollado se centra en la integración de métodos para el análisis estructural y modelado de proteínas con métodos de acoplamiento para la predicción de complejos físicos de proteína-proteína. Esta integración en el futuro podría facilitar el uso de estos métodos para ayudar en la reconstrucción de la red biológica.

CONTENTS

Chapter 1. Introduction

This chapter begins by introducing the research areas and concepts that are relevant for the work presented in the current thesis. Relevant research from the areas of Systems Biology and Bioinformatics is introduced and discussed to motivate the work I did. Next, I list the main goals of my work and briefly describe the contributions that allowed me to achieve these goals, as well as some collaborations that are relevant to these goals. Finally, I present the organization of the remaining chapters of this thesis.

1.1 Overview

Molecular systems biology is an approach to biomedical and biological scientific research where computational methods play an important role. This approach aims at analyzing the molecular components of the cells and their behavior in an integrated manner in order to understand how they function when they are assembled and discover emerging properties of cells, tissues and organisms.

Systems biology also aims at analyzing the emerging properties of the interacting molecular components of cells to discover design principles in molecular and cellular circuits. These principles are extracted from different studies that identify the topologies [1–4], the range of parameters [5–7] and the dynamic behavior [2,8] of a particular biological circuit. Such studies correlate the action of natural selection on the alternative designs for the circuits to the effect of the design variations on the "fitness" of an organism [9,10], explaining selection of alternative designs for the same function in different organisms.

However, the identification of design principles in biological circuits requires knowing what alternatives are available for natural selection to work with. In many cases this information is unknown, even at the level of the circuit composition and topology. For example, on average, 20% of the genes annotated in a fully sequenced genome for a new organism are hypothetical or have unknown function. This implies that there are molecular circuits that we have yet to discover.

Before aiming at a systematic identification of design principles in molecular circuits we first need to reconstruct those circuits. This implies identifying the individual function of genes with unknown function, understanding their contribution to the functioning of the circuits of interest, and cataloging the variations that are observed in those circuits over the set of organisms in the tree of life. It is only after getting to this stage that we can study the design principles of molecular circuits. The research presented in this thesis contributes to these steps that are preliminary to, and yet fundamental for, the study of design principles.

Currently, the preliminary reconstruction of circuits can rely on an increasing number of techniques, methods and tools that enable us to collect comprehensive data sets, such as whole genome gene expression changes, gene or protein sequences, structures, and interactions, etc. These techniques are leading to the accumulation of large datasets at a rate that makes it impossible for any one person to analyze, organize, and integrate all the available information.

In this **thesis** I focus on the **development and implementation of a set of data mining techniques and tools** that facilitate the **analysis and integration of biological information** derived from these large datasets **with the aim of assisting in circuit reconstruction**. Developing such tools is one of the main aims of **Bioinformatics** [11–13]. Bioinformaticians develop methods to mine, integrate, and represent the information out of the various biological datasets, and build software applications that do so in a user-friendly way. The integration of these data is hard, due to the difficulty of defining and implementing identification standards and functional classifications of proteins that are universally valid and accepted. Because of this there is a significant amount of redundant and sometimes contradictory information between different databases that is difficult to curate automatically. This highlights the importance of creating universally accepted classifications, examples of which are Gene Ontology (GO) [14] and the Protein Naming Utility [15].

The creation of such classifications facilitated the development of different applications that integrate data from different sources in order to reconstruct biological circuits [16–21]. The public distribution of these software applications to the research community potentiates appropriate processing, analysis, and display of the information available in large datasets in a simple, informative, and organized way. Thus, the appropriate integration of that information facilitates *in silico* reconstruction of the biological circuits, pathways, and networks [22]. Furthermore, different types of datasets are available for such *in silico* reconstruction. I briefly describe each of those types (summarized in Table 1) and discuss different methods and tools that are most commonly used to extract information that aids in the reconstruction of molecular networks.

Table 1. Types of data sets available for *in silico* reconstruction.

Types of datasets	Description of the type of data	Main Repositories
Bibliomic data	Literature published in scientific journals and textbook	Medline [1], Pubmed[1], Biomed Central [2] and PLoS[3]
Sequence data	Genome sequences, gene/protein sequences and functional annotations	NCBI[1], KEGG[4]
Structural data	Protein structures and structure classifications	PDB[5], CATH [6], SCOP[7]
Proteomic data	Protein information about activity, concentrations, participation, localization and posttranslational modifications.	UniProt [8], InterPro [9], Expasy[10]
Metabolomic data	Information regarding changes of metabolic fluxes and concentrations over time or under different conditions.	Biological Magnetic Resonance Data Bank[11]
Gene Expression data	Information about how gene expression changes over time under different conditions	SAGE[12], GEO[1]

[1] http://www.ncbi.nlm.nih.gov/
[2] http://www.biomedcentral.com/
[3] http://www.plosone.org/
[4] http://www.genome.jp/kegg/
[5] http://www.rcsb.org/
[6] http://www.cathdb.info/
[7] http://scop.bic.nus.edu.sg/
[8] http://www.uniprot.org/
[9] http://www.ebi.ac.uk/interpro/
[10] http://www.expasy.org/
[11] http://www.bmrb.wisc.edu/metabolomics/external_metab_links.html
[12] http://www.sagenet.org/

1.2 Types of data sets for *in silico* reconstruction

1.2.1 Bibliomic data

An increasingly large body of scientific literature has accumulated for the last century. Databases such as Medline [23] collect data from these published documents in an organized way, facilitating researchers' access to the literature. Thanks to text-mining applications and Natural Language Processing (NLP) techniques, these documents can be automatically mined in order to extract information about genes, proteins, biological processes, chemicals or drugs. This information can then be used to reconstruct the network of interactions and regulation that mediates biological processes of interest.

The recent development of tools like iHOP [16] or STRING [18] permitted using those techniques in a user friendly way for automated reconstruction of networks of co-occurrence for genes and proteins in the scientific literature. The assumption underlying the application of these tools is that if genes or proteins co-occur in the same document(s) they are likely to be functionally related among them. This automatic detection of networks from text-mining has been used as an starting point to reconstruct biological pathways [1,2], although it needs to be curated and complemented with other types of information.

There are several limitations to the text-mining approaches used for network reconstruction. One of these limitations is the type of documents that are analyzed. For example, many applications mine Medline abstracts. This is the case of iHOP and STRING. If one mines only Medline, a large fraction of the information contained in a paper is not analyzed. Hence there is room for improvement by analyzing full documents extracting much more information than in abstracts. Initiatives such as Pubmed, Biomed Central or PLoS are crucial in making available the full text of scientific papers, trying to reduce the problem of the copyrighted nature of full paper which prevents their used [24–27]. Another limitation is the strategy for database analysis. Both iHOP and STRING rely on having the documents analyzed previously, and storing the

Protein-Protein Interactions (PPIs) found in each document in a database (precompiled information). This reduces user execution time because is not necessary to parse the documents at the time of execution. However, the results are not up-to-date, as updates to their databases are not as frequent as those in Medline or Pubmed.

Mining information about the names of chemical compounds and their possible biological interactions is another area of intense text mining research. This interest has arisen with the emergence of freely available chemical databases such as PubChem [28] or DrugBank [29]. These databases permitted the implementation of novel tools that use text-mining and NLP techniques to identify chemical compound names in text. Some of these tools are publicly available, like OSCAR [30] and ChemSpot [31]. The automatic identification of such names is the base for the creation of new tools to extract information related to the pharmaceutical treatment of diseases and identify relationships between chemical compounds and genes/proteins that can help to understand how those compounds modulate gene/protein activities.

The detection of chemical compounds is challenging because they are mentioned in the literature using different nomenclatures, each with specific morphological features. The same compound can have widely varying names, depending on the nomenclature one uses. Aspirin provides an elucidating example : (1) ***trivial*** or ***brand*** name is "*Aspirin*"; (2) ***systematic*** or ***IUPAC*** name is "*2-(acetyloxy)benzoic acid*"; (3) SMILES ***identifier*** is (*CC(=O)Oc1ccccc1C(O)=O*), InChi ***identifier*** is (*InChI=1S/C9H8O4/c1-6(10)13-8-5-3-2-4-7(8)9(11)12/h2-5H,1H3,(H,11,12)*), CAS numbers are (*50-78-2*), etc.; (4) ***family*** name is "*Salicylates*"; (5) ***molecular formula*** is "$C_9H_8O_4$"; and (6) ***abbreviations*** is "*ASA*".

While OSCAR and ChemSpot are developed to identify all types of chemicals names with no indication of what type they are, there is no tool publicly available that explicitly distinguishes between the different types of nomenclature. In particular, specifically identifying systematic or IUPAC chemical names is an important feature that these two applications lack. This is a limitation, as such names are the most commonly used in important scientific

documents such as patents. The chemical structure of a compound can be easily derived from the IUPAC name, which facilitates understanding the reactivity of the compound and its possible interactions in disease treatments.

Due to the increasing interest in the detection of chemical compounds names in text and the lack of corpora annotated, the recent BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology) IV Challenge [32] had one of its tracks focused on the identification of chemical compound names in text, providing annotated corpora that, after the challenge, will be publicly available for all the community.

1.2.2 Sequence data

Due to the decreasing of costs, genome sequencing projects are now come place. Because of this, the DNA sequence of thousands of organisms and individuals have been decoded and stored in a few central repositories that contain most of that sequence information. Many of these, such as the Sanger center [33], KEGG [34–38] and the NCBI repositories [39–45], are freely accessible via their web pages. They contain genome sequences and complete gene/protein sequences and annotations.

The association of sequence to function facilitates the annotation of subsequently sequenced genomes [46–49]. This relies on the assumption that proteins with similar sequences often carry out similar functions [50]. In other words, the more similar two sequences are, the more likely it is that they have a similar function. Thus, by comparing sequences with associated annotation to those of the newly discovered genes from freshly sequenced genomes functional information is transferred from the already annotated sequence to the new one. Today, tools such as BLAST [51] or HMMER [52] are widely used to aid in the search for homologues of a sequence.

1.2.3 Structural data

As stated above, on average 20% of genes annotated in new genomes are either predicted proteins or have homologues with unknown function. If this is so and

structural information is available for the proteins that they code, that information can aid attributing general or specific functions to that gene.

Owing to the abundance of structural genomics projects, the amount of available structural data is increasing continuously. However, the number of available protein structures is still far behind the number of protein sequence due to the difficulties with the experimental structure determination. It is known that strong correlations exist between structure conservation and functional conservation, and between sequence conservation and structure conservation [50,53,54]. Thus, sequence implies structure which implies function. This knowledge can be used to reduce the huge gap between the number of sequence and structures leading to implement computational methods that aid in the prediction of protein tertiary (3D) structures, also known as structural protein modeling.

Available structural data is collected in central repositories such as Protein Data Bank (PDB) [55], SCOP [56], CATH [57] and FireDB [58]. The PDB contains information about the 3D structures of large biological molecules, including proteins and nucleic acids, SCOP and CATH contain structure classifications and FireDB contains PDB structures and their associated ligands.

The way that two proteins interact physically (the formation of the complex) has functional consequences. Because of this it is important to predict how two proteins forms a complex in order to study, for instance, if any mutation in one of the protein sequences can changes the native conformation of the complex and affect biological function. There are also computational docking methods that can be used to predict protein interactions *in silico*.

Structural analysis

There are three main computational methods for structural analysis: the template-based homology modeling, the fold-recognition and the so-called *ab initio* prediction. The best strategy for protein 3D structure prediction first involves homology modeling followed by fold recognition, and if not successful, *ab initio* prediction (see Figure 1).

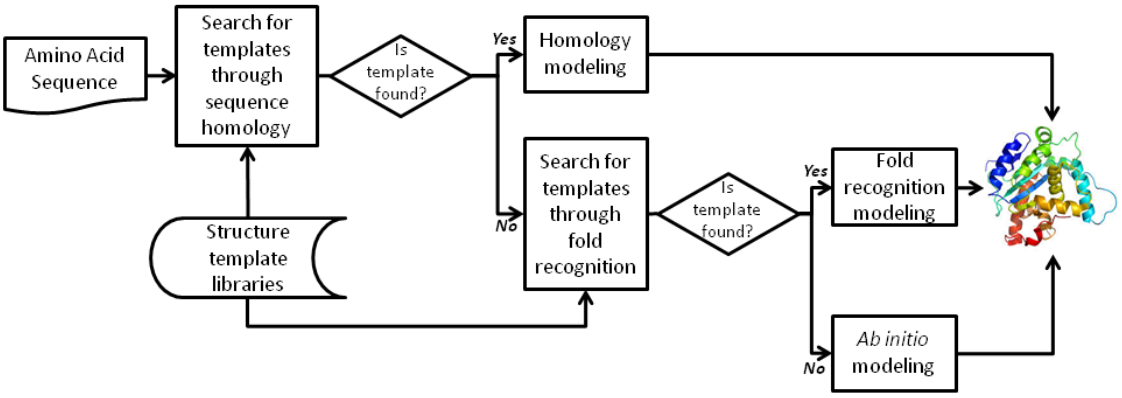


Figure 1. A typical strategy for modeling protein structure. This strategy works as follows: Search templates via sequence homology of the target sequence. If a template is found use homology modeling. Otherwise, search for templates via fold recognition. If a template is found use fold-recognition modeling. Otherwise, use *ab initio* modeling.

Homology modeling exploits the strong correlation that exists between sequence and structure to predict the structure of a protein. This prediction relies on a preexisting template structure of a protein that is homologous to the protein whose structure one wants to predict. There are a set of tools that implements the homology modeling approach such as MODELLER and SWISS-MODEL [59–64]. BLAST and other similar tools assist in this approach by facilitating the identification of homologous sequences.

Fold recognition modeling methods assist in the recognition and assignation of the correct fold when the sequence comparison methods are not sensitive enough to recognize a sufficiently similar sequence homologue. Tools that use fold-recognition methods, such as Phyre2 [65], have two main functions in common: (1) a function to align the target sequence with the fold of the structures contained in a library of representative or unique structures and (2) an energy function. Depending of which algorithms are used the fold-recognition can be divided into four classes. More details about each of these classes can be found in the literature [66].

Ab initio modeling methods provide candidate conformations for the protein structure one wants to predict. These conformations are built using methods that fragment the query sequence into short sequence of amino acids (typically nine residues) [67]. Then, candidate structures for these fragments are generated using template based techniques similar to the ones used in

homology modeling, and are stochastically sampled and assembled to construct a low energy protein conformation. Once these conformations are built, the method chooses one among them by comparing thermodynamic stabilities and energy states. Most of the tools that implement this approach, such as ROSETTA [68], have the three following factors in common: (1) an energy function that compares all possible conformations and identifies the most thermodynamically stable state as the native protein structure; (2) a search method to identify the low-energy states through conformational search; (3) a selection method of native-like models from a collection of conformations.

In brief, homology modeling builds a protein 3D structure from a template and it is the most accurate and successful structural modeling approach, especially when the template used to create the protein 3D structure has a high sequence identity to the target protein. On the other end of the structural modeling spectrum *ab initio* methods are, potentially, the least accurate approach to predict protein structure. In addition, they are limited by the high computational cost and by the bottlenecks in the energy functions. However, when no suitable template is available, *ab initio* methods can play an important role identifying new folds.

Protein docking

Docking is the process by which one predicts how two proteins might form a physical complex (see Figure 2). It can be conceptually simplified as a problem of "*lock-and-key*", where one is interested in find the right orientation of the *key* that will open the *lock*.

The aim of docking is to simulate the molecular recognition process that leads to specific interaction between two or more molecules. There are two main docking approaches. The first approach analyzes the structure of the proteins and presents the most likely docking configuration that maximizes shape complementary in the interface between two proteins [69–71]. The second approach simulates the physical process of docking, for example by using molecular dynamics calculations to predict how the two proteins to be docked

might approach in space and what the final complex between the two might look like [72]. Both types of approach follow the same three step-process: (1) representation of the system, (2) conformational space search and (3) ranking of potential solutions. The docking problem also requires an efficient search procedure and a good scoring function (See [73] for more details). There are several tools that permit protein docking of either experimentally determined or computationally predicted protein structures, with the latter type of docking being much more error-prone and less accurate. Some of the most accurate tools are HEX [74–76], GRAMM-X [77,78] and RosettaDocking [79,80].

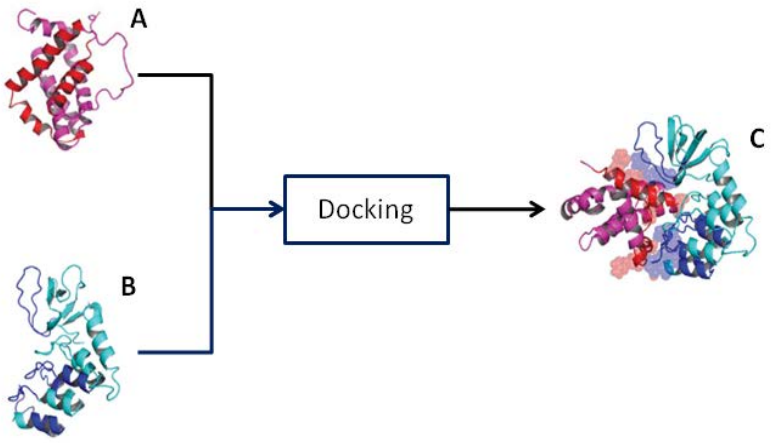


Figure 2. Scheme of protein docking prediction. A and B are two different proteins. C is the complex predicted by the docking approach of how the two proteins physically interact. Docking simulate how two proteins interact physically looking for different configurations evaluating them using appropriate scoring functions.

1.2.4 Proteomics data

Proteomics is the study of proteins, their activity, concentrations, participation in protein complexes, their localization, their interactions and posttranslational modifications. The experimental evolution in proteomics has led to the development, integration, and automation of various high throughput proteomics techniques and equipments, such as mass spectrometry and protein microarrays. Those techniques and methods permit to divide, identify, quantify and characterize proteins, collecting and organizing the resulting information in different databases such as UniProt [81,82], InterPro [83–86], and ExPASy [87,88], where various tools are available to mine the data.

Proteomics data is fundamental to assess the *in vivo* functionality of proteins. These data can assist in pathway reconstruction, providing quantitative information about proteins and their regulation and facilitating the reconstruction of gene circuits and PPI networks [89–93].

1.2.5 Metabolomics data

The term metabolomics was coined at 1997 and defined as "*the comprehensive, qualitative, and quantitative study of all the small molecules in an organism*" [94]. Since then, metabolomics studies generate large amounts of data with a complex structure. These data required developing a variety of processing techniques for its analysis [95,96]. Depending on the purpose of one's experiment, the most adequate technique will differ. There are three main approaches for the analysis of the metabolome: metabolite profiling which aims to identify and quantify metabolites [96,97], metabolic fingerprinting which is used in tissue comparisons, and metabolomics which focuses on the metabolic response of organisms to pathophysiological stimuli or genetic modification [98,99]. Examples of metabolomics studies are plenty. For example, metabolite profiling of blood plasma samples of individuals subjected to toxicological stimuli is used to measure or detect any physiological changes caused by any chemical or drug. Additionally, observed metabolic changes can sometimes be related to specific syndromes, and that fact is particularly relevant for the pharmaceutical companies which are waiting to test the potential and toxicity of new drugs. Findings from this type of study may have large social and economic implications. This is clearly seen if one considers that detecting adverse toxic effects of a drug [100] before clinical trials may save lives and save development costs.

The data obtained from metabolomic approaches allow, in principle, to reconstruct a metabolic network, including the regulatory influences of the metabolites on the different reactions of that network. Any change in the concentrations and/or fluxes going through that network and related to an external factor can be associated to the physiology of the cell.

1.2.6 Gene Expression data

The concept of gene expression describes a complex process in which the information contained within the genome is translated into the phenotype of the organism. DNA microarrays and NGS (Next Generation Sequencing) technologies can be used to obtain high throughput gene expression profiles that give information about how the expression of genes changes over time under different conditions and elucidate the correlation between gene expression and biochemical pathways. The datasets from these measurements are often deposited at SAGE [101,102], GEO [40,41], or other repositories, and made freely available to the community. From these data one can infer which genes are significantly related with specific cellular responses, because genes that belong to a common pathway are usually regulated in a similar way. Similarities in expression patterns can thus be used to assign general function to unknown genes and to assist in pathway network reconstruction [103].

1.3 Goals

Identifying the biological design principles of a molecular circuit requires that one knows how the circuit is wired and works. Thus, before analyzing design principles one must have a large set of reconstructed molecular circuits. The systematic and automated reconstruction of those circuits requires integrating various types of data with different origins. This reconstruction is one of the main goals/problems of Systems Biology today. **It is the general objective of this thesis to contribute for the development of tools implementing *in silico* methods that facilitate that reconstruction.**

To achieve this, the thesis has the following specific goals:

1. Develop a tool for analyzing scientific documents from different sources, such as Biomed Central, Pubmed, among others. This tool will identify co-occurrence between genes/proteins to reconstruct networks of those involved in the same biological process. The tool also will identify biological processes and pathways to calculate co-occurrences between those and genes/proteins. These relations will give clues that will assist the researcher in identifying which genes or proteins are related to a set of biological processes or pathways.

2. Develop a tool for the identification of chemical compounds and drugs in documents to infer regulation that might occur in the regulatory circuits. As a first step to this I built a tool for identifying those chemicals and drugs that can affect or change the regulation of the biological function of genes and proteins. Specifically, I focus on the identification of the systematic chemical names due to their prevalence in pharmaceutical patents which studied the toxicity of this compound and the functional changes that it can make. However, this tool also permits the identification of other types of chemical names.

3. Implement an application that integrates structural information and predictions to infer causal interactions in molecular circuits. This tool will use structural information extracted from different resources (PDB, SWISS-MODEL, Phyre2, and MODELLER), to perform *in silico* docking using alternative protocols (RosettaDocking, Gramm-x, and Hex), and infer causal interactions in the molecular circuits.

1.3.1 Achieving the specific goals

To achieve the first specific goal I created a tool, Biblio-MetReS, that addresses the two issues described in the bibliomic data section: (1) a large fraction of information contained in a full-text document is not analyzed if one mines only Medline, and (2) a strategy of document analysis that is purely based on precompiled documents leads to results that are not as up to date as they should be. I sort out both issues by implementing a tool that mines the full-text of publications from different information sources on the fly (see Figure 3). This tool combines a strategy of pre-compilation for documents it has found in the past with an on-the-fly analysis of newly found documents.

To achieve the second specific goal, I implement a tool, CheNER, to identify chemical names in scientific texts. In one of its configurations this tool specifically identifies systematic or IUPAC chemical names (see Figure 4). CheNER was used to participate in the BioCreAtIvE IV challenge, focused in the identification of all type of chemical names.

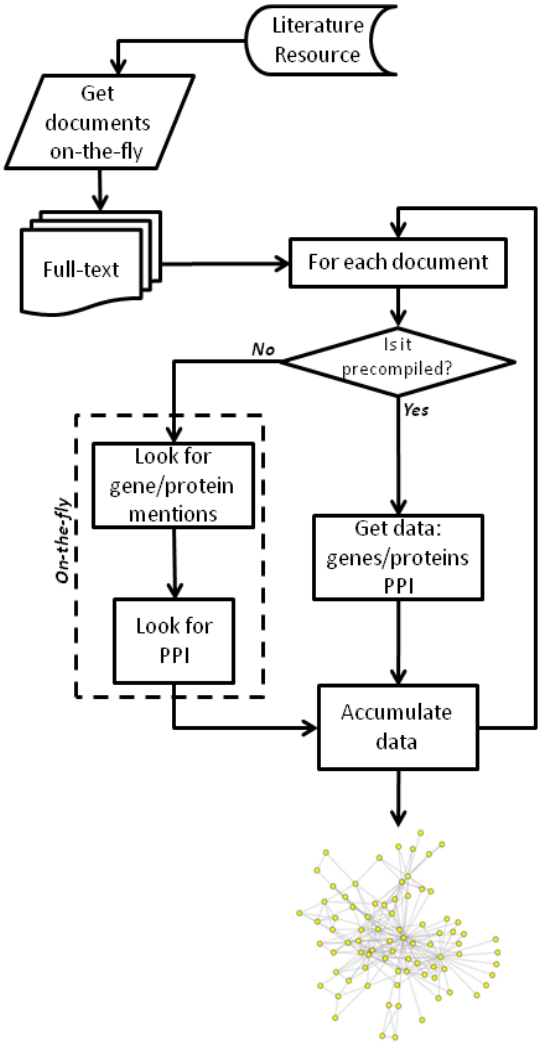


Figure 3. Biblio-MetReS algorithm**.** Workflow of the proposed strategy combining the on-the-fly and the precompiled strategy in one tool.

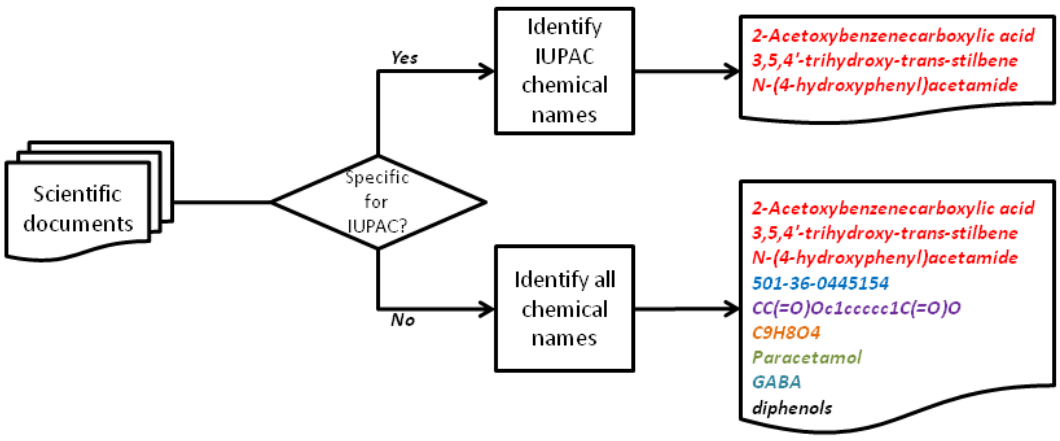


Figure 4. CheNER configuration options. The system allows to identify just IUPAC chemical compound names or all types of chemicals. Each color is used to show different types of chemical nomenclature.

To achieve the third and final specific goal of this thesis, I created a meta-tool, Protein-MetReS that improves the integration of several available servers that perform structural analysis and docking of proteins (see Figure 5).

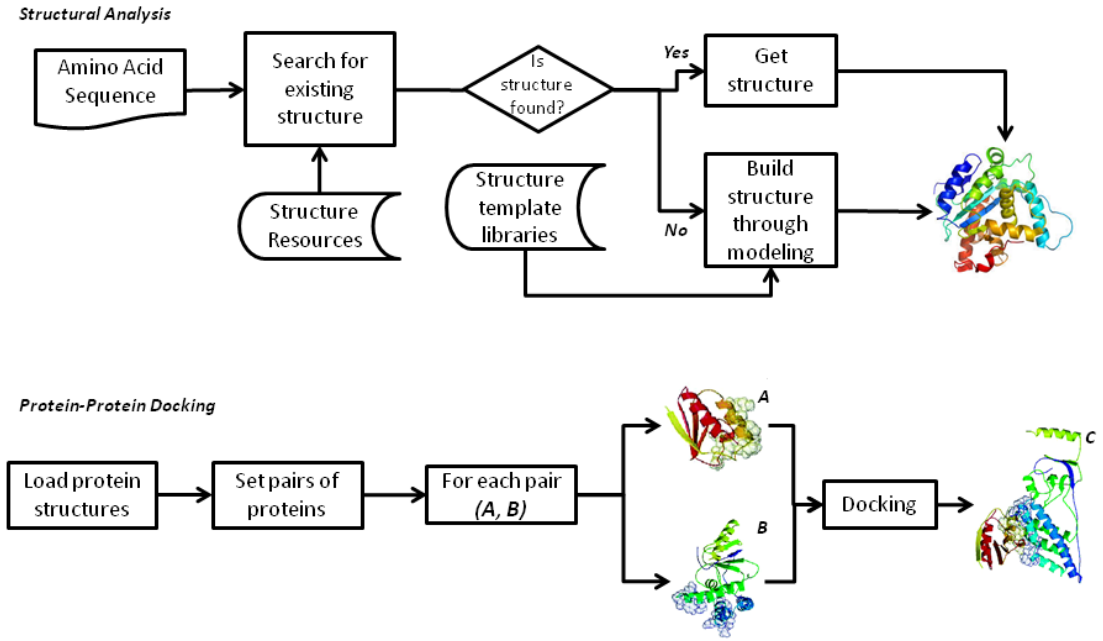


Figure 5. Protein-MetReS modules. The top half of the figure represents the program's structural analysis function that work as follows: if the structure/model for the sequence target already exist within either the PDB or the SWISS-MODEL Repository, load the structure/model,; otherwise, build the model using at least one of the modeling servers integrated in the application (MODELLER, SWISS-MODEL or Phyre2). The bottom half of the figure represents the program's docking functionality that works as follows: at least two protein structures/models must be loaded in order to start execution. If there is more than two structure proteins, then the different combinations of protein pairs are set up. Each pair of structures for which docking predictions are required are submitted to the docking tools (HEX, Gramm-X).

1.3.2 Additional contributions

During this thesis I have contributed in a parallel project of the lab that shared its general goal with this thesis. That parallel project is described in the doctoral thesis of Hiren Karathia [104]. The tool Homol-MetReS was developed in that project. This tool uses comparative functional genomics to facilitate network reconstruction of proteins involved in specific biological processes. Homol-MetReS is a user-friendly web application that facilitates: (1) Management of molecular information for each organism, including protein/gene sequences and function/processes information assigned to the sequence. (2) (Re)assignment and integration of functional information to proteins or genes as per standard terms defined in GO, EC Numbers, KEGG pathways databases

or used defined terms. (3) Creation of organisms - centric clusters of orthologs, homologues and absent genes. (4) Visual comparison between organisms of sets of proteins involved in different specific processes and appropriate choice of model [104]. My contribution to this work focused on the design of both, the database that underlies Homol-MetReS and Biblio-MetReS, and the user interface for Homol-MetReS.

1.4 Organitzation

I now describe the organization of the remaining chapters of this thesis.

Chapter 2 introduces the implementation of Biblio-MetReS, a user-friendly tool developed in order to find interactions between genes/proteins in scientific documents using text-mining techniques. One of the most important aspects is that the identifications of genes is done on-the fly which ensures that the documents found are always up-to-date. Biblio-MetReS provides a paper-centric view and display the interactions found between genes/proteins in an interaction network. Also, in this chapter is discussed the limitations and the possible improvements that can be made.

Chapter 3 describes improvements made to the original version of Biblio-MetReS. In brief, the improvements are related to the reduction of the execution time and the identification of biological processes in order to add more biological information to the interaction network displayed.

Chapter 4 details the methodology used to develop a tool to identify chemical compounds and drugs that is called CheNER. This tool arises from a short stay at CNIO (Centro Nacional de Investigaciones Oncológicas) and it was developed in collaboration with the group of Alfonso Valencia. The configuration of CheNER described in this chapter focuses mostly on the identification of IUPAC or systematic chemicals names.

Chapter 5 introduces the more general configuration of CheNER and describes the participation in the IV BioCreAtIvE Challenge. This configuration implements a software solution to the problem described in Track 2, CHEMDNER, which is focused on the identification of different types of chemical compounds such as commercial numbers, abbreviations, trade names, families, etc.

Chapter 6 details the beta implementation of Protein-MetReS, the tool that integrates different standalone and web server applications for the prediction of protein structures and for protein docking.

Chapter 7 discusses the possible future directions that this work opens.

Finally, the last chapter presents the conclusions of my work .

1.5 References

1. Alves R, Savageau MA (2003) Comparative analysis of prototype two-component systems with either bifunctional or monofunctional sensors: differences in molecular structure and physiological function. Mol Microbiol 48: 25–51.
2. Igoshin OA, Alves R, Savageau MA (2008) Hysteretic and graded responses in bacterial two-component signal transduction. Mol Microbiol 68: 1196–1215. doi:10.1111/j.1365-2958.2008.06221.x.
3. Alves R, Sorribas A (2007) In silico pathway reconstruction: Iron-sulfur cluster biogenesis in Saccharomyces cerevisiae. BMC Syst Biol 1: 10. doi:10.1186/1752-0509-1-10.
4. Ma W, Trusina A, El-Samad H, Lim WA, Tang C (2009) Defining Network Topologies that Can Achieve Biochemical Adaptation. Cell 138: 760–773. doi:10.1016/j.cell.2009.06.013.
5. Savageau MA, Coelho PMBM, Fasani RA, Tolla DA, Salvador A (2009) Phenotypes and tolerances in the design space of biochemical systems. Proc Natl Acad Sci U S A 106: 6435–6440. doi:10.1073/pnas.0809869106.
6. Vilaprinyo E, Alves R, Sorribas A (2006) Use of physiological constraints to identify quantitative design principles for gene expression in yeast adaptation to heat shock. BMC Bioinformatics 7: 184. doi:10.1186/1471-2105-7-184.
7. Zelezniak A, Pers TH, Soares S, Patti ME, Patil KR (2010) Metabolic Network Topology Reveals Transcriptional Regulatory Signatures of Type 2 Diabetes. PLoS Comput Biol 6: e1000729. doi:10.1371/journal.pcbi.1000729.
8. Ulitsky I, Shamir R (2007) Identification of functional modules using network topology and high-throughput data. BMC Syst Biol 1: 8. doi:10.1186/1752-0509-1-8.
9. Savageau MA (1971) Concepts relating the behavior of biochemical systems to their underlying molecular properties. Arch Biochem Biophys 145: 612–621.
10. Savageau MA, Fasani RA (2009) Qualitatively distinct phenotypes in the design space of biochemical systems. FEBS Lett 583: 3914–3922. doi:10.1016/j.febslet.2009.10.073.
11. Hucka M, Finney A, Bornstein BJ, Keating SM, Shapiro BE, et al. (2004) Evolving a lingua franca and associated software infrastructure for computational systems biology: the Systems Biology Markup Language (SBML) project. Syst Biol 1: 41–53.
12. Kitano H (2002) Systems Biology: A Brief Overview. Science 295: 1662–1664. doi:10.1126/science.1069492.
13. Tanaka R, Csete M, Doyle J (2005) Highly optimised global organisation of metabolic networks. Syst Biol 152: 179–184.
14. Gene Ontology Consortium (2004) The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32: 258D–261. doi:10.1093/nar/gkh036.

15. Goll J, Montgomery R, Brinkac LM, Schobel S, Harkins DM, et al. (2010) The Protein Naming Utility: a rules database for protein nomenclature. Nucleic Acids Res 38: D336–339. doi:10.1093/nar/gkp958.
16. Fernández JM, Hoffmann R, Valencia A (2007) iHOP web services. Nucleic Acids Res 35: W21–26. doi:10.1093/nar/gkm298.
17. Beagley N, Stratton KG, Webb-Robertson B-JM (2010) VIBE 2.0: visual integration for bayesian evaluation. Bioinforma Oxf Engl 26: 280–282. doi:10.1093/bioinformatics/btp639.
18. Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, et al. (2009) STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 37: D412–416. doi:10.1093/nar/gkn760.
19. Caspi R, Altman T, Dreher K, Fulcher CA, Subhraveti P, et al. (2012) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res 40: D742–753. doi:10.1093/nar/gkr1014.
20. Okuda S, Yamada T, Hamajima M, Itoh M, Katayama T, et al. (2008) KEGG Atlas mapping for global analysis of metabolic pathways. Nucleic Acids Res 36: W423–426. doi:10.1093/nar/gkn282.
21. Reyes-Palomares A, Montañez R, Real-Chicharro A, Chniber O, Kerzazi A, et al. (2009) Systems biology metabolic modeling assistant: an ontology-based tool for the integration of metabolic data in kinetic modeling. Bioinforma Oxf Engl 25: 834–835. doi:10.1093/bioinformatics/btp061.
22. Alves R, Vilaprinyo E, Sorribas A (2008) Integrating Bioinformatics and Computational Biology: Perspectives and Possibilities for In Silico Network Reconstruction in Molecular Systems Biology. Curr Bioinforma 3: 98–129. doi:10.2174/157489308784340694.
23. Richards D (2006) Medline at thirty-five. Evid Based Dent 7: 89. doi:10.1038/sj.ebd.6400440.
24. Afifi M (2007) PubMed-indexed duplicate publications in the last decade, 1996-2006. Ann Saudi Med 27: 302–304; author reply 304.
25. Falagas ME, Pitsouni EI, Malietzis GA, Pappas G (2008) Comparison of PubMed, Scopus, Web of Science, and Google Scholar: strengths and weaknesses. FASEB J Off Publ Fed Am Soc Exp Biol 22: 338–342. doi:10.1096/fj.07-9492LSF.
26. Gass A, Doyle H (2005) PLOS position on NIH public access policy. Science 308: 356. doi:10.1126/science.308.5720.356a.
27. Butler D (2000) BioMed Central boosted by editorial board. Nature 405: 384. doi:10.1038/35013218.
28. Li Q, Cheng T, Wang Y, Bryant SH (2010) PubChem as a public resource for drug discovery. Drug Discov Today 15: 1052–1057. doi:10.1016/j.drudis.2010.10.003.
29. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, et al. (2007) DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res 36: D901–D906. doi:10.1093/nar/gkm958.

30. Jessop D, Adams S, Willighagen E, Hawizy L, Murray-Rust P (2011) OSCAR4: a flexible architecture for chemical text-mining. J Cheminformatics 3: 41. doi:10.1186/1758-2946-3-41.
31. Rocktäschel T, Weidlich M, Leser U (2012) ChemSpot: A Hybrid System for Chemical Named Entity Recognition. Bioinformatics. Available: http://bioinformatics.oxfordjournals.org/content/early/2012/04/11/bioinformatics.bts183. Accessed 10 October 2012.
32. BioCreAtIvE IV (n.d.). Available: http://www.biocreative.org/events/biocreative-iv/CFP/.
33. Wellcome trust sanger institute (n.d.). Available: http://www.sanger.ac.uk.
34. Kanehisa M (2002) The KEGG database. Novartis Found Symp 247: 91–101; discussion 101–103, 119–128, 244–252.
35. Kanehisa M, Goto S, Hattori M, Aoki-Kinoshita KF, Itoh M, et al. (2006) From genomics to chemical genomics: new developments in KEGG. Nucleic Acids Res 34: D354–357. doi:10.1093/nar/gkj102.
36. Kanehisa M, Goto S, Kawashima S, Nakaya A (2002) The KEGG databases at GenomeNet. Nucleic Acids Res 30: 42–46.
37. Kanehisa M, Goto S, Kawashima S, Okuno Y, Hattori M (2004) The KEGG resource for deciphering the genome. Nucleic Acids Res 32: D277–280. doi:10.1093/nar/gkh063.
38. Nikitin F, Rance B, Itoh M, Kanehisa M, Lisacek F (2004) Using protein motif combinations to update KEGG pathway maps and orthologue tables. Genome Informatics Int Conf Genome Informatics 15: 266–275.
39. Barrett T, Edgar R (2006) Gene expression omnibus: microarray data storage, submission, retrieval, and analysis. Methods Enzymol 411: 352–369. doi:10.1016/S0076-6879(06)11019-8.
40. Edgar R, Barrett T (2006) NCBI GEO standards and services for microarray data. Nat Biotechnol 24: 1471–1472. doi:10.1038/nbt1206-1471.
41. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, et al. (2007) NCBI GEO: mining tens of millions of expression profiles--database and tools update. Nucleic Acids Res 35: D760–765. doi:10.1093/nar/gkl887.
42. Maglott D, Ostell J, Pruitt KD, Tatusova T (2011) Entrez Gene: gene-centered information at NCBI. Nucleic Acids Res 39: D52–57. doi:10.1093/nar/gkq1237.
43. Ostell JM, Kans JA (n.d.) The NCBI Data Model. In: Baxevanis AD, Ouellette BFF, editors. Methods of Biochemical Analysis. Hoboken, NJ, USA: John Wiley & Sons, Inc., Vol. 39. pp. 121–144. Available: http://doi.wiley.com/10.1002/9780470110607.ch6. Accessed 8 August 2013.
44. Pruitt KD, Tatusova T, Maglott DR (2003) NCBI Reference Sequence project: update and current status. Nucleic Acids Res 31: 34–37.
45. Pruitt KD, Tatusova T, Maglott DR (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res 35: D61–65. doi:10.1093/nar/gkl842.

46. Kuroda M, Hiramatsu K (2004) Genome sequencing and annotation: an overview. Methods Mol Biol Clifton NJ 266: 29–45. doi:10.1385/1-59259-763-7:029.
47. Besemer J, Borodovsky M (2005) GeneMark: web software for gene finding in prokaryotes, eukaryotes and viruses. Nucleic Acids Res 33: W451–454. doi:10.1093/nar/gki487.
48. Cawley SL, Pachter L (2003) HMM sampling and applications to gene finding and alternative splicing. Bioinforma Oxf Engl 19 Suppl 2: ii36–41.
49. Chatterji S, Pachter L (2005) Large multiple organism gene finding by collapsed Gibbs sampling. J Comput Biol J Comput Mol Cell Biol 12: 599–608. doi:10.1089/cmb.2005.12.599.
50. Lee D, Redfern O, Orengo C (2007) Predicting protein function from sequence and structure. Nat Rev Mol Cell Biol 8: 995–1005. doi:10.1038/nrm2281.
51. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25: 3389–3402.
52. Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39: W29–W37. doi:10.1093/nar/gkr367.
53. Whisstock JC, Lesk AM (2003) Prediction of protein function from protein sequence and structure. Q Rev Biophys 36: 307–340.
54. Devos D, Valencia A (2000) Practical limits of function prediction. Proteins 41: 98–107.
55. Berman HM, Bhat TN, Bourne PE, Feng Z, Gilliland G, et al. (2000) The Protein Data Bank and the challenge of structural genomics. Nat Struct Biol 7 Suppl: 957–959. doi:10.1038/80734.
56. Murzin AG, Brenner SE, Hubbard T, Chothia C (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247: 536–540. doi:10.1006/jmbi.1995.0159.
57. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, et al. (1997) CATH--a hierarchic classification of protein domain structures. Struct Lond Engl 1993 5: 1093–1108.
58. Maietta P, Lopez G, Carro A, Pingilley BJ, Leon LG, et al. (2013) FireDB: a compendium of biological and pharmacologically relevant ligands. Nucleic Acids Res. Available: http://nar.oxfordjournals.org/lookup/doi/10.1093/nar/gkt1127. Accessed 27 December 2013.
59. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, et al. (2007) Comparative protein structure modeling using MODELLER. Curr Protoc Protein Sci Editor Board John E Coligan Al Chapter 2: Unit 2.9. doi:10.1002/0471140864.ps0209s50.
60. Sánchez R, Sali A (1997) Evaluation of comparative protein structure modeling by MODELLER-3. Proteins Suppl 1: 50–58.

61. Bordoli L, Kiefer F, Arnold K, Benkert P, Battey J, et al. (2009) Protein structure homology modeling using SWISS-MODEL workspace. Nat Protoc 4: 1–13. doi:10.1038/nprot.2008.197.
62. Arnold K, Bordoli L, Kopp J, Schwede T (2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinforma Oxf Engl 22: 195–201. doi:10.1093/bioinformatics/bti770.
63. Kiefer F, Arnold K, Künzli M, Bordoli L, Schwede T (2009) The SWISS-MODEL Repository and associated resources. Nucleic Acids Res 37: D387–392. doi:10.1093/nar/gkn750.
64. Peitsch MC (1995) Protein Modeling by E-mail. Bio/Technology 13: 658–660. doi:10.1038/nbt0795-658.
65. Kelley LA, Sternberg MJE (2009) Protein structure prediction on the Web: a case study using the Phyre server. Nat Protoc 4: 363–371. doi:10.1038/nprot.2009.2.
66. Zhang Z (2003) An Overview of Protein Structure Prediction: From Homology to Ab Initio. Available: http://biochem218.stanford.edu/Projects.html.
67. Kim DE, Chivian D, Baker D (2004) Protein structure prediction and analysis using the Robetta server. Nucleic Acids Res 32: W526–531. doi:10.1093/nar/gkh468.
68. Rohl CA, Strauss CEM, Misura KMS, Baker D (2004) Protein structure prediction using Rosetta. Methods Enzymol 383: 66–93. doi:10.1016/S0076-6879(04)83004-0.
69. Goldman BB, Wipke WT (2000) QSD quadratic shape descriptors. 2. Molecular docking using quadratic shape descriptors (QSDock). Proteins 38: 79–94.
70. Meng EC, Shoichet BK, Kuntz ID (1992) Automated docking with grid-based energy evaluation. J Comput Chem 13: 505–524. doi:10.1002/jcc.540130412.
71. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, et al. (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J Comput Chem 19: 1639–1662. doi:10.1002/(SICI)1096-987X(19981115)19:14<1639::AID-JCC10>3.0.CO;2-B.
72. Feig M, Onufriev A, Lee MS, Im W, Case DA, et al. (2004) Performance comparison of generalized born and Poisson methods in the calculation of electrostatic solvation energies for protein structures. J Comput Chem 25: 265–284. doi:10.1002/jcc.10378.
73. Halperin I, Ma B, Wolfson H, Nussinov R (2002) Principles of docking: An overview of search algorithms and a guide to scoring functions. Proteins 47: 409–443. doi:10.1002/prot.10115.
74. Ritchie DW (2005) High-order analytic translation matrix elements for real-space six-dimensional polar Fourier correlations. J Appl Crystallogr 38: 808–818. doi:10.1107/S002188980502474X.
75. Mustard D, Ritchie DW (2005) Docking essential dynamics eigenstructures. Proteins 60: 269–274. doi:10.1002/prot.20569.

76. Ritchie DW, Kemp GJ (2000) Protein docking using spherical polar Fourier correlations. Proteins 39: 178–194.
77. Tovchigrechko A, Vakser IA (2006) GRAMM-X public web server for protein-protein docking. Nucleic Acids Res 34: W310–314. doi:10.1093/nar/gkl206.
78. Tovchigrechko A, Vakser IA (2005) Development and testing of an automated approach to protein docking. Proteins 60: 296–301. doi:10.1002/prot.20573.
79. Lyskov S, Gray JJ (2008) The RosettaDock server for local protein-protein docking. Nucleic Acids Res 36: W233–238. doi:10.1093/nar/gkn216.
80. Lyskov S, Chou F-C, Conchúir SÓ, Der BS, Drew K, et al. (2013) Serverification of molecular modeling applications: the Rosetta Online Server that Includes Everyone (ROSIE). PloS One 8: e63906. doi:10.1371/journal.pone.0063906.
81. Leinonen R, Diez FG, Binns D, Fleischmann W, Lopez R, et al. (2004) UniProt archive. Bioinformatics 20: 3236–3237. doi:10.1093/bioinformatics/bth191.
82. The UniProt Consortium (2007) The Universal Protein Resource (UniProt). Nucleic Acids Res 35: D193–D197. doi:10.1093/nar/gkl929.
83. Apweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, et al. (2000) InterPro--an integrated documentation resource for protein families, domains and functional sites. Bioinformatics 16: 1145–1150. doi:10.1093/bioinformatics/16.12.1145.
84. Mulder N, Apweiler R (2007) InterPro and InterProScan: tools for protein sequence classification and comparison. Methods Mol Biol Clifton NJ 396: 59–70. doi:10.1007/978-1-59745-515-2_5.
85. Hunter S, Apweiler R, Attwood TK, Bairoch A, Bateman A, et al. (2009) InterPro: the integrative protein signature database. Nucleic Acids Res 37: D211–D215. doi:10.1093/nar/gkn785.
86. Hunter S, Jones P, Mitchell A, Apweiler R, Attwood TK, et al. (2011) InterPro in 2011: new developments in the family and domain prediction database. Nucleic Acids Res 40: D306–D312. doi:10.1093/nar/gkr948.
87. Gasteiger E, Gattiker A, Hoogland C, Ivanyi I, Appel RD, et al. (2003) ExPASy: The proteomics server for in-depth protein knowledge and analysis. Nucleic Acids Res 31: 3784–3788.
88. Artimo P, Jonnalagedda M, Arnold K, Baratin D, Csardi G, et al. (2012) ExPASy: SIB bioinformatics resource portal. Nucleic Acids Res 40: W597–W603. doi:10.1093/nar/gks400.
89. Zanzoni A, Montecchi-Palazzi L, Quondam M, Ausiello G, Helmer-Citterich M, et al. (2002) MINT: a Molecular INTeraction database. FEBS Lett 513: 135–140.
90. Persico M, Ceol A, Gavrila C, Hoffmann R, Florio A, et al. (2005) HomoMINT: an inferred human network based on orthology mapping of protein interactions discovered in model organisms. BMC Bioinformatics 6 Suppl 4: S21. doi:10.1186/1471-2105-6-S4-S21.

91. Ceol A, Chatr-aryamontri A, Santonico E, Sacco R, Castagnoli L, et al. (2007) DOMINO: a database of domain-peptide interactions. Nucleic Acids Res 35: D557–560. doi:10.1093/nar/gkl961.
92. Xenarios I, Salwínski L, Duan XJ, Higney P, Kim S-M, et al. (2002) DIP, the Database of Interacting Proteins: a research tool for studying cellular networks of protein interactions. Nucleic Acids Res 30: 303–305.
93. Rho S, You S, Kim Y, Hwang D (2008) From proteomics toward systems biology: integration of different types of proteomics data into network models. BMB Reports 41: 184–193.
94. Oliver SG, Winson MK, Kell DB, Baganz F (1998) Systematic functional analysis of the yeast genome. Trends Biotechnol 16: 373–378.
95. Shulaev V (2006) Metabolomics technology and bioinformatics. Brief Bioinform 7: 128–139. doi:10.1093/bib/bbl012.
96. Hall RD (2006) Plant metabolomics: from holistic hope, to hype, to hot topic. New Phytol 169: 453–468. doi:10.1111/j.1469-8137.2005.01632.x.
97. Metabolomics: Current analytical platforms and methodologies (2005). TrAC Trends Anal Chem 24: 285–294. doi:10.1016/j.trac.2004.11.021.
98. Moradi F, Buračas GT, Buxton RB (2012) Attention strongly increases oxygen metabolic response to stimulus in primary visual cortex. Neuroimage 59: 601–607. doi:10.1016/j.neuroimage.2011.07.078.
99. McKelvie JR, Whitfield Åslund M, Celejewski MA, Simpson AJ, Simpson MJ (2013) Reduction in the earthworm metabolomic response after phenanthrene exposure in soils with high soil organic carbon content. Environ Pollut Barking Essex 1987 175: 75–81. doi:10.1016/j.envpol.2012.12.018.
100. Naranjo CA, Busto U, Sellers EM, Sandor P, Ruiz I, et al. (1981) A method for estimating the probability of adverse drug reactions. Clin Pharmacol Ther 30: 239–245. doi:10.1038/clpt.1981.154.
101. Dai JL (2005) Serial analyses of gene expression (SAGE). Methods Mol Med 103: 161–174.
102. Knox DP, Skuce PJ (2005) SAGE and the quantitative analysis of gene expression in parasites. Trends Parasitol 21: 322–326. doi:10.1016/j.pt.2005.05.011.
103. Khatri P, Draghici S (2005) Ontological analysis of gene expression data: current tools, limitations, and open problems. Bioinformatics 21: 3587–3595. doi:10.1093/bioinformatics/bti565.
104. Karathia H (2012) Development and application of computational methdologies for Integrated Molecular Systems Biology. Available: http://www.tdx.cat/handle/10803/110518.

Chapter 2. Biblio-MetReS I

BMC Bioinformatics 2011, **12**:387
doi:10.1186/1471-2105-12-387

Biblio-MetReS: A bibliometric network reconstruction application and server

Anabel Usié, Hiren Karathia, Ivan Teixidó, Joan Valls, Xavier Faus, Rui Alves and Francesc Solsona

Abstract

Background: Reconstruction of genes and/or protein networks from automated analysis of the literature is one of the current targets of text mining in biomedical research. Some user-friendly tools already perform this analysis on precompiled databases of abstracts of scientific papers. Other tools allow expert users to elaborate and analyze the full content of a corpus of scientific documents. However, to our knowledge, no user friendly tool that simultaneously analyzes the latest set of scientific documents available on line and reconstructs the set of genes referenced in those documents is available.

Results: This article presents such a tool, Biblio-MetReS, and compares its functioning and results to those of other user-friendly applications (iHOP, STRING) that are widely used. Under similar conditions, Biblio-MetReS creates networks that are comparable to those of other user friendly tools. Furthermore, analysis of full text documents provides more complete reconstructions than those that result from using only the abstract of the document.

Conclusions: Literature-based automated network reconstruction is still far from providing complete reconstructions of molecular networks. However, its value as an auxiliary tool is high and it will increase as standards for reporting biological entities and relationships become more widely accepted and enforced. Biblio-MetReS is an application that can be downloaded from http://metres.udl.cat/. It provides an easy to use environment for researchers to reconstruct their networks of interest from an always up to date set of scientific documents.

2.1 Background

Reconstructing molecular networks that are responsible for regulating biological processes is a fundamental task in molecular biology, if one is to understand how the different components of those networks contribute to each process. In recent years many alternative types of methods have been proposed to achieve such a reconstruction [1,2]. One type of method relies on the automated analysis of published literature to identify genes and proteins that co-occur in the same document(s) [3–11]. It has been assumed that if two genes or proteins are cited in the same document, there is the likelihood that they functionally interact. In fact, many algorithms, methods and tools have been proposed and implemented in order to reconstruct the network of genes associated with a given gene of interest, by automated mining of the published literature [3–31].

Only a small number of these tools are more widely cited (and likely used) by molecular biologists (Table 1). Out of these, iHOP [3] and STRING [5] have a usage that is at least one order of magnitude higher than that of other applications, as estimated by the number of times that the different applications are cited (Table 1). These two web servers preprocess documents that are published in Medline and PubMed, looking for words that match the names of genes from the different organisms in the web server's database. Once they have identified the genes that co-occur in those documents, they provide different functionality to the user. While iHOP allows the user to choose exactly which genes s/he wants to add to the interaction network, STRING automatically establishes a threshold score above which all genes are included in the model for the network.

A shortcoming of both these tools is that, in terms of literature, they only analyze the information contained in Medline or PubMed abstracts and their databases require constant update. Given that policies for publication and access to scientific papers are changing and, as a consequence, an increasing number of scientific publications are becoming freely available over the

internet, iHOP and STRING ignore a growing source of information about possible interactions between genes [20,32,33].

Table 1. Number of citations for text mining programs in the Web of Science database as of June 2011

Program	Total Number of Citations
STRING	949
iHOP	274
Whatizit	41
Alibaba	37
Reflect	16
iProLink	11
SciMiner	4
BioLMiner	1
Linguamatics I2E	1
Akane RE	0
Laitor	0
PahtText	0

Currently, other tools that analyze full documents without pre-processing in order to reconstruct molecular gene networks are either still experimental, applicable only to a document or documents supplied by the user or present in PubMed [6,9,11,34,35] and/or require a high level of computational expertise for their use [6,34,35].

Thus, there is a need for a tool that a) analyzes full documents as they are made available on the world wide web and before they are included in databases such as PubMed, b) analyzes documents and literature corpora that have not been manually annotated, and c) is user-friendly. We developed Biblio-MetReS http://metres.udl.cat/ to meet these demands, allowing for an on-the-run full text analysis for automated reconstruction of literature gene/protein networks in an intuitive way. Biblio-MetReS relies on a database that contains lists of all annotated genes of organisms with fully sequenced genomes from the KEGG database. The tool allows users to select different sources of information from where to compile data for the reconstruction of the molecular networks responsible for regulating and executing biological processes.

Here we present the tool and benchmark it against STRING and iHOP, using genes that participate in well characterized metabolic processes of organisms with fully sequenced genomes. The three tools have comparable results when Biblio-MetReS searches are limited to Medline. When this limitation is removed, Biblio-MetReS finds networks that are more complete than those found by iHOP and STRING.

2.2 Implementation

2.2.1 Underlying database & Biblio-MetReS implementation

Biblio-MetReS relies on an in-house database of organisms and genes that was built using the list of organisms with fully sequenced genomes available in KEGG [36]. The database of gene names and their synonyms is built and regularly updated by matching the KEGG gene names and synonyms to their NCBI [37] names and synonyms, followed by removing of redundant terms. The databases are implemented using Zope technology, which is based on MySQL and Python.

The application itself was implemented in JAVA, using the NetBeans IDE. Swing was used to implement the Graphical User Interface (GUI). Swing was also used to create the parsers for the different documents to be analyzed, with the exception of PDF files. These files are parsed using the PDFBox library. We implemented parsers for HTML documents, PDF and ASCII. HTML documents are transformed into plain text as follows: paragraphs are detected in the HTML code, using a parsing library to navigate through the tags, followed by extraction of the text within those tags. PDF documents are transformed into plain text using the Pdfbox library, which extracts the text within the document while ignoring the images. Once the text is extracted, we parse for paragraphs by looking for punctuation signs that signal the end of a sentence followed by the new line escape character. These punctuation signs are used to split sentences, controlling to make sure that we are not splitting decimal figures, e-mail addresses, web pages, and others.

The results are stored in a file with XML format that is generated at the end of each search. The processing of the XML files is done using the JDOM API. The JGraph API is used for the graphical representation of the network results in 2D

2.2.2 Document search analysis

Biblio-MetReS implements a meta-search engine that compiles results from the search engines selected by the user (see Figure 1, panel 3 for a list of document sources). The search that is launched to each search engine includes all genes selected by the user, as well as the name of the organism of interest. As the search is completed by the relevant search motor (or motors if the user selected more than one data source), Biblio-MetReS collects the URLs of all documents found by each of the search engines. The treatment of these URLs goes as follows. First, the application eliminates redundant URLs. Then, for results from scientific databases and journals, it analyzes the doi number for each document, eliminating further duplicates. When the non-redundant list of documents is ready, Biblio-MetReS identifies if the full text of the document is freely available (either because the text is free to all users or because the institution providing the web connection has access agreements with the content provider) or if it is protected. In the latter case, the application discards this document and analyzes only the freely available abstract. Once all this pruning procedure is done, the application analyzes each document in search of co-occurrence of any genes or proteins in sentences, paragraphs and entire documents. Exact name matching is used and all synonyms for a gene are searched for. The dictionary of synonyms we use is a merge from those of NCBI and KEGG.

The analysis of co-occurrence in sentences and paragraphs is done because, when analyzing the full text of scientific documents, one must consider some form of proximity measurement. Otherwise, the co-occurrence of genes in different sections of the same document will introduce a significant amount of noise in the network of possible interactions [20].

2.2.3 Metrics

We need to define appropriate metrics in order to provide some degree of biological significance to the fact that if two genes or proteins co-occur in a document they do not do so by pure chance. To do this we consider different aspects of co-occurrence.

First, we measure how frequently the different proteins or gene pairs co-occur in sentences, paragraphs and/or documents. We then take the odds ratio of the frequency of occurrences in the first two categories with respect to that of the third. The closer to one these odds ratios are, the more frequent it is that both genes are mentioned only in the same sentences or paragraphs of a document, rather than appearing haphazardly in different sections of the text.

Second, we calculate how much information we gain by having the two genes co-occur, when compared to the individual occurrences of the two genes. To estimate this we use information theory. The individual probability of occurrence of a gene is denoted as $p(Gi)$ and it is formally defined as $p(Gi)=\frac{a}{n}$, where $\boldsymbol{a}$ is the number of documents where gene i appears, and $\boldsymbol{n}$ is the total number of documents.

The joint probability of co-occurrence of two genes, $p(Gi,Gj)$, is defined as $p(Gi,Gj)=\frac{b}{n}$, where $\boldsymbol{b}$ is the number of documents where genes i and j simultaneously appear, and $\boldsymbol{n}$ is the total number of documents.

Having defined how to calculate the various probabilities, the mutual information, $MI(Gi,Gj)$ is calculated as follows:

$$MI(Gi,Gj) = p(Gi,Gj)\log\left(\frac{p(Gi,Gj)}{p(Gi)p(Gj)}\right)$$

where the applied logarithm is in natural base.

Finally, and in order to attribute some form of statistical significance to the co-occurrence of a pair of genes, we analyze contingency tables for those co-

occurrences. The analysis is as follows. Consider a set of n sentences (paragraphs, documents) [1 ..., n]. For a given gene k define

$$y_{ik}\begin{cases}1 \Leftarrow \text{gene } \boldsymbol{k} \text{ occurs in sentences} \\ \quad\quad \text{(paragraph,document) } \boldsymbol{i} \\ 0 \Leftarrow \text{otherwise}\end{cases}$$

Now, for genes k1 and k2 define

$$\phi_{k1,k2} = y_{i,k1} y_{i,k2}$$

which has value 1 when both genes co-occur and 0 otherwise.

Both these variables have a Bernoulli distribution. If the occurrence of genes k1 and k2 is independent, then $p(\varphi_{k1,k2}) = p(y_{k1})(y_{k2})$ would be expected, where $p(y_{k\cdot})$ is the relative frequency of occurrence of gene $y_{k\cdot}$ and $p(\varphi_{k1,k2})$ is the relative frequency of co-occurrence of genes k1 and k2 in the total number n of sentences (paragraphs, documents). Then, a Pearson statistic can be used to test for independence of occurrence between k1 and k2 by comparing the observed frequencies $n_1 = p(\varphi_{k1,k2})n$, and $n_2 = (1 - p(\varphi_{k1,k2}))n$ with the expected frequencies under the null hypothesis of independence, which would be $m_1 = p(y_{k1})p(y_{k2})n$ and $m_2 = (1 - p(y_{k1})p(y_{k2}))n$. The Pearson statistic is computed as follows:

$$X^2 = \sum_{i=1}^{2} \frac{(n_i - m_i)^2}{m_i}$$

This statistics follows a chi-square distribution with one degree of freedom, i.e. $\chi_1^2 \sim \chi^2$; hence, the p-value can be calculated as $p = \Pr(\chi_1^2 > \chi^2)$ to assess whether the observed co-occurrence is higher than the one expected by pure chance.

2.3 Results

2.3.1 The workflow

Figure 1 summarizes the workings of Biblio-MetReS. For security reasons users need to register before their first use, in order for the application to be able to access the central database. Once they have registered and logged in, an

organism is chosen to work with. The application loads all genes from this organism that are present in the central database. Once the loading is finished, the user is presented with a window where s/he has to select the data sources for the analysis as well as the genes that will start the analysis. There are three types of data sources to choose from: General Engines (Yahoo, ...), Literature Database (Medline, ...) and Journals (Nature, ...). Once the choices are made and the search is started, the tool identifies the documents that contain the gene names provided by the user and their synonyms. Then, it extracts the full text from each document, and analyses for the co-occurrence of any pair of genes from the organism. All this processing is done on the fly.

The results of the analysis are presented to the user in several forms (Figure 1). First, Biblio-MetReS provides identifying information about each document that it analyzed, together with a list of links to those documents. If the user clicks on any of these links, the documents will open in their default browser. The user is also provided with a list of all genes and gene pairs that were found in each document.

Second, Biblio-MetReS presents the results of co-occurrence as tables. In these tables, the program provides information about absolute and relative frequencies of gene co-occurrence, linked to mutual information and p-values. The tables also provide links to gene and pathway information from other databases.

Third, the results are also presented as two graphs. These graphs provide alternative representations of co-occurrence. One graph presents the co-occurrence of genes in sentences, while the other presents the co-occurrence of genes in paragraphs and documents. In these graphs, each node or vertex is a gene/protein and each edge refers to the interaction between genes/proteins. The thickness of the edge is proportional to the mutual information between two genes and the colour of the edge is proportional to the p-value for the co-occurrence between the two genes or proteins. The colour scale changes in a continuous manner between red (non-significant) and green (significant).

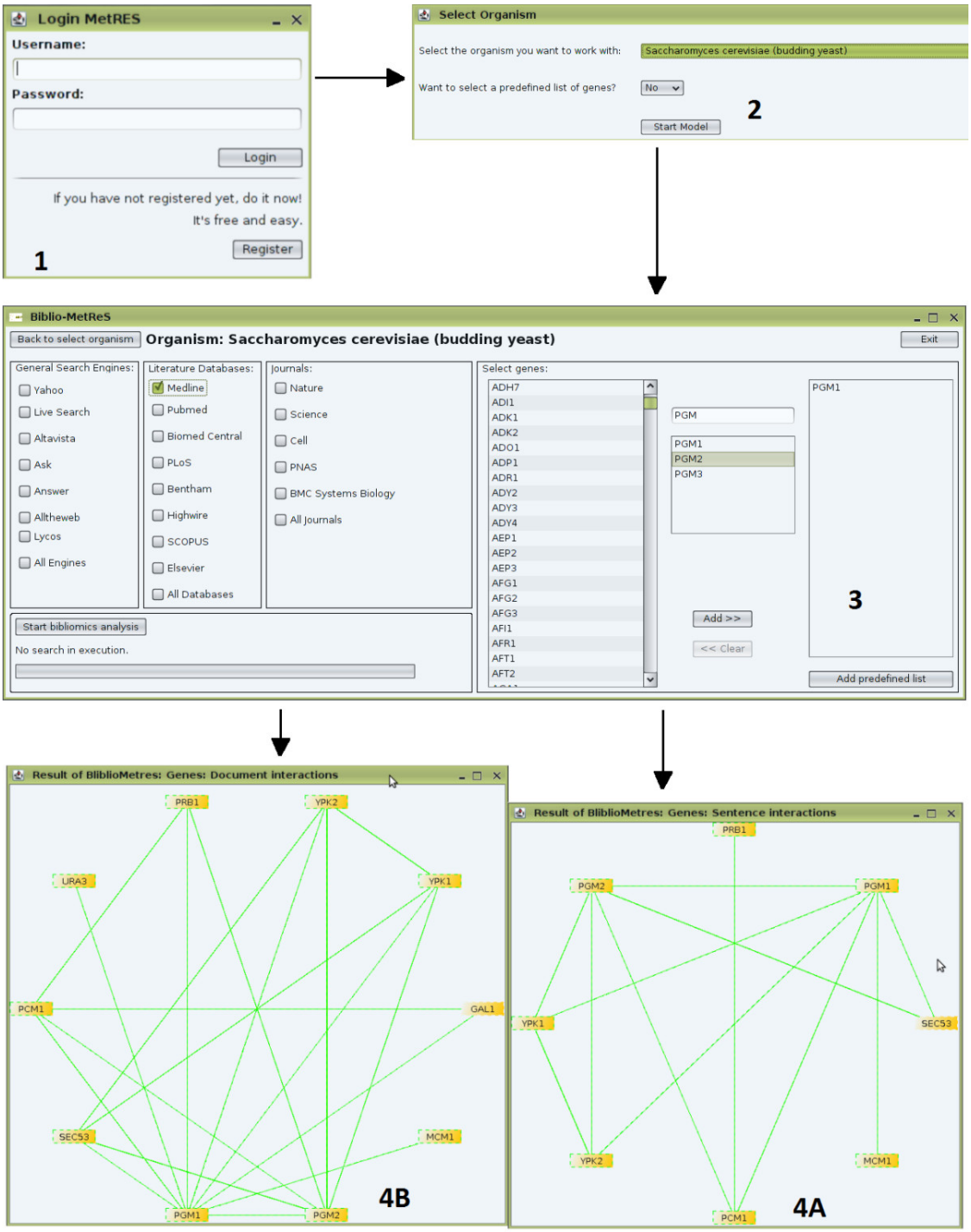


Figure 1. Workflow of Biblio-MetReS. The user registers, logs in (Panel 1) and selects an organism (Panel 2). Once selected, the program either loads the full list of genes from which the user will select the genes to analyze (Panel 3) or allows the user to directly insert the genes s/he wants to analyze (Panel 2.1). Then, the user must select the databases and web searchers s/he wants to use (Panel 3). The program then starts the search and when finished, it generates a series of outputs for the results. First, the list of documents that was analyzed is shown, together with links to the document and to the list of genes fount in each document (Panel 4A). Second, a list of all genes that were found is given (Panel 4B). Each gene is linked to its KEGG webpage, where the gene is associated with other databases and biological pathways. Third, a tabular analysis of gene co-occurrence is shown

2.3.2 Comparing Biblio-MetReS to iHOP and STRING

Given that Biblio-MetReS is intended for an audience similar to that of iHOP and STRING, we need to compare how the results of the three tools differ amongst each other. To do this, we selected three pathways described in KEGG for four different organisms (Supplementary Table 1). In each organism, and starting from a set of three or four genes per pathway, we performed a network reconstruction for each of the three pathways under different conditions (Supplementary Table 1).

iHOP and STRING only search Medline or PubMed abstracts that are pre-processed and stored internally by each program. Because of this, a comparison between the results of these applications and those from Biblio-MetReS require that the set of documents analyzed by Biblio-MetReS is restricted to those contained in Medline. Furthermore, because Biblio-MetReS always analyzes the most recent update of Medline at NCBI, it was run to analyze only the 20 most relevant abstracts from Medline, to avoid an unfair advantage. Our analysis led to the following observations.

First, Biblio-MetReS, iHOP and STRING generate different results, even though the literature corpus that they analyze is, in principle, the same (Figure 2, Supplementary Figure 1). This is likely to be the result of a) different processing of PubMed abstracts (either because the two tools update their databases at different times or because they process abstract content differently), and b) dictionaries that provide synonyms to the standard gene names that do not fully overlap in each of the three tools. In particular STRING uses internal precompiled synonym dictionaries, iHOP uses Entrez Genes, FlyBase, UniProt and the classification from the HUGO nomenclature Committee, and Biblio-MetReS uses KEGG, UniProt and NCBI nomenclature. We cannot control or further investigate a), as this would require access to the inner workings of each program. However, we controlled for b) by checking by hand if all genes we found in one dataset had synonyms in the other two or not, but many of the differences remained (Figure 2).

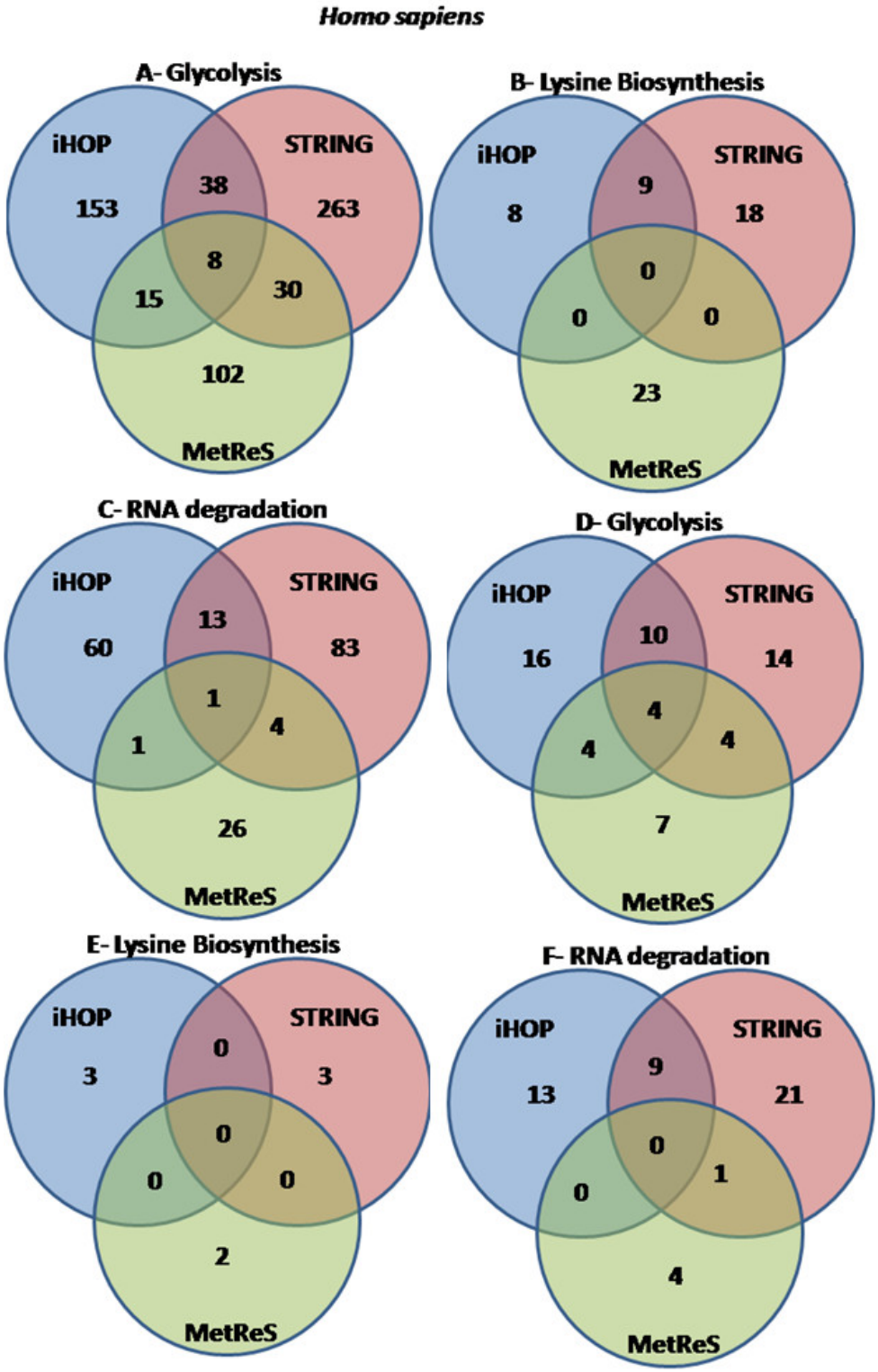


Figure 2. Comparison of results between Biblio-MetReS, iHOP and STRING. Representation of the number of common genes found for the different pathways in Homo sapiens using Biblio-MetReS, iHOP and STRING. This figure shows all genes found for each test. Additional File 3 shows the results for the other organisms, as well as for the genes that are not considered to be in the canonical pathways. A.- Glycolysis, all genes. B.- Lysine metabolism, all genes. C.- RNA degradation, all genes. D.-.Glycolysis, only genes known to belong to the canonical pathway. E.- Lysine metabolism, only genes known to belong to the canonical pathway. F.- RNA degradation, only genes known to belong to the canonical pathway.

Second, even with the self-imposed limitation of using only the 20 more relevant abstracts, Biblio-MetReS always found a number of genes that is comparable to that found by either iHOP or STRING (Figure 2, Supplementary Figure 1).

Third, and as a way to control for the quality of the result from each program, we analyzed how many of the genes that are found by each application are known to be a part of the pathways, as defined in KEGG. No applications find all genes that are associated with the different pathways. In fact, only between 5% and 30% of all genes that were found by the three applications are annotated in KEGG as being a part of the relevant canonical pathway. The application that finds the largest number of genes associated with a canonical pathway varies and is case-dependent (compare Supplementary Figure 1 and Supplementary Figure 2). No single application performs best neither in all pathways of a given organism nor in all organisms for a single pathway. In addition, all application finds several genes that are not associated with the canonical KEGG pathways but co-occur with pathway genes in the literature. In fact between 70% and 95% of all genes identified by iHOP, STRING, or Biblio-MetReS belong to this category. This reveals one of the benefits of these applications, that of finding associations that are not commonly considered. However, this benefit is also associated with the risk of misidentification of functionally interacting genes (see below).

2.3.3 Contribution of different data sources

Given that one of the added values of Biblio-MetReS is its capacity to search and analyze full text documents, we tested how different sources of information added to the number of genes that were found. In these tests, we use the different types of source information ("Literature Databases", "Journals" and "General Engines") in order to find out how much information the different sources add to the reconstruction process. Supplementary Table 1 contains a summary of the tests performed for this analysis.

First, our results suggest that using general search engines for this type of network reconstruction should be done sparingly, if at all. In every test case these engines found files with the entire fully annotated set of genes from the relevant organism. This means that the sensitivity of these search engines for the job of finding co-occurring genes in documents is very high. However, their selectivity is null. Therefore, we do not recommend using these engines when reconstructing a gene network. Because of this we performed the remainder of the benchmark tests using only the search engines from the Literature Databases and Journals panes of Biblio-MetReS (see Figure 1 panel 3).

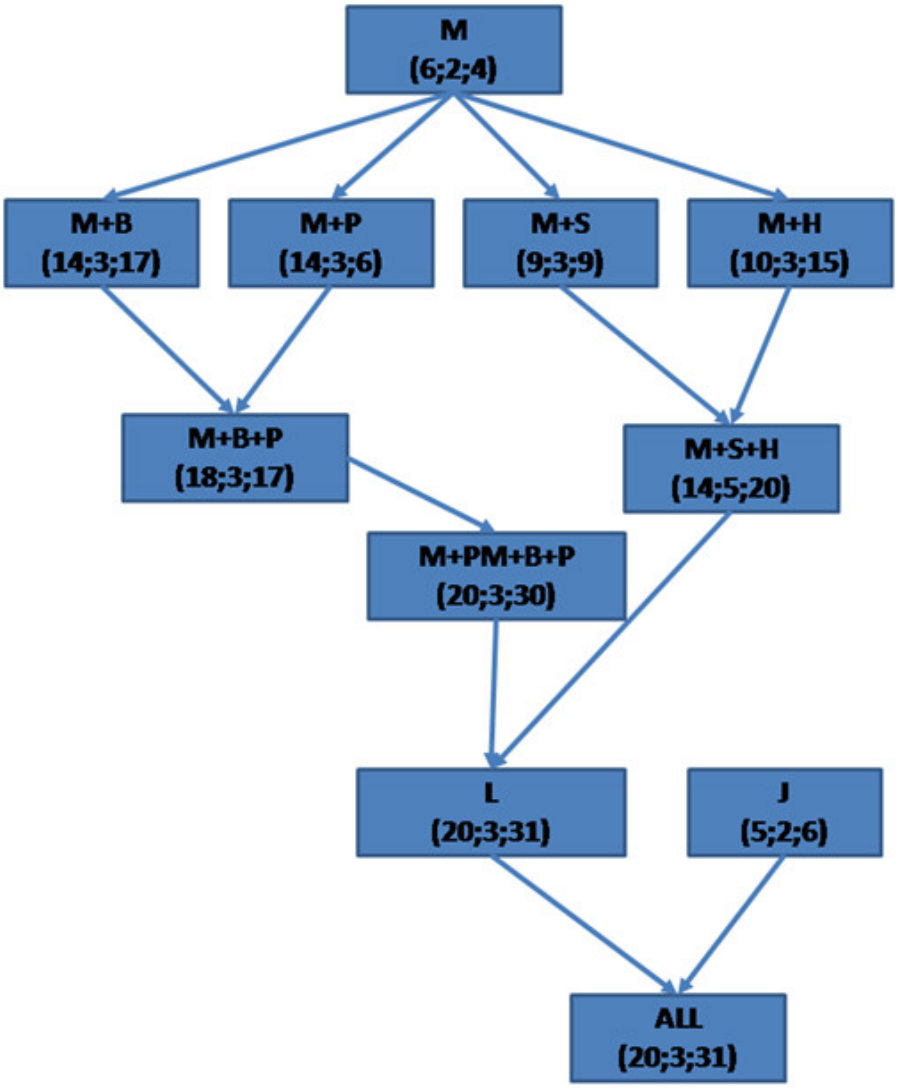


Figure 3. Homo sapiens: Representation of the number of additional genes found by Biblio-MetReS that are known to belong to the canonical pathways under analysis, as we add more data sources to Medline. Each panel shows three numbers in each square. The first number represents the number of genes found for glycolysis. The second number shows the number of genes found for lysine metabolism. The third number shows the number of genes found for RNA degradation. M: Medline, B:Biomed Central, P: PLoS, S: Scopus, H: Highwire, PM: PubMed, L: All literature databases, J: All journals and ALL: all information sources.

Second, we compared the sensitivity of Biblio-MetReS using different databases for scientific documents (Figure 3, Supplementary Figure 3 and Supplementary Figure 4). In general, Medline is the database in which a smaller number of genes is found. When Medline analysis is compared to analysis of databases containing the full text of scientific papers from individual journals or publishing houses, more genes that belong to the relevant pathways are almost always found in the latter case. This suggests that, many times, the information gain provided by analyzing the full text of scientific papers of a given publisher more than offsets the loss of information caused by only having access to a fraction of the scientific literature.

Nevertheless, as is the case when comparing iHOP, STRING and Biblio-MetReS (using Medline), each literature database generates a set of genes that, in many cases, is only partially overlapping. Therefore, we analyzed how much is gained by combining the different literature sources. Supplementary Figure 3 and Supplementary Figure 4 summarize the results of this analysis.

We find that, in general, searching the set of individual journals that we include in Biblio-MetReS discovers a smaller number of gene interactions than using Medline. We also find that, as we combine larger databases, the number of genes that belong to the network of interest increases. However, so does the number of genes that are not recognized by KEGG as being associated with the pathway. In general, a search in literature databases identifies all the genes that are also identified when searching the set of individual journals. However, in some cases, the sets of genes found in the two types of databases are absolutely complementary. This is the case of the genes for glycolysis in *Drosophila melanogaster*.

Another aspect of interest that needs to be analyzed is that of discrimination between genes that are known to belong to the different canonical pathways under analysis and genes whose association to those of the pathway is indirect. Supplementary Figure 3 shows how many of the genes found by Biblio-MetReS are annotated as belonging to the relevant pathways in KEGG. For example, compare the squares marked M (Medline) in each panel of Supplementary

Figure 4 to the subsequent squares in the same panel. You can see that Biblio-MetReS now finds between 1.5 and 6-7 times more genes associated with the canonical pathway than any of the applications in benchmark 1. In contrast, Supplementary Figure 4 shows the total number of genes found during the analysis. We find that most of the genes that are found by the program in the different combinations of databases are not directly associated with the canonical pathway being tested. This was also the case in the first benchmark tests for the three applications being compared (Biblio-MetReS, iHOP, and STRING). The percentage of the total genes that are outside the canonical pathway increases with the number of documents being analyzed.

One way to filter many of the interactions with additional genes that may be irrelevant is by analyzing the graph of genes that co-occur in sentences. The sentence co-occurrence network has a much smaller number of interactions between genes (compare panels 4A and 4B in Figure 1). These interactions are enriched in interactions between genes that belong to the canonical pathway. Furthermore, it is easier for the user to identify if a gene association in this network is important for the work at hand, because Biblio-MetReS shows the relevant sentences.

2.4 Discussion

Automated text mining efforts with the goal of extracting biological information is a booming field. Many issues still need to be solved in order for this extraction to be as good as it can be. On one hand, reporting of biological entities and concepts still needs to be standardized and standards need to be fully accepted and implemented by both journals and researchers. On the other, more efficient methods also need to be developed. The BioCreAtIvE challenge has been established to evaluate how well the different methods perform in both identifying biological entities and relationships between these entities [38]. The BioCreAtIvE challenge, as any control experiment should do, performs an evaluation of different tools in well curated datasets. However, while more developed methods are being further developed, biological researchers can still

benefit from prototypical applications that assist them in many the large majority of the scientific literature, which is not curated at all. Efforts to mine this body of literature in order to reconstruct networks of interacting genes started as early as in the end of the nineties [39]. In the first decade of the twenty first century, a few tools have been developed to enable this reconstruction. Most of these require a non-trivial amount of computational knowledge if they are to be used. Some, such as iHOP and STRING, are widely used and user-friendly. Each of these applications searches a database of scientific documents that was previously analyzed and processed. This pre-processing strategy makes the identification of co-occurring genes a faster process at the cost of disregarding documents present in PubMed and/or Medline but not yet processed by the pipeline underlying the applications. Biblio-MetReS, which is developed to fit in this user friendly category, provides the following added value with respect to iHOP and STRING:

1. Our reconstruction is done live and with the latest available documents on the internet. In contrast, iHOP and STRING use a precompiled database of documents for their search. This means that our results will be more up to date than those of the other two applications.

2. While iHOP, STRING, and Biblio-MetReS search for gene interactions in abstracts of Medline and PubMed documents, Biblio-MetReS can additionally search full documents from other scientific and general data sources. This increases the number of gene associations that can be found. Nevertheless, it has been reported that the analysis of complete scientific documents may increase the noise in gene associations that are found [20,32].

3. A third additional functionality provided by Biblio-MetReS with respect to iHOP and STRING permits filtering out some of the noise that may arise from the analysis of complete documents. Our tool distinguishes between co-occurrence of genes and proteins in sentences, paragraphs and whole documents. The analysis of sentences

decreases the probability of detecting spurious associations between genes that are found in different parts of the documents and may have little to do with one another.

Both pre-processing of documents strategies, as done by iHOP and STRING, and on-the fly analysis strategies, as done by Biblio-MetReS or Reflect, have disadvantages. This first strategy has the cost of using information that is almost never quite up to date, while the latter has the cost of becoming potentially very slow. One way to sidestep these disadvantages is by combining both strategies in the same tool. We are working on an implementation of Biblio-MetReS that will do this. In fact, the next version of Biblio-MetReS is being implemented in such a way that the results of each search will be stored and compiled. Thus, if a new search finds a document that has been analyzed before, it will retrieve the processed data from our local database. Only new documents will be processed on the fly. This approach will combine the advantages of on-the-fly processing and pre-processing strategies, enabling the application to speed up searches, analyses, and reconstruction of networks. It will also facilitate implementing methods to better predict the confidence in the different interactions that are found, based for example on Bayesian networks [40].

Our tool, together with iHOP and STRING, is limited by the non-standardized nomenclature that exists in biology. Each application finds a different set of genes for each benchmarked network, with only partial overlap between the genes that are identified. Furthermore, no application finds all genes that belong to the canonical pathway defined in the KEGG server. This fact is a consequence not only of non-standard nomenclature but also of the limitations of the various datasets, where not all possible experiments and associations have been reported. Furthermore, many of these associations are reported in older papers that have yet to be made available over the web. Nevertheless, the results also emphasize the usefulness of those tools, as they tag a number of genes that interact with the benchmarked pathways but do not belong to it. The usefulness of this kind of network reconstruction will increase over time, as the

nomenclature of genes and biological concepts becomes more standardized and widely used and the number of scientific documents that associate genes to biological function increases.

2.5 Conclusions

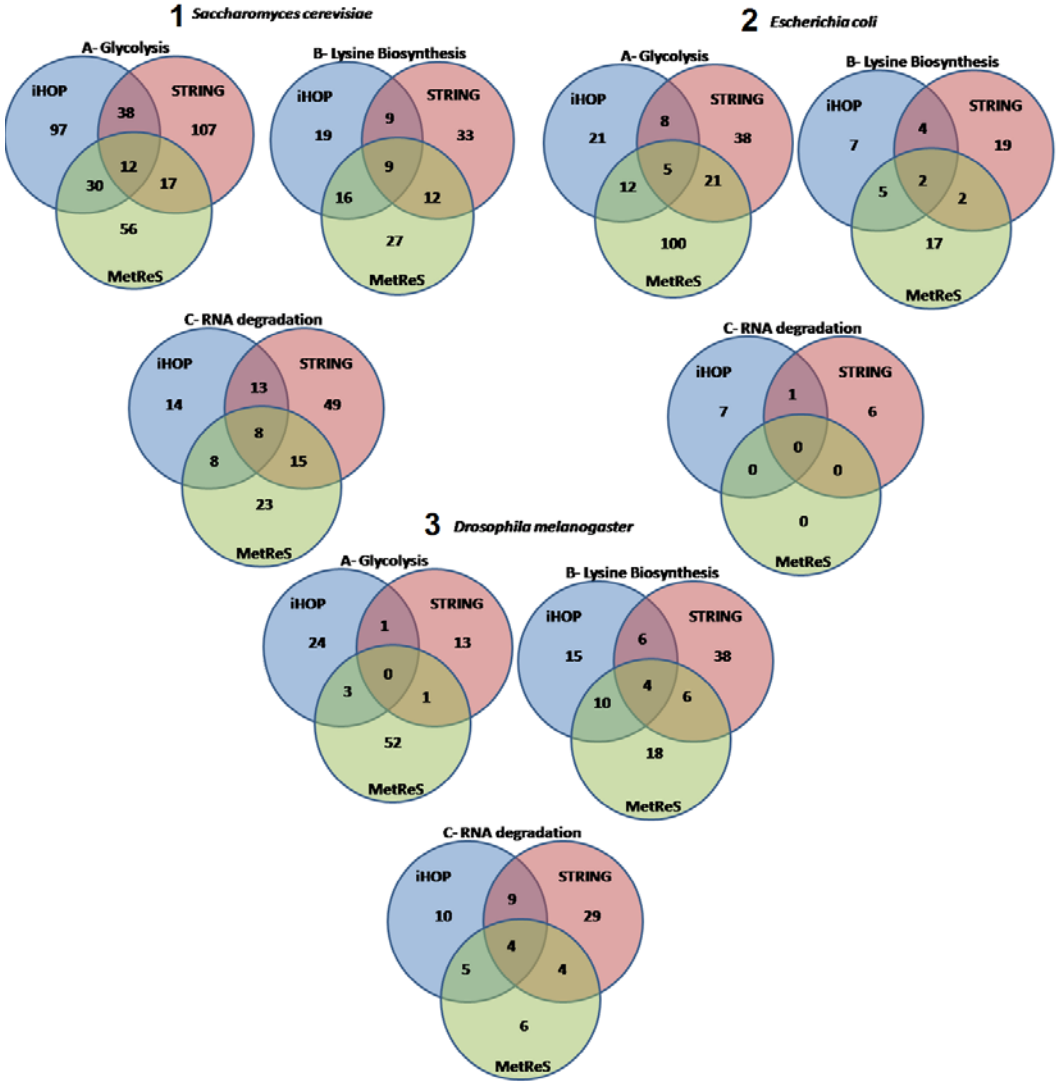
Biblio-MetReS is a new user-friendly tool for text-based network reconstruction that is comparable in function to iHOP and STRING. Biblio-MetReS is more flexible than both, iHOP and STRING, in at least two aspects, while being equally user-friendly. First, it includes all sources of information used by iHOP and STRING, always analyzing the most up to date version of these sources. Second, the user can choose different sources of information to search from simply by checking boxes. Neither iHOP nor STRING allow for this. Furthermore, it permits analyzing the full text of scientific documents, rather than only mining the information contained in abstracts.

2.6 Supplementary Materials

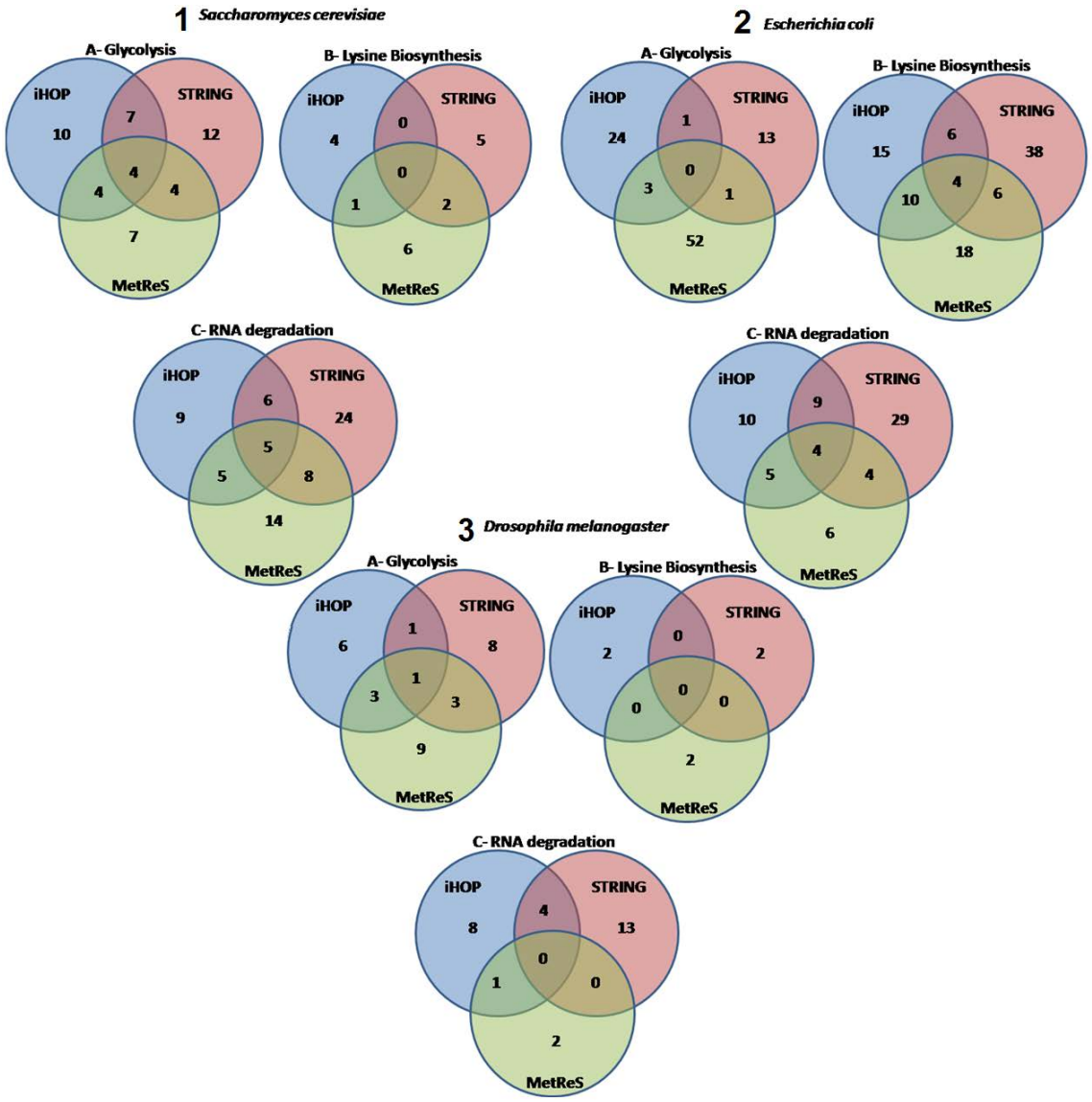
These materials contain the following information:

- Supplementary Figure 1.
- Supplementary Figure 2.
- Supplementary Figure 3.
- Supplementary Figure 4
- Supplementary Table 1

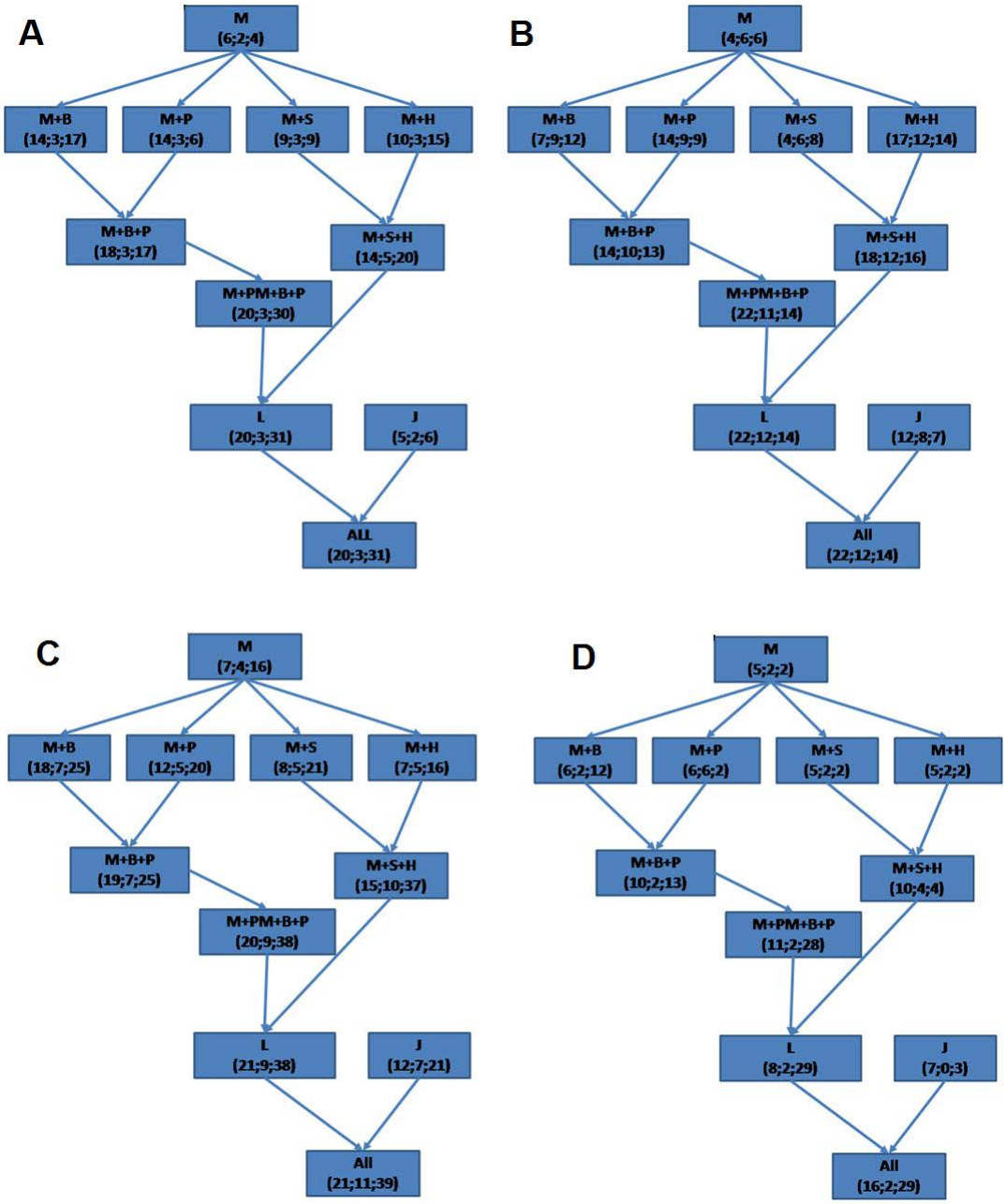
Supplementary Figures



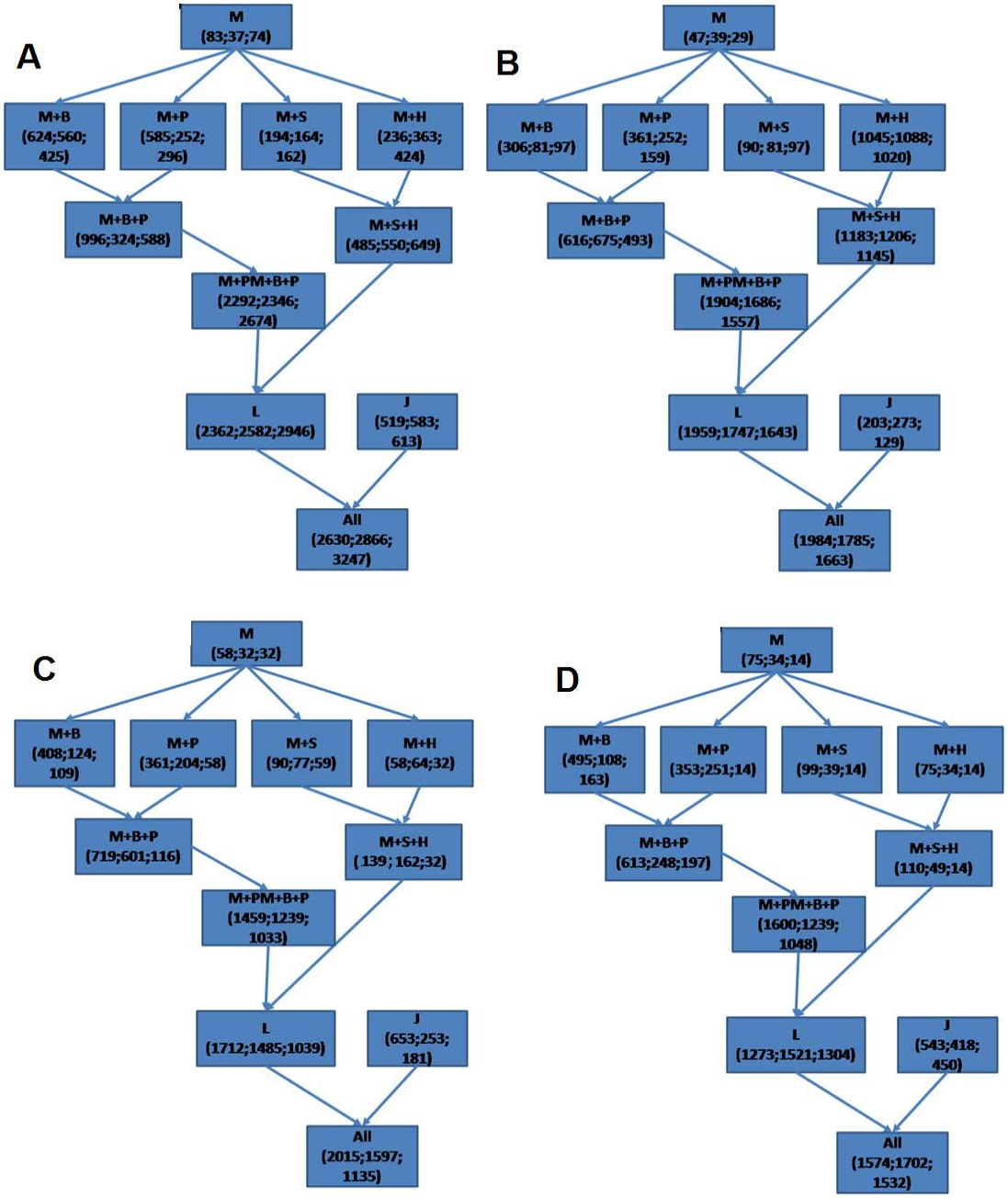
Supplementary Figure 1. Representation of the number of common genes found for the different pathways in *Saccharomyces cerevisiae(1), Escherichia coli(2)*, and *Drosophila melanogaster(3)* using Biblio-MetReS, iHOP and STRING. A - Glycolysis, B - Lysine metabolism, C - RNA degradation.



Supplementary Figure 2. Representation of the number of common genes found for the different pathways in *Saccharomyces cerevisiae(1), Escherichia coli(2)*, and *Drosophila melanogaster(3)* using Biblio-MetReS, iHOP and STRING.A - Glycolysis, genes known to be in the pathway, B - Lysine metabolism, genes known to be in the pathway, C - RNA degradation, genes known to be in the pathway.



Supplementary Figure 3. Representation of the number of additional genes that are found by Biblio-MetReS as we add more data sources to Medline. Each panel shows three numbers in each square. The first number represents the number of genes found for glycolysis. The second number shows the number of genes found for lysine metabolism. The third number shows the number of genes found for RNA degradation. A - *Homo sapiens*. B - *Escherichia coli*. C - *Saccharomyces cerevisiae*. D - *Drosophila melanogaster*. In this figure we represent only the genes that are known to belong to the canonical pathways as defined in KEGG. M: Medline, B:Biomed Central, P: PLoS, S: Scopus, H: Highwire, PM: PubMed, L: All literature databases, J: All journals and ALL: all information sources.



Supplementary Figure 4. Representation of the number of additional genes found by Biblio-MetReS that are known to belong to the canonical pathways under analysis as we add more data sources to Medline. Each panel shows three numbers in each square. The first number represents the number of genes found for glycolysis. The second number shows the number of genes found for lysine metabolism. The third number shows the number of genes found for RNA degradation. A - *Homo sapiens*. B - *Escherichia coli*. C - *Saccharomyces cerevisiae*. D - *Drosophila melanogaster*. In this figure we represent all genes found during the automated analysis. M: Medline, B:Biomed Central, P: PLoS, S: Scopus, H: Highwire, PM: PubMed, L: All literature databases, J: All journals and ALL: all information sources.

Supplementary Tables

Supplementary Table 1. Benchmarcking of the application

Organisms	Pathways	Seed Genes	Benchmark 1*	Benchmark 2**
Saccharomyces cerevisiae	Glycolysis	PGM1, FBA1, CDC19	iHOP vs. STRING vs. Biblio-MetReS	T1-T12
	Lysine biosynthesis	LYS21, ARO8, LYS9	iHOP vs. STRING vs. Biblio-MetReS	T1-T12
	RNA degradation	MTR3, MPP6, CAF16, RRP41	iHOP vs. STRING vs. Biblio-MetReS	T1-T12
Homo sapiens	Glycolysis	PGM1, ALDOA, PKLR	iHOP vs. STRING vs. Biblio-MetReS	T1-T12
	Lysine biosynthesis	AADAT, ASSDH, AASS	iHOP vs. STRING vs. Biblio-MetReS	T1-T12
	RNA degradation	MTR3, MPP6, CNOT4, RRP41	iHOP vs. STRING vs. Biblio-MetReS	T1-T12
Escherichia coli	Glycolysis	Pgm, fbaB, PykF	iHOP vs. STRING vs. Biblio-MetReS	T1-T12
	Lysine biosynthesis	thrA, dapB, dapF	iHOP vs. STRING vs. Biblio-MetReS	T1-T12
	RNA degradation	rppH, rlhE, rnr	iHOP vs. STRING vs. Biblio-MetReS	T1-T12
Drosophila melanogaster	Glycolysis	PGM, ALD, PYK	iHOP vs. STRING vs. Biblio-MetReS	T1-T12
	Lysine degradation	LKR, CG9547, GPP	iHOP vs. STRING vs. Biblio-MetReS	T1-T12
	RNA degradation	RRP42, MPP6, CNOT4, RRP41	iHOP vs. STRING vs. Biblio-MetReS	T1-T12

* In this benchmark we have compared the ability of three different servers to reconstruct the molecular networks regulating three types of well characterized cellular processes. We have run Biblio-MetReS using only the Medline database and used iHOP and STRING to reconstruct the networks.
** In this benchmark we have compared the differences in equivalent networks reconstructed using different information sources from the same starting set of genes. T1.- Reconstruction using general search engines. We neither discuss nor show results for these benchmark tests, because they are very non-specific. T2.- Reconstruction using Medline abstracts. T3.- Reconstruction using Medline abstract and Biomed central documents. T4.- Reconstruction using Medline abstracts and PLoS documents. T5.- Reconstruction using Medline abstracts and documents from SCOPUS. T6.- Reconstruction using Medline abstracts and documents from the Highwire database. T7.- Reconstruction using Medline abstracts and documents from Biomed central and PLoS. T8.- Reconstruction using Medline abstracts and documents from the SCOPUS and HIGHWIRE databases. T9.- Reconstruction using Medline abstracts and documents from Pubmed, PLoS and Biomed central. T10.- Reconstruction using Medline and documents from PLoS, Biomed central, Pubmed, SCOPUS and HIGHWIRE. T11.- Reconstruction using documents from all the databases in the Journals pane of Biblio-MetReS. T12.- Reconstruction using all information sources from the scientific literature and journals panes of Biblio-MetReS.

2.7 References

1. Alves R, Sorribas A (2007) In silico pathway reconstruction: Iron-sulfur cluster biogenesis in Saccharomyces cerevisiae. BMC Syst Biol 1: 10. doi:10.1186/1752-0509-1-10.
2. Markowetz F, Spang R (2007) Inferring cellular networks--a review. BMC Bioinformatics 8 Suppl 6: S5. doi:10.1186/1471-2105-8-S6-S5.
3. Hoffmann R, Valencia A (2005) Implementing the iHOP concept for navigation of biomedical literature. Bioinformatics 21: ii252–ii258. doi:10.1093/bioinformatics/bti1142.
4. Hoffmann R, Valencia A (2004) A gene network for navigating the literature. Nat Genet 36: 664–664. doi:10.1038/ng0704-664.
5. Von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, et al. (2007) STRING 7—recent developments in the integration and prediction of protein interactions. Nucleic Acids Res 35: D358–D362. doi:10.1093/nar/gkl825.
6. Barbosa-Silva A, Soldatos TG, Magalhães IL, Pavlopoulos GA, Fontaine J-F, et al. (2010) LAITOR - Literature Assistant for Identification of Terms co-Occurrences and Relationships. BMC Bioinformatics 11: 70. doi:10.1186/1471-2105-11-70.
7. Kemper B, Matsuzaki T, Matsuoka Y, Tsuruoka Y, Kitano H, et al. (2010) PathText: a text mining integrator for biological pathway visualizations. Bioinformatics 26: i374–i381. doi:10.1093/bioinformatics/btq221.
8. Walport M, Kiley R (2006) Open access, UK PubMed Central and the Wellcome Trust. J R Soc Med 99: 438–439. doi:10.1258/jrsm.99.9.438.
9. Pafilis E, O'Donoghue SI, Jensen LJ, Horn H, Kuhn M, et al. (2009) Reflect: augmented browsing for the life scientist. Nat Biotechnol 27: 508–510. doi:10.1038/nbt0609-508.
10. Rebholz-Schuhmann D, Arregui M, Gaudan S, Kirsch H, Jimeno A (2007) Text processing through Web services: calling Whatizit. Bioinformatics 24: 296–298. doi:10.1093/bioinformatics/btm557.
11. Plake C, Schiemann T, Pankalla M, Hakenberg J, Leser U (2006) AliBaba: PubMed as a graph. Bioinforma Oxf Engl 22: 2444–2445. doi:10.1093/bioinformatics/btl408.
12. Krallinger M, Leitner F, Valencia A (2010) Analysis of biological processes and diseases using text mining approaches. Methods Mol Biol Clifton NJ 593: 341–382. doi:10.1007/978-1-60327-194-3_16.
13. Krallinger M, Valencia A, Hirschman L (2008) Linking genes to literature: text mining, information extraction, and retrieval applications for biology. Genome Biol 9 Suppl 2: S8. doi:10.1186/gb-2008-9-s2-s8.
14. Hahn U, Valencia A (2006) Semantic Mining in Biomedicine (Introduction to the papers selected from the SMBM 2005 Symposium, Hinxton, U.K., April 2005). Bioinformatics 22: 643–644. doi:10.1093/bioinformatics/btl084.

15. Yuryev A, Mulyukov Z, Kotelnikova E, Maslov S, Egorov S, et al. (2006) Automatic pathway building in biological association networks. BMC Bioinformatics 7: 171. doi:10.1186/1471-2105-7-171.
16. Overby C, Tarczy-Hornoch P, Demner-Fushman D (2009) The potential for automated question answering in the context of genomic medicine: an assessment of existing resources and properties of answers. BMC Bioinformatics 10: S8. doi:10.1186/1471-2105-10-S9-S8.
17. Krallinger M, Morgan A, Smith L, Leitner F, Tanabe L, et al. (2008) Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge. Genome Biol 9 Suppl 2: S1. doi:10.1186/gb-2008-9-s2-s1.
18. Hu Z-Z, Mani I, Hermoso V, Liu H, Wu CH (2004) iProLINK: an integrated protein resource for literature mining. Comput Biol Chem 28: 409–416. doi:10.1016/j.compbiolchem.2004.09.010.
19. De Bruijn B, Martin J (2002) Getting to the (c)ore of knowledge: mining biomedical literature. Int J Med Inf 67: 7–18.
20. Shah PK, Perez-Iratxeta C, Bork P, Andrade MA (2003) Information extraction from full text scientific articles: where are the keywords? BMC Bioinformatics 4: 20. doi:10.1186/1471-2105-4-20.
21. Nuzzo A, Mulas F, Gabetta M, Arbustini E, Zupan B, et al. (2010) Text Mining approaches for automated literature knowledge extraction and representation. Stud Health Technol Inform 160: 954–958.
22. Song Y-L, Chen S-S (2009) Text mining biomedical literature for constructing gene regulatory networks. Interdiscip Sci Comput Life Sci 1: 179–186. doi:10.1007/s12539-009-0028-7.
23. Ananiadou S, Pyysalo S, Tsujii J, Kell DB (2010) Event extraction for systems biology by text mining the literature. Trends Biotechnol 28: 381–390. doi:10.1016/j.tibtech.2010.04.005.
24. Laakso M, Hautaniemi S (2010) Integrative platform to translate gene sets to networks. Bioinformatics 26: 1802–1803. doi:10.1093/bioinformatics/btq277.
25. Bandy J, Milward D, McQuay S (2009) Mining protein-protein interactions from published literature using Linguamatics I2E. Methods Mol Biol Clifton NJ 563: 3–13. doi:10.1007/978-1-60761-175-2_1.
26. Hur J, Schuyler AD, States DJ, Feldman EL (2009) SciMiner: web-based literature mining tool for target identification and functional enrichment analysis. Bioinforma Oxf Engl 25: 838–840. doi:10.1093/bioinformatics/btp049.
27. Saetre R, Yoshida K, Miwa M, Matsuzaki T, Kano Y, et al. (2010) Extracting protein interactions from text with the unified AkaneRE event extraction system. IEEEACM Trans Comput Biol Bioinforma IEEE ACM 7: 442–453. doi:10.1109/TCBB.2010.46.
28. Kolchinsky A, Abi-Haidar A, Kaur J, Hamed AA, Rocha LM (2010) Classification of Protein-Protein Interaction Full-Text Documents Using Text and Citation Network Features. IEEE/ACM Trans Comput Biol Bioinform 7: 400–411. doi:10.1109/TCBB.2010.55.

29. Dai H-J, Lai P-T, Tsai RT-H (2010) Multistage gene normalization and SVM-based ranking for protein interactor extraction in full-text articles. IEEEACM Trans Comput Biol Bioinforma IEEE ACM 7: 412–420. doi:10.1109/TCBB.2010.45.
30. Chen Y, Liu F, Manderick B (2010) BioLMiner System: interaction normalization task and interaction pair task in the BioCreative II.5 challenge. IEEEACM Trans Comput Biol Bioinforma IEEE ACM 7: 428–441. doi:10.1109/TCBB.2010.47.
31. Ohta T, Matsuzaki T, Okazaki N, Miwa M, Sætre R, et al. (2010) Medie and Info-pubmed: 2010 update. BMC Bioinformatics 11: P7. doi:10.1186/1471-2105-11-S5-P7.
32. Lin J (2009) Is searching full text more effective than searching abstracts? BMC Bioinformatics 10: 46. doi:10.1186/1471-2105-10-46.
33. McIntosh T, Curran JR (2009) Challenges for automatically extracting molecular interactions from full-text articles. BMC Bioinformatics 10: 311. doi:10.1186/1471-2105-10-311.
34. Lourenço A, Carreira R, Carneiro S, Maia P, Glez-Peña D, et al. (2009) @Note: a workbench for biomedical text mining. J Biomed Inform 42: 710–720. doi:10.1016/j.jbi.2009.04.002.
35. Lourenço A, Carreira R, Glez-Peña D, Méndez JR, Carneiro S, et al. (2010) BioDR: Semantic indexing networks for biomedical document retrieval. Expert Syst Appl 37: 3444–3453. doi:10.1016/j.eswa.2009.10.044.
36. Aoki KF, Kanehisa M (2005) Using the KEGG database resource. Curr Protoc Bioinforma Ed Board Andreas Baxevanis Al Chapter 1: Unit 1.12. doi:10.1002/0471250953.bi0112s11.
37. Geer LY, Marchler-Bauer A, Geer RC, Han L, He J, et al. (2009) The NCBI BioSystems database. Nucleic Acids Res 38: D492–D496. doi:10.1093/nar/gkp858.
38. Leitner F, Mardis SA, Krallinger M, Cesareni G, Hirschman LA, et al. (2010) An Overview of BioCreative II.5. IEEEACM Trans Comput Biol Bioinforma IEEE ACM 7: 385–399.
39. Stapley BJ, Benoit G (2000) Biobibliometrics: information retrieval and visualization from co-occurrences of gene names in Medline abstracts. Pac Symp Biocomput Pac Symp Biocomput: 529–540.
40. Steele E, Tucker A, 't Hoen PAC, Schuemie MJ (2009) Literature-based priors for gene regulatory networks. Bioinforma Oxf Engl 25: 1768–1774. doi:10.1093/bioinformatics/btp277.

Chapter 3. Biblio-MetReS II

Submitted

Biblio-MetReS for user friendly mining of genes and biological processes in scientific documents

Anabel Usié, Hiren Karathia, Ivan Teixidó, Rui Alves and Francesc Solsona

Abstract

One way to facilitate the reconstruction of molecular circuits is by using automated text-mining techniques. Developing more efficient methods for such reconstruction is a topic of active research, and those methods are typically included by bioinformaticians in pipelines used to mine and curate large literature datasets. Nevertheless, experimental biologists have a limited number of available user-friendly tools that use text-mining for network reconstruction and require no programming skills to use. One of these tools is Biblio-MetReS.

Originally, this tool permitted an on-the-fly analysis of documents contained in a number of web-based literature databases to identify co-occurrence of proteins/genes. This approach ensured results that were always up-to-date with the latest live version of the databases. However, this "up-to-dateness" came at the cost of large execution times.

In this paper we report the current version of Biblio-MetReS, including new functionality that identifies co-occurrence of biological processes, pathways, and proteins. In addition, the program now combines on-the-fly document analysis with preprocessing of previously analyzed documents. This strategy simultaneously maximizes "up-to-dateness" of the results and minimizes run time of the analysis.

3.1 Introduction

The reconstruction of molecular circuits and their behavior is an important research goal in the biological sciences. One of the ways to achieve that circuit reconstruction is by using automated text-mining techniques, due to the increased number of scientific documents that are collected in biological databases [1,2].

A gold standard of these databases, MEDLINE, contains more than 19×10^6 records, with 2000-4000 new entries being added each day [3]. Extracting biological information from such large databases requires text-mining methods and tools that are able to automatically integrate and summarize useful biological information across the database records.

The development of text-mining methods that enable circuit reconstruction from scientific documents is an area of active development [4–11] The performance of those methods for automated identification of the circuits [7], of their components (genes/proteins), and of the inter-component relationships, has been systematically evaluated over the last few years, for example through the BioNLP [8,9] and BioCreAtIvE initiatives [10,11].

To briefly summarize, there are three large classes of methods that have been employed for circuit reconstruction: dictionary-based methods, morphology-based methods, and context-based methods [12]. Dictionary-based methods rely on compiled list of the terms that are to be recognized, and implementing a matching method to find those terms in the text of the documents [13]. Morphology-based methods rely on the morphological structure of specific classes of words to single them out in documents [14,15]. Finally, context-based methods can be divided into Machine Learning or Natural Language Processing techniques: The former identify patterns in the structure of the text that help to recognize the presence of the relevant entities in documents; the later draw from our knowledge of language grammar and how natural language is formed to recognize those entities. These three general approaches can be combined in order to improve NER (for example see [16]).

In general, methods participating in evaluations such as BioCreAtIvE or BioNLP are implemented in tools that can be included in web-services and assist curators in the maintenance of large databases of biological knowledge. Examples of this are given in [17]. In most cases, using these methods and tools requires that one becomes an expert computer user and learns how to program.

Experimental scientists that are interested in being users of, without becoming experts in, text-mining methods to directly reconstruct molecular circuits for their genes of interest from scientific documents have a much smaller set of available user-friendly tools. The first that became available was iHOP [4], which was later joined by STRING [5]. These web applications allow anyone to reconstruct the network of co-occurrences contained in Medline abstracts for their proteins/genes of interest in a user friendly and intuitive way. There are two caveats of using these applications. First, they rely on preprocessed versions of the Medline database, which means that they are always out of date. Second, they disregard the analysis of full text documents.

Recognizing these limitations, Biblio-MetReS (**Biblio**metric **Met**abolic network **Re**construction **S**erver [6]) was implemented for the same target audience as STRING or iHOP, but relying on two differential features. The first was that it would search databases and analyze documents on the run, thus providing the users with the most up-to-date results available on the web. The second was that full text documents were also analyzed, as were other databases besides Medline. These two features made Biblio-MetReS significantly slower than STRING and iHOP.

Here, we report Biblio-MetReS v2.0. This new version combines on-the-fly analysis of new documents with preprocessed information from documents that were encountered in previous analysis. This combination simultaneously optimizes program run time and "up-to-dateness" of the analysis. In addition, this new version allows users to also identify GO terms and KEGG or Panther pathways that might be associated to their genes/proteins of interest in the documents that they analyze. This further facilitates the identification of functional relationships between proteins and aids in identifying the biological

processes and circuits in which those proteins may be involved. To our knowledge, no other user-friendly tool is available to simultaneously analyzes genes/proteins and GO terms/pathways document co-occurrences.

3.2 Methods

3.2.1 Pre-processing of documents

Biblio-MetReS uses exact matching to an internal dictionary of synonyms approach to identify co-mentions of genes/proteins from more than 1200 organisms in the text of scientific records, as described in section 1 of supplementary materials (see also [6]). The database containing the organisms and their gene names is updated every three months using information compiled automatically from NCBI. In addition, Biblio-MetReS v2.0 implements a precompilation approach that works in the following way. Any search done will identify a given number of documents in the database(s) selected by the user(s) for an organism of interest. If a given document has not been found in any previous search by any user, with the same organism, Biblio-MetReS will analyze it as described in section 1 of supplementary materials (see also [6]) and all information contained in that document and relevant for the analysis will be stored in a central database (see section 3 of supplementary materials for detailed information). If a given document has been previously found by any user, its information will be directly accessed from our central database, and the document will not be reanalyzed. This means that newly found documents are mined on the fly by the program to find and count mentions of relevant entities, while mentions in documents that have been previously found are simply looked up in our central database.

3.2.2 GO and Pathway entities

In addition to identifying genes/proteins, as we mentioned in the introduction, v2.0 of Biblio-MetReS also searches for all entities from the GO ontology biological process categories [18] and from the complete joint sets of KEGG [19]

and Panther pathways [20]. This allows it to identify co-occurrence among GO/Pathway entities and between GO/Pathway entities and gene/protein entities in sentences, paragraphs or documents. Identification of the GO/Pathway entities is done using a dictionary approach with exact matching.

3.3 Results

3.3.1 Biblio-MetReS and Biblio-MetReS Player

Biblio-MetReS v2.0 can be used to identify genes/proteins from more than 1200 different organisms in records stored in a variety of databases. Users download the application and run it locally. An internet connection and JAVA software is required. Upon starting the program, users login to the central Biblio-MetReS database and choose which organism they are interested in and whether they want to search only for co-occurrence of genes/proteins or if they also want to include biological pathways and/or GO biological processes in the analysis. Once this choice is made, the program loads the necessary information from the central database. Taking this approach, instead of including all the data locally in the program installation, permits making an application that is much smaller in size and needs less RAM to function properly. Subsequently, users select the source of documents that they want to analyze, as well as the genes/proteins and/or pathways/GO biological processes that they want to search for. They can also include their own handpicked list of documents to be searched. Once the search is launched, Biblio-MetReS will identify documents in the relevant databases that contain mentions to the relevant search items. After identifying these documents, the application fully analyzes them to identify mentions for any additional gene or Pathway/GO biological process via a dictionary matching approach. The co-occurrence of the different entities is analyzed at the level of the whole document, of individual paragraphs and of individual sentences, and the significance of this co-occurrence is calculated as described in [6]. The information that is relevant for the co-occurrence calculations is stored in the central Biblio-MetReS database. Any subsequent searches that

identify the same document will not reanalyze it; instead, these numbers are directly retrieved from that database. Once the analysis is complete, the users can visualize it in graphical and textual form. Links to the documents and sentences where co-occurrences are found are provided. Graphical visualization of the results can be done in different ways. Users can create graphs for the global co-occurrences network and for gene- or pathway/process-centric co-occurrences at the document/paragraph and sentence levels. Significance and Mutual Information of each co-occurrence is also provided in tabular form. The graphical representation of the networks is automatically stored in local xml files. These files can be opened using a small app, Biblio-MetReS Player, which can be downloaded from the Biblio-MetReS website. This permits reviewing previously obtained networks without having to redo the search. All this process is summarized in Figure 1.

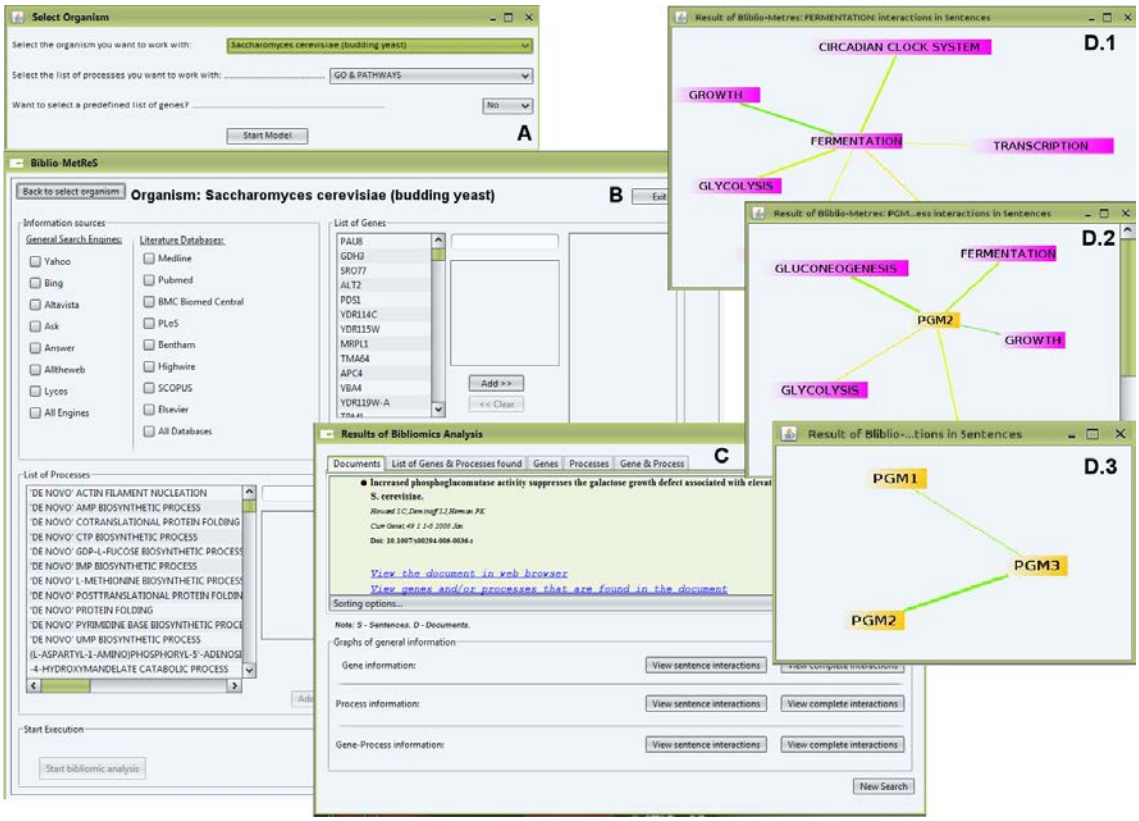


Figure 1. Workflow for Biblio-MetReS.

3.3.2 Improvements with respect to previous versions

One of the improvements in the current version of Biblio-MetReS is the possibility of identifying pathways and biological processes. This brings the application one step closer to helping the user in the reconstruction of the molecular circuits in which the genes of interest are involved. However, the main improvement has to do with the implementation of precompilation in the searches, as described in methods.

To test the effect of this change on the run time we performed the same benchmarks as described in [6] (also, see Figure 2 details and section 2 of supplementary materials). In brief, we used the name of three or four genes involved in glycolysis, lysine biosynthesis, or RNA metabolism and performed co-occurrence analyses ensuring that the database did not include any of the documents included in the searches. Subsequently we repeated the searches, with the database now containing the information from previously found documents. In all cases, the time decreased by around two orders of magnitude (Figure 1 and Supplementary Table 1 and Supplementary Figure 2 in supplementary materials).

3.4. Discussion

Here we present the new version of Biblio-MetReS, a user friendly tool for the identification of gene/protein co-occurrence networks in scientific documents. The major changes with respect to version 1.0 have to do with the search and analysis process of the documents, which can now be up to 95% faster than in the previous version. In addition, the tool now also searches for co-occurrences of biological processes and pathways, to help users to more easily establish the biological circuits in which their genes of interest may be involved in.

The methods used by the application to identify genes and proteins in the documents are dictionary-based. These methods perform on par with iHOP and STRING for gene and protein identification [6].

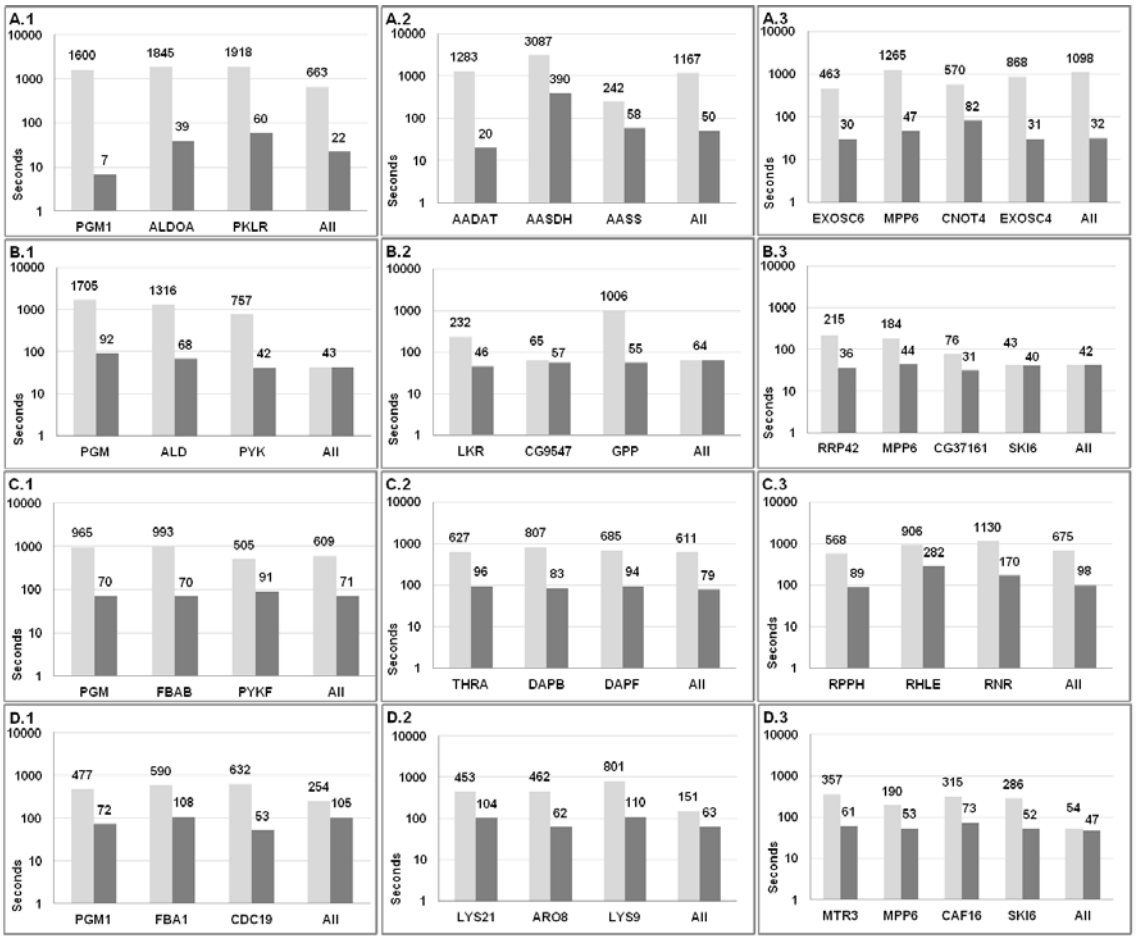


Figure 2. Effect of preprocessing documents on Biblio-MetReS' run time. In brief, genes from three KEGG-defined pathways are used for this test. Panels A.x show experimental results for glycolysis genes. Panels B.x show experimental results for Lysine biosynthesis genes. Panels C.X show experimental results for RNA degradation genes. Three organisms are used in this benchmark. Panels Y.1 show results for *Homo sapiens*, panels Y.2 show results for *Drosophila melanogaster*, panels Y.3 show results for *Escherichia coli,* and panels Y.4 show results for *Saccharomyces cerevisiae*. These pathways and organisms were chosen to remain consistent with the tests performed in [6]. Searches were done selecting all the databases in the application. Graphs can be interpreted as follows. Light gray bars indicate the run time for Biblio-MetReS when the corresponding gene is searched for the first time. In this case the program has to do a full document analysis on the fly and no information has been pre-processed. Darker gray bars indicate the run time for Biblio-MetReS when the search for the corresponding gene is repeated, and pre-processed information is already present in Biblio-MetReS' central database. The column "All" indicates the run-time for searching all genes in the graph simultaneously, after individual searches for each gene had already been done and results pre-processed and stored.

Taking into account the results from the BiocreAtIvE initiative, dictionary matching methods applied to GO term identification have a high precision (that is, the terms that are identified are mostly correct) and low recall (many terms that should also be identified are not) [7]. However, other classes of methods proposed in the same initiative have average recalls between 10% and 20% for the same type of entities and lower precision [7]. Because we choose to have higher precision in our identification of GO and Pathway terms, we choose to also identify these terms using a dictionary approach. However, a quantitative evaluation of the performance for this approach was not done in this work. This is so because such an evaluation was done in BioCreAtIvE, as stated above. Re-evaluating the same approach using the same corpora used in those initiatives would be inefficient and add little, if anything to the work presented here.

As is demonstrated by the BioCreAtIvE challenge [21], the problem of identifying entities in scientific texts is far from solved. Although Biblio-MetReS aims at giving non-expert users the possibility of performing such identification and use that identification to extract biological knowledge, there is much room for improvement. We are implementing an offline system to automatically search, analyze, and store information about gene/protein and pathway/biological processes co-occurrences in the documents. This will contribute to decrease the dependence of Biblio-MetReS on the users and their searches to preprocess information and make searches faster.

3.5 Supplementary Materials

These materials contain the following information:

- 1: Document Search and analysis.
- 2: Benchmarking.
- 3: Storage of Information.
- Supplementary Figure 1.
- Supplementary Figure 2.
- Supplementary Table 1.
- Supplementary Table 2.

3.5.1. Document search and analysis

The different information sources accessed by Biblio-MetReS to create a meta-search engine are divided into two groups: Literature Databases (Medline, Pubmed, Plos One, etc) and General Engines (Yahoo, Ask, etc.) [6]. The organism of interest and the gene(s) and/or biological processes selected by the user are assembled into a query that is launched to each selected information source.

Once the search is completed, Biblio-MetReS eliminates duplicate documents by comparing URLs and doi numbers. Once a non-redundant list of documents is assembled, Biblio-MetReS selects the full text of the document to analyze, unless only the abstract is available. It discards all the documents for which the user IP address has no access. Then the tool checks if any of the flagged documents has been previously analyzed and stored in the database, retrieving preprocessed data from that database, if available. At the end of this process, the application analyzes on the fly the documents that were not found in the database, looking for co-occurrence of any gene or protein, and/or for any biological process (if the user included them in the analysis) in sentences, paragraphs, and entire documents. It then stores the relevant analysis of new documents. We note that even though PubMed frequently updates the content of its documents, we assume that these updates will not significantly affect the

entities that are found in any given document. Therefore, once analyzed and stored, a document will not be reanalyzed again live unless it is removed first from our database of preprocessed documents.

The method used to find both types of entities (genes/proteins and biological processes/pathways) is exact matching, including all the synonyms for a gene. The dictionary of genes and their synonyms we use is a merge from those of NCBI and KEGG. The dictionary of biological processes we use is a merge of GO, KEGG and Panther. See the workflow in Supplementary Figure 1 for a summary of the process.

The decision to analyze the co-occurrence in sentences and paragraph relies on the consideration that, when analyzing full text documents, proximity between the entities implies a more direct relationship between them. If this is not taken into account, a significant amount of noise can be included in the network of possible interactions [22].

3.5.2. Benchmarking

The benchmarking of the program and its improvements with respect to version 1.0 was carried out using four organisms of interest: *Saccharomyces cerevisiae*, *Homo sapiens*, *Escherichia coli* and *Drosophila melanogaster*. For each organism we used as a search seed the following genes belonging to Glycolysis, Lysine biosynthesis and RNA Degradation pathways. Details are shown in Supplementary Table 1.

The benchmark is divided into two different tests. The first test is done using only the Pubmed database. The second test is done using all the Literature databases. In both tests we used all the seed genes from Supplementary Table 1. Each seed is used by Biblio-MetReS as a query search. This query search is launched twice. The first search is done with a database that contains no preprocessed document. The information in documents is analyzed on-the-fly and stored. The second time, the search is repeated with the documents now stored in the database. This allows us to estimate the percentage of run-time

saved by preprocessing the documents. The results are shown in Supplementary Table 2.

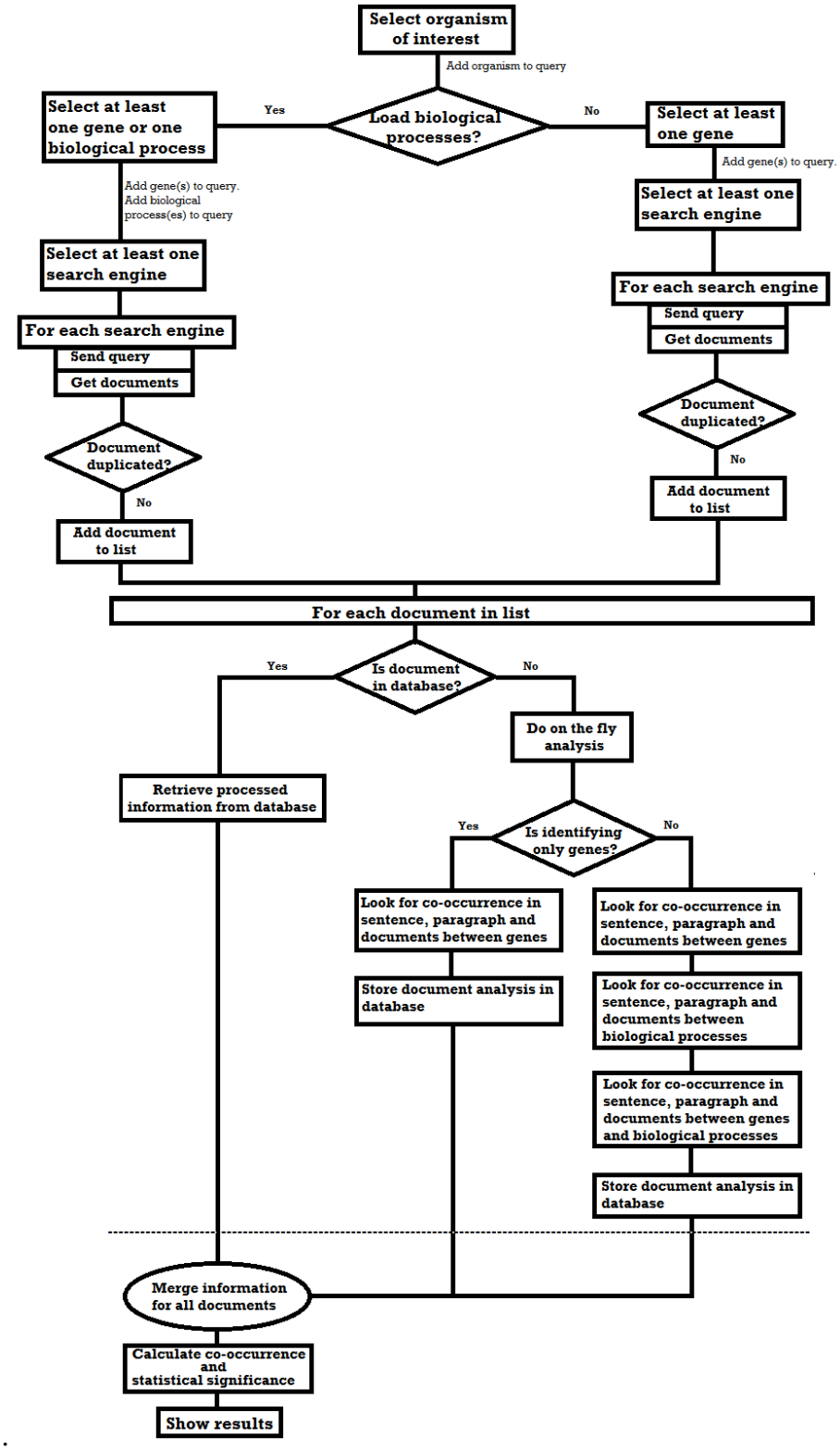
3.5.3. Precompiled information in the database

To save the information related to each document we created a set of seven tables in the database. These tables allow Biblio-MetReS to reduce the execution time and reproduce the same result given by the on-the-fly analysis. The tables are:

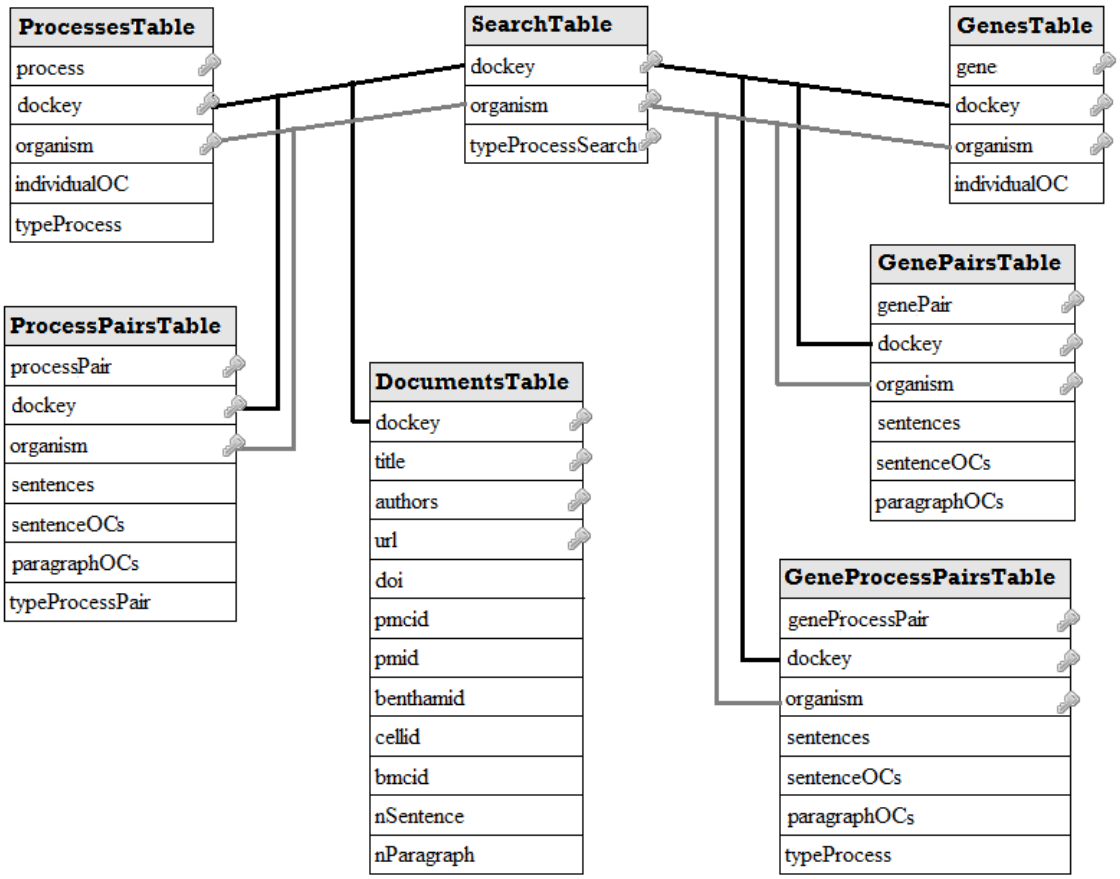
1. DocumentsTable
2. SearchTable
3. GenesTable
4. ProcessesTable
5. GenePairsTable
6. ProcessPairsTable
7. GeneProcessPairsTable

Supplementary Figure 2 details the relationship between the tables and their variables. There are some aspects about the variables in each table that have to be clarified. First, a variable in plural refers to a list of elements. In *SearchTable*, the variable *typeProcessSearch* can assume the following values: 0 - the search is done mapping genes and biological processes from GO, Pathways and Panther; 1 - the search is done mapping genes and biological processes from GO; 2- the search is done mapping genes and biological processes from Pathways and Panther; 3- the search is done only mapping genes. In *ProcessesTable* and *GeneProcessPairTable*, the variable *typeProcess* can assume the following values: 0- the biological process is from GO; 1.- the biological process is from Pathway or Panther. Finally, in *ProcessPairsTable* the variable *typeProcessPair* can assume the following values: 0- both biological processes in the pair are from GO; 1- both biological processes in the pair are from Pathway or Panther; 2- one biological process in the pair is from GO and the other is from Pathway or Panther.

Supplementary Figures



Supplementary Figure 1. Algorithmic workflow for the process of network reconstruction in Biblio-MetReS



Supplementary Figure 2. Database used to store preprocessed information by Biblio-MetReS. Tables and their relationships. See section 3 of these supplementary materials for details.

Supplementary Tables

Supplementary Table 1. Orgaisms and genes used for benchmarking the application.

Organisms	Pathway	Genes to start reconstruction
Saccharomyces cerevisiae	Glycolysis	PGM1; FBA1; CDC19; ***ALL***
	Lysine biosynthesis	LYS21; ARO8; LYS9; ***ALL***
	RNA degradation	MTR3; MPP6; CAF16; RRP41 ***ALL***
Homo sapiens	Glycolysis	PGM1; ALDOA; PKLR; ***ALL***
	Lysine biosynthesis	AADAT; AASDH; AASS; ***ALL***
	RNA degradation	MTR3; MPP6; CNOT4; RRP41; ***ALL***
Escherichia coli	Glycolysis	Pgm; fbaB; pykF; ***ALL***
	Lysine biosynthesis	thrA; dapB; dapF; ***ALL***
	RNA degradation	rppH; rhlE; rnr; ***ALL***
Drosophila melanogaster	Glycolysis	Pgm; Ald; PyK; ***ALL***
	Lysine degradation	Lkr; CG9547; Gpp; ***ALL***
	RNA degradation	Rrp42; Mpp6; Cnot4; Rrp41; ***ALL***

Supplementary Table 2. Percentage of reduction in Biblio-MetReS run time due to pre-processing in controlled experiments.

Saccharomyces cerevisiae													
	Glycolysis				**Lysine Biosynthesis**				**RNA Degradation**				
	PGM1	FBA1	CDC19	ALL	LYS21	ARO8	LYS9	ALL	MTR3	MPP6	CAF16	RRP41	ALL
Pubmed	97.02	99.99	97.63	0	94.80	97.54	97.55	0	98.46	98.73	98.54	64.47	0
All DB	84.80	81.70	91.65	58.57	77.39	85.56	86.35	57.91	82.85	71.90	76.70	81.78	12.48
Homo sapiens													
	Glycolysis				**Lysine Biosynthesis**				**RNA Degradation**				
	PGM1	ALDOA	PKLR	ALL	AADAT	AASDH	AASS	ALL	MTR3	MPP6	CNOT4	RRP41	ALL
Pubmed	99.34	98.96	99.18	0	97.73	98.64	96.69	0	94.05	99.11	89.94	98.65	0
All DB	99.57	97.89	96.88	96.64	98.41	87.36	76.16	95.69	93.45	96.31	85.58	96.46	97.11
Drosophila melanogaster													
	Glycolysis				**Lysine Biosynthesis**				**RNA Degradation**				
	PGM	ALD	PYK	ALL	LKR	CG9547	GPP	ALL	RRP42	MPP6	CNOT4	RRP41	ALL
Pubmed	99.30	99.41	99.10	75.69	98.89	98.99	99.34	0	99.31	98.61	98.45	99.04	0
All DB	94.62	94.84	94.47	0	80.12	12.14	94.58	0	83.27	76.18	59.47	6.23	0
Escherichia coli													
	Glycolysis				**Lysine Biosynthesis**				**RNA Degradation**				
	PGM	FBAB	PYKF	ALL	THRA	DAPB	DAPF	ALL	RRPH	RHLE	RNR		ALL
Pubmed	97.46	97.19	91.21	0	97.93	97.37	94.95	0	96.24	94.57	96.52		0
All DB	92.69	93.00	81.94	88.63	84.76	89.65	86.31	87.15	84.42	68.82	84.97		85.43

Pubmed row results for searching Pubmed exclusively. ***All databases*** results for simultaneous search of all databases available in the application. For each gene and organism, Biblio-MetReS analyzes the documents that are found and stores the analysis. Then, the experiment is repeated and Biblio-MetReS find the same documents but now does not need to reanalyze them, as they have been pre-processed by the previous search. The percentage of time reduction between the first experiment and the second is calculated and the number is presented in the tables. Columns **ALL** represents the results obtained from searching for all the genes together. When all documents found in a new search have been preprocess, repeating this search will not further reduce its run-time, as we can see in some of the entries of the columns **ALL**.

3.6 References

1. Alves R, Sorribas A (2007) In silico pathway reconstruction: Iron-sulfur cluster biogenesis in Saccharomyces cerevisiae. BMC Syst Biol 1: 10. doi:10.1186/1752-0509-1-10.
2. Markowetz F, Spang R (2007) Inferring cellular networks--a review. BMC Bioinformatics 8 Suppl 6: S5. doi:10.1186/1471-2105-8-S6-S5.
3. NCBI (2013) MEDLINE Factsheet.
4. Hoffmann R, Valencia A (2005) Implementing the iHOP concept for navigation of biomedical literature. Bioinforma Oxf Engl 21 Suppl 2: ii252–258. doi:10.1093/bioinformatics/bti1142.
5. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, et al. (2011) The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res 39: D561–568. doi:10.1093/nar/gkq973.
6. Usié A, Karathia H, Teixidó I, Valls J, Faus X, et al. (2011) Biblio-MetReS: A bibliometric network reconstruction application and server. BMC Bioinformatics 12: 387. doi:10.1186/1471-2105-12-387.
7. Camon EB, Barrell DG, Dimmer EC, Lee V, Magrane M, et al. (2005) An evaluation of GO annotation retrieval for BioCreAtIvE and GOA. BMC Bioinformatics 6 Suppl 1: S17. doi:10.1186/1471-2105-6-S1-S17.
8. Kim J-D, Ohta T, Pyysalo S, Kano Y, Tsujii J (n.d.) Overview of BioNLP'09 shared task on event extraction. Proceedings of the Workshop on Current Trends in Biomedical Natural Language Processing: Shared Task. pp. 1–9.
9. Kim J-D, Pyysalo S, Ohta T, Bossy R, Nguyen N, et al. (2011) Overview of BioNLP Shared Task. BioNLP Shared Task '11 Proceedings of the BioNLP Shared Task 2011 Workshop. pp. 1–6.
10. Chen Y, Liu F, Manderick B (2010) BioLMiner and the BioCreative II.5 challenge. BMC Bioinformatics 11: P6. doi:10.1186/1471-2105-11-S5-P6.
11. Huang M, Ding S, Wang H, Zhu X (2008) Mining physical protein-protein interactions from the literature. Genome Biol 9: S12. doi:10.1186/gb-2008-9-s2-s12.
12. Vazquez M, Krallinger M, Leitner F, Valencia A (2011) Text Mining for Drugs and Chemical Compounds: Methods, Tools and Applications. Mol Informatics 30: 506–519. doi:10.1002/minf.201100005.
13. Yang Z, Lin H, Li Y (2008) Exploiting the contextual cues for bio-entity name recognition in biomedical literature. J Biomed Informatics 41: 580–587. doi:10.1016/j.jbi.2008.01.002.
14. Peng F, Schuurmans D (2003) Combining Naive Bayes and n-Gram Language Models for Text Classification. In: Sebastiani F, editor. Advances in Information Retrieval. Lecture Notes in Computer Science. Springer Berlin / Heidelberg, Vol. 2633. pp. 547–547. Available: http://www.springerlink.com/content/mturdqvxh3vv7k4f/abstract/. Accessed 10 October 2012.
15. Malouf R (2002) Markov models for language-independent named entity recognition. proceedings of the 6th conference on Natural language

learning - Volume 20. COLING-02. Stroudsburg, PA, USA: Association for Computational Linguistics. pp. 1–4. Available: http://dx.doi.org/10.3115/1118853.1118872. Accessed 10 October 2012.
16. Sasaki Y, Tsuruoka Y, McNaught J, Ananiadou S (2008) How to make the most of NE dictionaries in statistical NER. BMC Bioinformatics 9: S5. doi:10.1186/1471-2105-9-S11-S5.
17. Lu Z, Hirschman L (2012) Biocuration workflows and text mining: overview of the BioCreative 2012 Workshop Track II. Database J Biol Databases Curation 2012: bas043. doi:10.1093/database/bas043.
18. Gene Ontology Consortium (2013) Gene Ontology annotations and resources. Nucleic Acids Res 41: D530–535. doi:10.1093/nar/gks1050.
19. Kotera M, Hirakawa M, Tokimatsu T, Goto S, Kanehisa M (2012) The KEGG databases and tools facilitating omics analysis: latest developments involving human diseases and pharmaceuticals. Methods Mol Biol Clifton NJ 802: 19–39. doi:10.1007/978-1-61779-400-1_2.
20. Mi H, Lazareva-Ulitsky B, Loo R, Kejariwal A, Vandergriff J, et al. (2005) The PANTHER database of protein families, subfamilies, functions and pathways. Nucleic Acids Res 33: D284–288. doi:10.1093/nar/gki078.
21. Wu CH, Arighi CN, Cohen KB, Hirschman L, Krallinger M, et al. (2012) BioCreative-2012 Virtual Issue. Database 2012: bas049–bas049. doi:10.1093/database/bas049.
22. Lin J (2009) Is searching full text more effective than searching abstracts? BMC Bioinformatics 10: 46. doi:10.1186/1471-2105-10-46.

Chapter 4. CheNER I

Bioinformatics 2013
doi:10.1093/bioinformatics/btt639

CheNER: Chemical Named Entity Recognizer

Anabel Usié, Rui Alves, Francesc Solsona, Miguel Vázquez and Alfonso Valencia

Abstract

Motivation: Chemical named entity recognition is used to automatically identify mentions to chemical compounds in text, and is the basis for more elaborate information extraction. However, only a small number of applications are freely available to identify such mentions. Particularly challenging and useful is the identification of IUPAC chemical compounds, which due to the complex morphology of IUPAC names requires more advanced techniques than that of brand names.

Results: We present CheNER, a tool for automated identification of systematic IUPAC chemical mentions. We evaluated different systems using an established literature corpus to show that CheNER has a superior performance in identifying IUPAC names specifically, and that it makes better use of computational resources.

Availability: http://metres.udl.cat/index.php/9-download/4-chener, http://chener.bioinfo.cnio.es/

Supplementary information: Both web sites above include the user manual for the software. Supplementary materials accompany this publication.

4.1 Introduction

Automated NER (Named Entity Recognition) of chemical compounds is receiving increased attention from researchers because it can facilitate the application of Information Extraction to the pharmaceutical treatment of diseases and to understanding how those compounds modulate gene/protein activities. Chemical NER draws from the experience in performing gene and protein NER [1], but differs from it in three ways.

First, catalogs of names and compositions of chemical compounds have been traditionally less accessible. Fortunately, freely available chemical databases such as PubChem [2] or DrugBank [3] are helping to correct this issue. This makes it possible to do NER of common drug names such as "*Aspirin*" or "*Acetone*" by using a dictionary-based approach.

Second, the complexities and the variability in the morphological structure of systematic IUPAC (Union of Pure and Applied Chemistry) chemical names [4] makes it impossible to create a finite dictionary of such names. This poses the main challenge for NER of chemical names [5]. IUPAC names can be simple words, or contain different punctuation marks, sequences of numbers separated by commas, etc. They can also be combined in different forms (for example "*18-bromo-12-butyl-11-chloro-4,8-diethyl-5-hydroxy-15-methoxy*"), making it impossible to enumerate them all. This means that NER of such names cannot be done using a dictionary matching, requiring alternative approaches.

Third, systematic nomenclatures of chemicals, like IUPAC, can be used directly to unambiguously derive their chemical structure.

The number of applications that are freely available to do NER of common and systematic names of chemical compounds is still incipient, and their usability, efficiency, and accuracy are far from perfect. To help alleviate these problems, in this work we present and benchmark CheNER, a machine learning application based on CRFs that performs NER of IUPAC chemical entities with improved performance over comparable tools.

4.2 Methods

CheNER uses linear Conditional Random Fields (CRFs) to predict the locations of IUPAC entity mentions in text. CRFs are a probabilistic framework for the labeling or segmentation of sequential data [6].

The training and benchmarking of the application was done using the corpora provided by Kolářik and Klinger [7,8]. The corpora are divided into a training corpus (***TrainC***, 463 abstracts, 5072 annotated entities), a Medline test corpus with a small number of entities (***MedlineC***, 1000 abstracts, 165 annotated entities), and an evaluation corpus with a large number of entities (***EvalC***, 100 abstracts, 1310 annotated entities). All corpora contain annotated chemical entities written using the IUPAC nomenclature as well as other types of chemical names. CheNER's CRF was trained on ***TrainC***. Its performance was subsequently evaluated independently on both, ***MedlineC*** and ***EvalC***.

In training our CRF we defined a set of features and tested different combinations of them, together with two types of tokenization (A: by spaces, B: by punctuation marks), different orders of CRF (1 or 2), and different sizes of offsets conjunction or sliding windows (0 or 1), which creates a new additional feature of a token by conjoining its features with those of the n (n=0, n=1) surrounding tokens. We then selected the best combination, indicated by the highest F-score value obtained in cross-validation over the training set, as a model to use in the evaluation. The selected model performs with an F-score value of 80.20% (Precision: 82,84%; Recall: 77.74%), uses a 2nd order CRF, an offset conjunction of 1, tokenization type A, and a particular set of features described in the supplementary materials. To mark chemical mentions and establish borders between tokens during training we used the IOB labeling scheme [5]. Details about the tested sets of features, training and evaluation corpora, training process, modeling assumption, performance, and selection, are described in sections 1 to 3 of the supplementary materials.

4.3 Results

4.3.1 Comparative performance for NER of chemical names

The predictive capability of CheNER for IUPAC names, was evaluated using the ***EvalC*** and the ***MedlineC*** corpora, performing the evaluation by comparing the system output to a gold standard in terms of the precision (p), recall (r) and F-score (F).

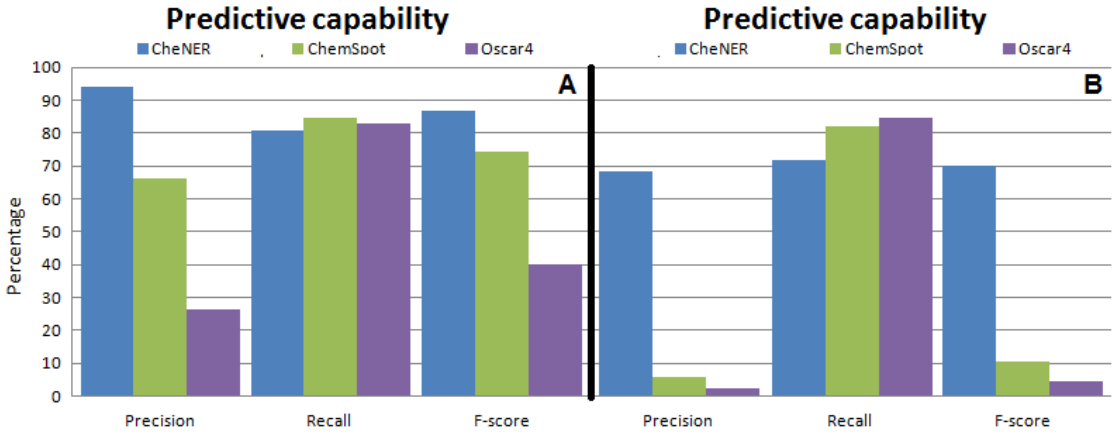


Figure 1. Predictive capability of the different tools identifying IUPAC entities over: (A) the EvalC corpus and (B) MedlineC corpus. We measure the ability of the three tools to specifically identify IUPAC chemical entities in the two corpora.

There are, to our knowledge only two other freely available tools for chemical NER. These are ChemSpot [9] and OSCAR4 [10,11]. To compare CheNER's performance to that of those tools, we use the three applications to independently annotate ***MedlineC*** and ***EvalC*** and compare the results. Our analysis shows that CheNER outperforms the other two applications in the experiments regarding IUPAC names alone (see Figure 1), due to the fact that it was trained specifically for them. Note that OSCAR and ChemSpot do not differentiate between IUPAC and other types of chemical entities and will detect entities that, albeit chemical, will not be IUPAC and will register as false positives. To make the three methods comparable we ignore non-IUPAC entities that are annotated in the corpora when evaluating performance. Unfortunately the MedlineC corpus does not annotate non-IUPAC entities, so this corpus can only be compared in terms of recall. We find that CheNER's

performs better than OSCAR4 and ChemSpot identifying IUPAC names. Details are given in section 4 of supplementary materials.

Given that CheNER has been trained in the specialized task of recognizing IUPAC names, it is not surprising that when applied to non-IUPAC names it does not perform at the levels of other systems (see section 4 of supplementary materials).

4.3.2 Comparative use of hardware resources

We also evaluated how efficiently ChemSpot, OSCAR4 and CheNER use computing resources. We found that CheNER requires less physical memory, running in computers that have less than 3 GB of RAM, compared with minimum of 3 and 12 GB of RAM required by OSCAR4 and ChemSpot respectively (see Supplementary Figure 3, Supplementary Figure 4 and section 4 of the supplementary materials for details).

4.4 Discussion

Because IUPAC names are the standard in important types of documents, such as patents, and the chemical structure is often derivable from the mention itself, it is important to have an application specifically devised for their identification. Given the potentially infinite number of IUPAC entities it is not feasible to develop a dictionary based approach to identify them, and NLP methods are more suitable to identify those entities. Thus, we developed CheNER, a named entity recognition approach for finding IUPAC names in text, using CRFs.We demonstrate that CheNER annotates IUPAC names in documents with a better F-score than ChemSpot and OSCAR4. CheNER is the only tool that is specifically developed to identify only such names while ChemSpot and OSCAR4 do not differentiate between entity types.

We also show that CheNER needs less memory and CPU than the others to perform the same tasks. In addition, CheNER is self-contained, requiring only that Java is installed to run, which makes it easier to integrate in other systems.

4.5 Supplementary Materials

This supplementary material contains the following:

- 1: Features used
- 2: Detailed information about the corpora used
- 3: The training process
- 4: Comparison of CheNER to other chemicals tools
- 5: Feature removal
- Supplementary Figure 1
- Supplementary Figure 2
- Supplementary Figure 3
- Supplementary Figure 4
- Supplementary Figure 5
- Supplementary Figure 6
- Supplementary Figure 7
- Supplementary Figure 8
- Supplementary Figure 9
- Supplementary Figure 10
- Supplementary Figure 11
- Supplementary Figure 12
- Supplementary Figure 13
- Supplementary Table 1
- Supplementary Table 2
- Supplementary Table 3
- Supplementary Table 4
- Supplementary Table 5

4.5.1 *Features set used*

The set of features used to train the CRF model is shown in Supplementary Table 1. Some of these features were identified by previous studies as being the most discriminative in the identification of gene and protein names [7].

4.5.2 Detailed information about the corpora used

These corpora include chemical entities of different types: IUPAC and PARTIUPAC employ multi-word systematic names and partial chemical names [4], MODIFIER names classify chemical modifiers [12], FAMILY type includes generic chemical entities such as "purine", or "sugar", [8], ABB type includes those chemicals that are abbreviations of names [13], SUM type includes the chemical and/or atomic formulas for the different compounds [4], and TRIVIAL are common names composed by single word terms. Supplementary Table 2 summarizes the entity composition of the three corpora.

While being an invaluable resource for this kind of work, and the best dataset to date for this task, close examination of these corpora reveals a caveat: while ***TrainC*** and ***MedlineC*** only have 3 entities in common, 46.16% of the entities in ***EvalC*** are also found in ***TrainC***. To our knowledge this fact had not been highlighted before.

4.5.3 The training process

First, we look for the combination of parameters and features sets (modeling assumptions) that create the model with the best F-Score performance. The performance of each modeling assumption was assessed using 5-fold cross-validation over the ***TrainC*** corpus. In cross validation, the dataset is randomly divided into several equally sized chunks or folds (5 in our case), and each fold is in turn used to validate a model trained using the other folds. Second, once we have determined the best model assumption as the one with the highest averaged F-Score performance over all 5 folds, it is used for the final training over the complete ***TrainC*** corpus.

The different modeling assumptions are detailed in Supplementary Table 3. Each modeling assumption explores combinations of 3 different characteristics. First, two tokenization types were used. In type A only blank spaces were used to delimit tokens, while in type B several other characters were also considered as possible token delimiters (see Supplementary Table 3 for details). Second,

CRFs of order 1 and 2 were used for the training. Third, two values of OC were considered (0: no context considered, and 1: features from tokens that immediately precede and follow the token of interest are considered in predicting that token's label). We obtained a top F-Score performance of 80.20% (Precision: 82.84%; Recall: 77.74%), using a 2nd order of CRF, an offset conjunction of 1, tokenization type A (by spaces), and the following set of features: MF, Toc, PS, W, WB, L, BNS and SW.

For each modeling assumption all the features set were used (see Supplementary Table 1). A: Uses blank spaces as delimiters of tokens. B: Uses blank space, punctuation marks (dots, dashes, etc) and parenthesis as delimiters if tokens.

4.5.4 Comparison of CheNER to other chemicals tools

A set of four experiments was conducted to benchmark the performance of different tools performing chemical NER: (1) IUPAC entities in ***EvalC***, (2) IUPAC entities in ***MedlineC***, (3) all entity types in ***EvalC***, and (4) all entity types in ***MedlineC***. The tools that were comparatively evaluated with respect to CheNER were OSCAR4 [10,11] and ChemSpot [9].

The results obtained in experiment 1 and 2, the identification of IUPAC names, are shown in Figure 1, respectively (see main document), and in Supplementary Table 4. CheNER has the best global performance in experiments 1 and 2, as measured by the F-score. To make the comparisons reliable, we eliminate all non-IUPAC entities annotated in each corpora by OSCAR4 and ChemSpot. This must be done to account for the fact that ChemSpot and OSCAR4 do not differentiate between the types of chemical entities they annotate, which would lead to an artificially low precision for these two applications. In addition we manually check the FP results of each tool. We find that most of these FP are real chemical entities that failed to be originally annotated in the corpora. When these are eliminated and only real FP entities are considered, both ChemSpot and OSCAR4 always reported a higher number of FP entities than CheNER. Thus, CheNER has more precision than

either OSCAR4 or ChemSpot, and a comparable recall to either of the other tools. Overall, CheNER has the highest F-score in identifying IUPAC chemical entities.

The results obtained in experiment 3 and 4, are shown in Supplementary Figure 1 and Supplementary Figure 2, respectively. As expected, CheNER has the worst F-score performance in experiment 3, as it mostly recognizes IUPAC names.

The low performance of ChemSpot and OSCAR4 in ***MedlineC*** is a consequence of the way that this corpus is annotated. While only IUPAC and MODIFIER chemical names are annotated, other types of chemicals are also present. A partial manual analysis of the results reveals that many of the false positive entities identified by OSCAR4 and ChemSpot are non IUPAC chemical entities. Therefore, the only real comparison of predictive capability that we can do between the three tools while tagging entities of ***MedlineC*** in this experiment is in term of recall. We show that CheNER has the worst recall in this experiment, mostly due to a failure in identifying MODIFIER entities.

Summarizing the results, CheNER is the application that more accurately identifies IUPAC chemicals names. It is also the only available tool that was specifically developed to identify this nomenclature. In this context, the three tools have a similar recall and CheNER's outperforms the other tools based on higher precision.

We also evaluated how efficiently ChemSpot, OSCAR4 and CheNER use available computing resources. To do so, we ran each application on the same machine (i7 processor, with four CPUs and 20GB of RAM) and monitored the consumption of main memory and CPU of the system during the annotation of the two evaluation corpora. Supplementary Figure 3 shows that CheNER uses less memory than the others (<1GB-3.5% of available RAM). In contrast ChemSpot uses the most amount of memory (>7GB-35.7%) and OSCAR4 needed an intermediate amount of memory (<2GB-5.5%). To estimate total memory requirements we must consider also the memory required to run the JVM (Java Virtual Machine). Supplementary Figure 3 shows that CheNER also

requires less CPU than the other applications. In addition, CPU usage is more constant over time than that of ChemSpot and OSCAR4.

4.5.5 Feature Removal

The feature set used to train our CRF is shown in Supplementary Table 1. Previous work identified some of these features as being the most discriminative in the identification of chemical names [7]. However not systematic study was done to understand how the different features interacted in the prediction of chemical names in text.

Here we performed such a study in the form of feature removal experiments. These feature removal experiments allowed us to establish the effect of the different features on the performance of the CRF. Each set of experiments is identified as leave- n -out, with n ranging from 0 to 8 and representing the number of features that are simultaneously removed from the feature set described in Supplementary Table 1 before retraining the CRF. In each set of experiments we remove all possible combinations of n features, retrain and then evaluate the performance of our CRF. Note that even with all the features removed, the actual token is also used by the CRF as a feature in the training; thus, even in the leave-8- out removal experiment the labels are still conditioned by the tokens themselves, by the two previous labels (2nd order linear-chain CRF) and by the preceded and followed token (OC parameter set to 1). This fact will be fundamental to understand the results of the feature removal experiments.

Our experiments indicate that the F-score performance of the system increased as more features where considered. This is summarized in Supplementary Figures 5 and 6, where one can see that the average F-score performance of the CRF increases as the number of features used to train the algorithm also increases. The detailed results of the experiments are presented in Supplementary Figure 7, on performance of progressively removing features, starting with the complete feature set, and ending with a 'naked' model, with

no features others than the tokens themselves for ***EvalC*** and ***MedlineC***, respectively.

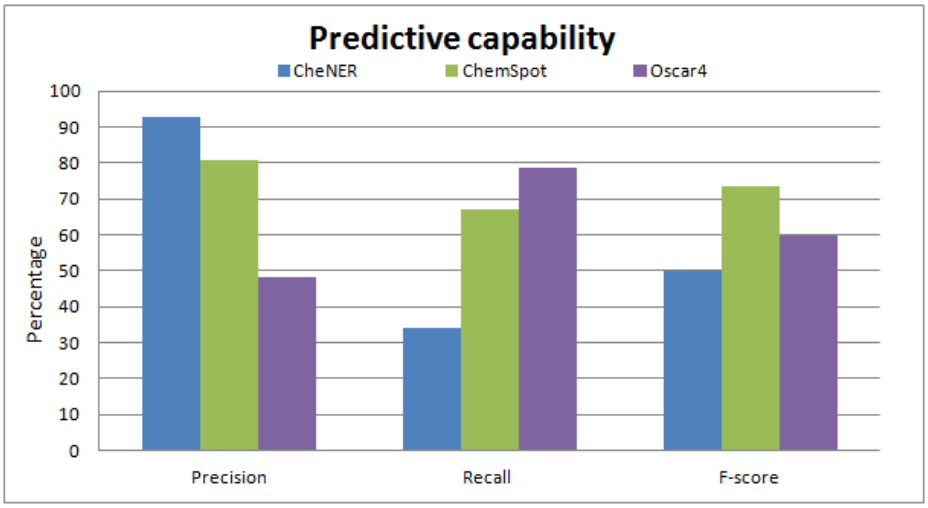
Our results show that the feature whose removal decreased performances the most is PS, which is consistent with previous findings [7]. This was also the feature that generates the best performance in absence of all others. The feature ToC and W contribute the least to performance. In fact, their simultaneous removal seems to improve results, even if this improvement is not statistically significant. Supplementary Table 5 shows the best and worst combinations of features for each removal set of experiments.

One of the most striking results of the feature removal experiments was the difference in performance between the two evaluation corpora. Although the performance over both corpora improved as more features where used to train the CRF, the degree of improvement was very different. The F-score performance in the ***MedlineC*** corpus ranged from 0% on the CRF trained with no features to 23% on the CRFs trained using just one feature. That performance further improved to 63% on the CRF trained using all 8 features. Note that all these CRFs include the "token literal" feature.

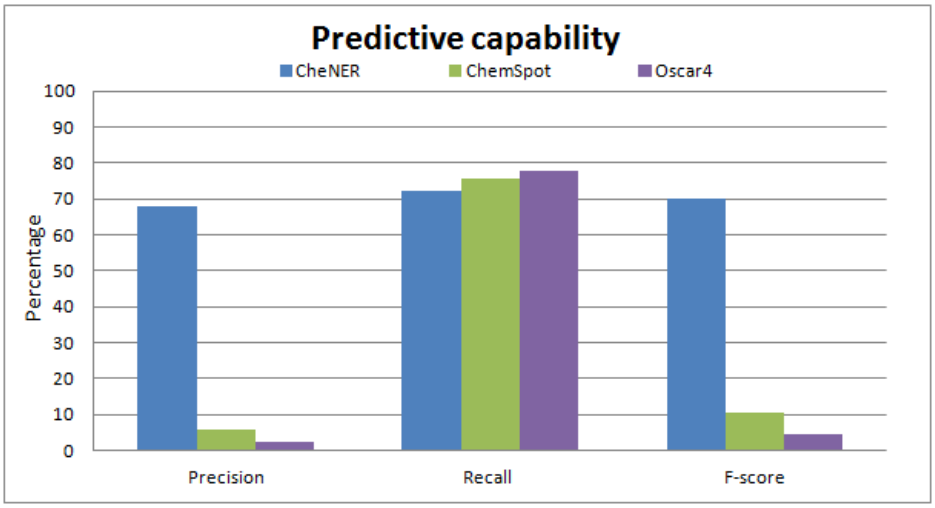
The ***EvalC***, however, presented a different behavior. The trained CRF had an F-score performance of around 70%, even in the leave-8-out experiment set. Further analysis revealed that almost half of the compounds mentioned in the evaluation set of the corpus appeared verbatim in the training set. This led us to hypothesize that the "token literal" features used to train the CRFs in all n-out experiments was sufficient to identify most chemical in this specific evaluation set. To test this hypothesis we performed an additional experiment and intentionally excluded this feature, retraining the leave-8-out CRF and reevaluating its performance on ***EvalC***. As predicted, this CRF is unable to correctly identify any chemical names (F-score=0). Notably, when the "token literal" feature is intentionally excluded, and the CRF is retrained in the leave-0-out experiment set, it retains the same F-score performance level observed in the leave-0-out experiment that included the "token literal" feature (F-score=77). This emphasizes that the CRF learns by combining two modes of

learning. One mode is by using the derived features to identify new chemical terms and the other is by memorizing the exact terms it finds on the training corpus.

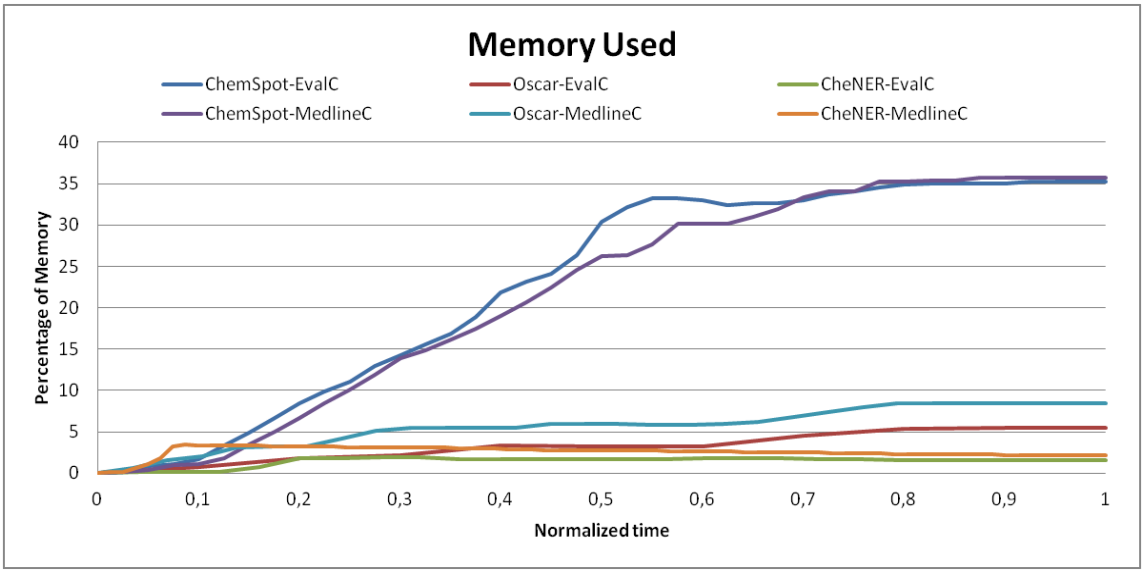
Supplementary Figures



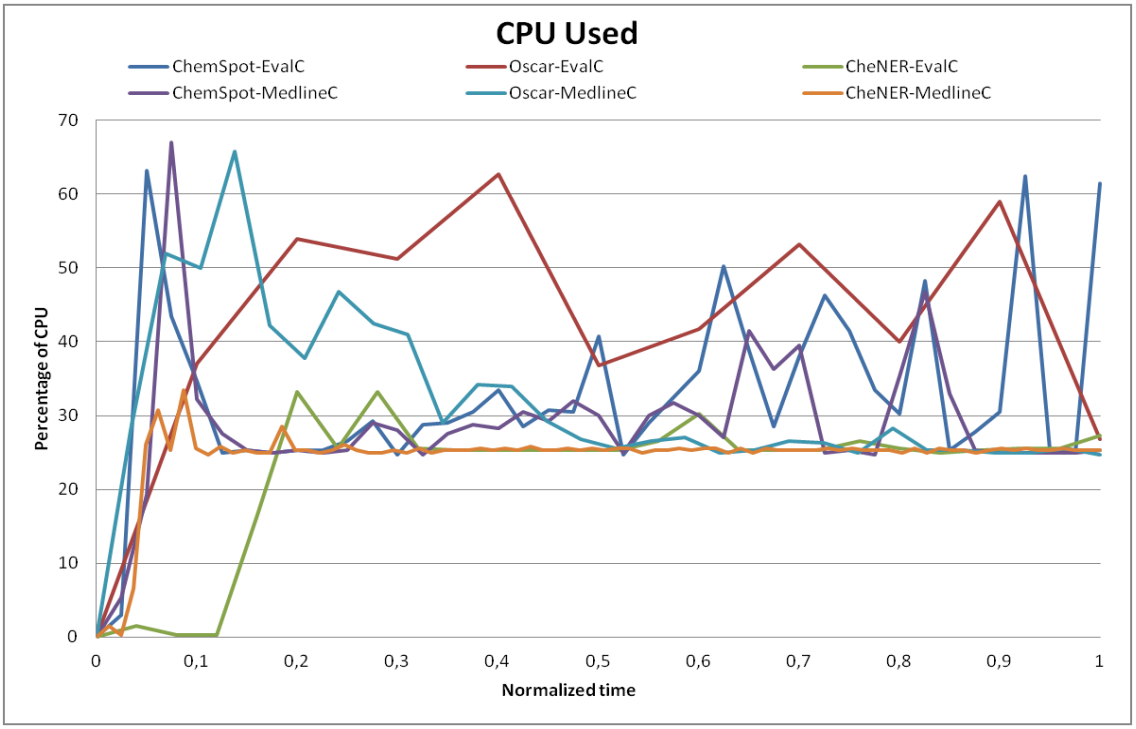
Supplementary Figure 1. Predictive capability of the different tools identifying all type of entities over the EvalC corpus in terms of precision, recall and F-score. In this experiment we measured the F-score of the three tools in identifying all types of chemical entities.



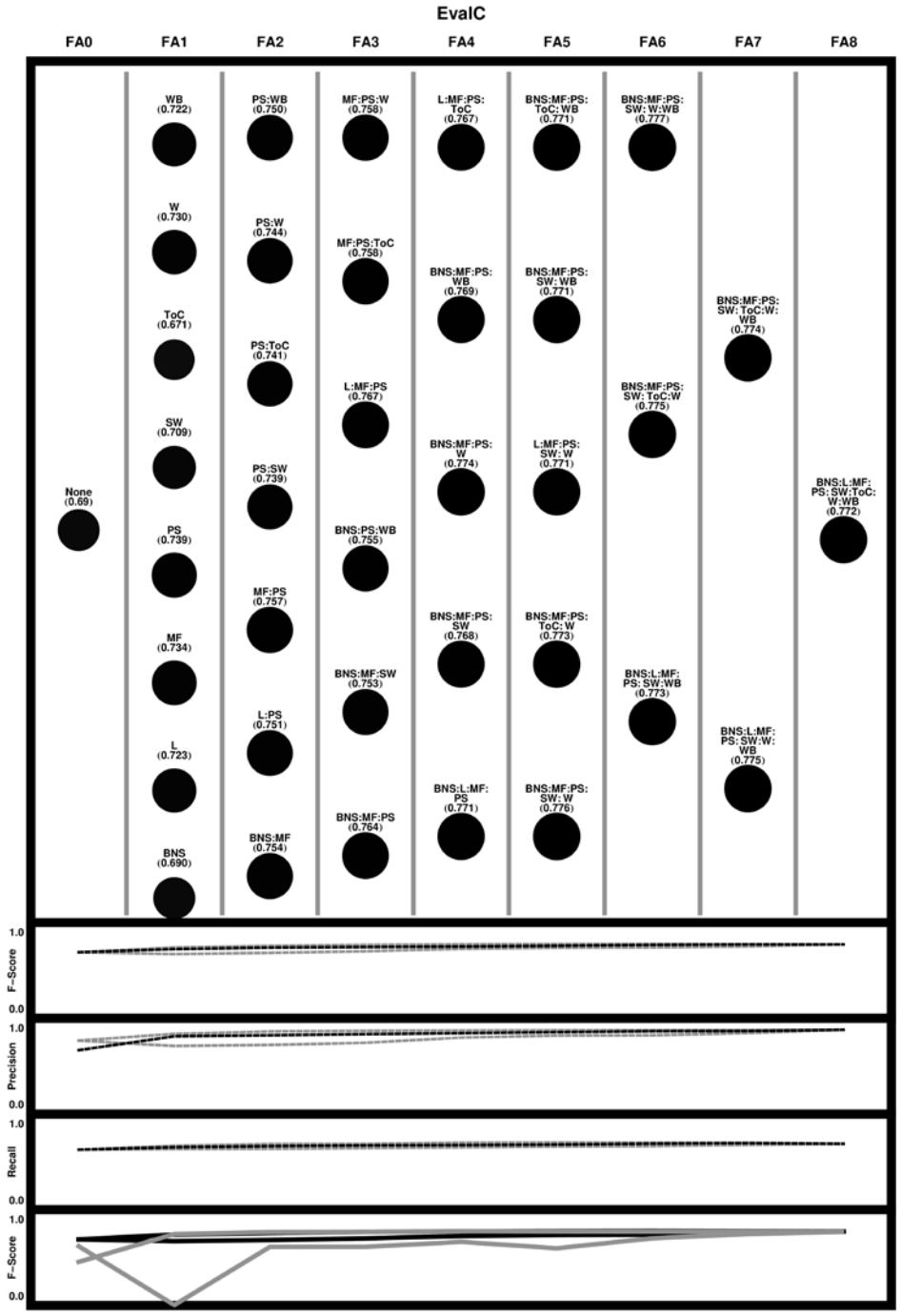
Supplementary Figure 2. Predictive capability of the different tools identifying all type of entities over the MedlineC corpus in terms of precision, recall and F-score. In this experiment we measured the F-score of the three tools in identifying all types of chemical entities.



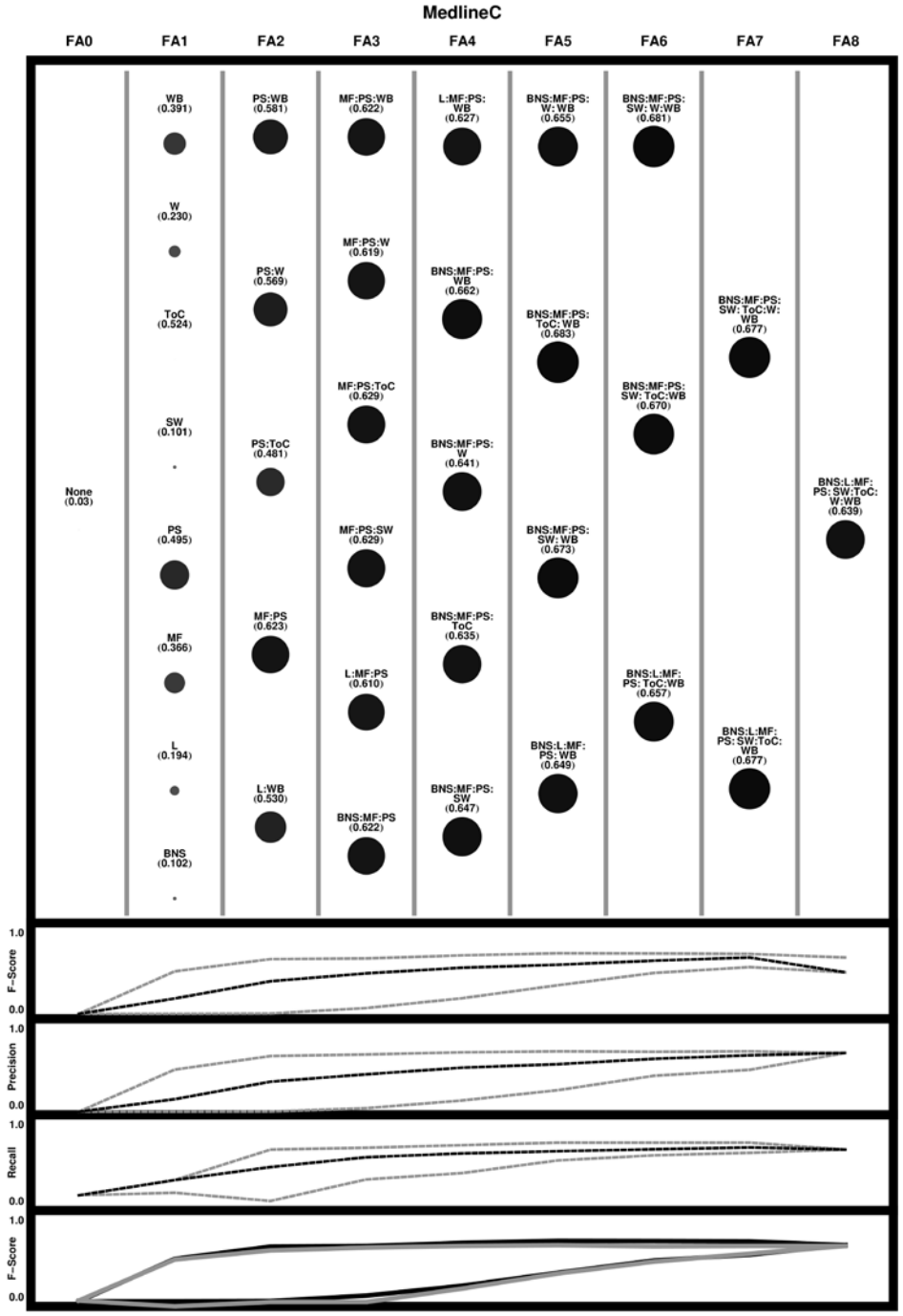
Supplementary Figure 3. Memory consumption for each application during the annotation of both corpora, EvalC and MedlineC. Time was normalized in each run. Absolute time for the same run is similar for the three programs**.**



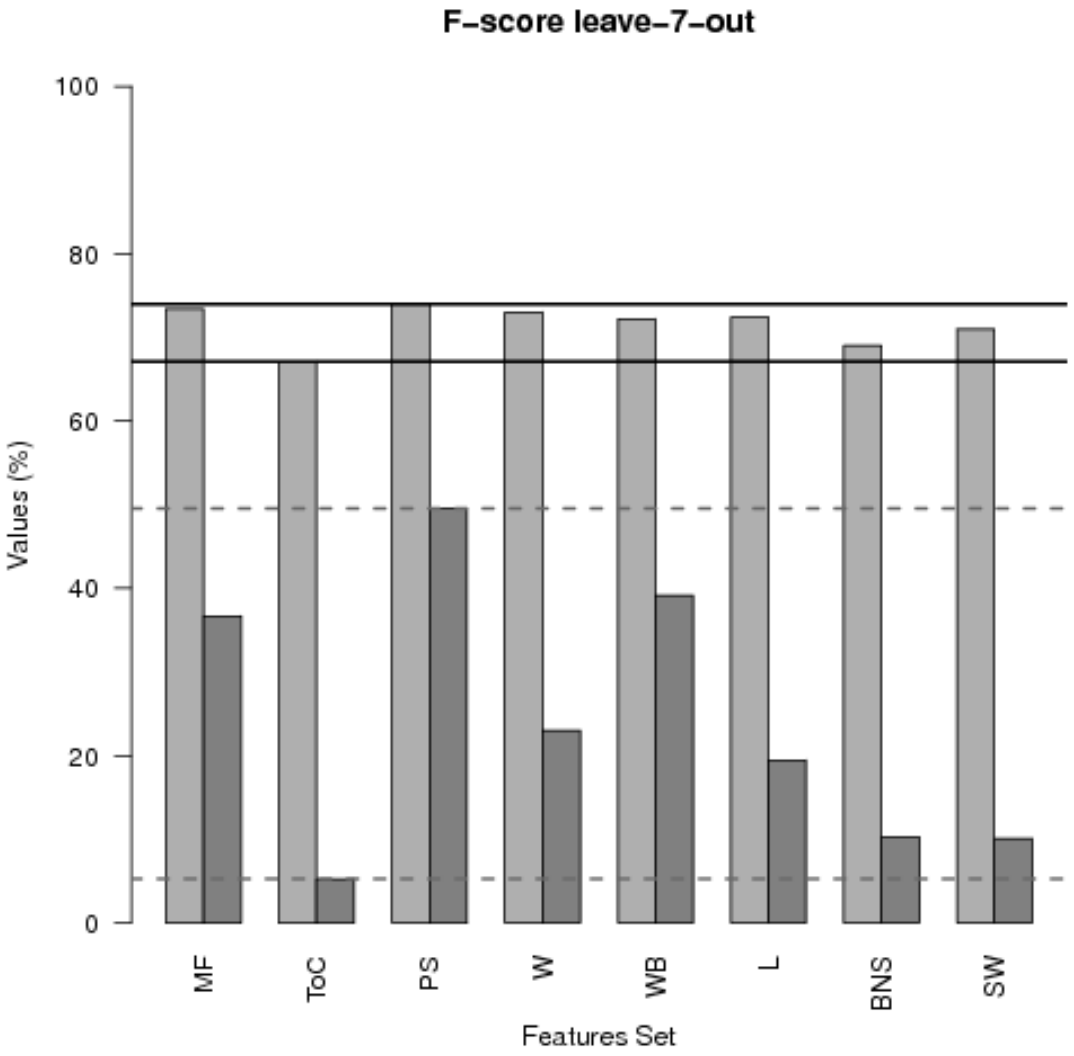
Supplementary Figure 4. CPU consumption for each application during the annotation of the EvalC and MedlineC corpora. Time was normalized in each run. Absolute time for the same run is similar for the three programs.



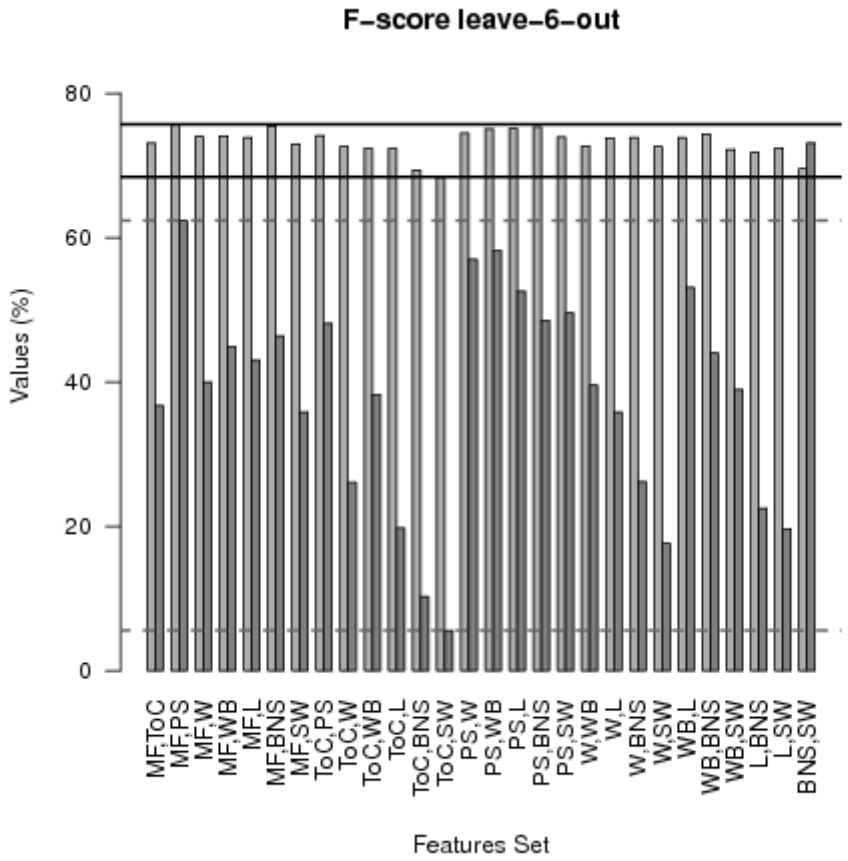
Supplementary Figure 5. Best CRF performers in the leave-n-out sets of experiments (1≤n≤7) over ***EvalC***. The FA0 column corresponds to the performance of the CRF trained without considering any additional feature. The FA1-FA7 columns correspond to the CRF performance in the leave-one-out to leave-seven-out sets of experiments. The FA8 column corresponds to the CRF performance when considering all features. Each column presents the maximum F-score for each of the eight features. The width of the circle for each feature is proportional to the F-score. The color can go from white (low F-score) to black (high F-score). The upper three lower panels indicate maximum (upper gray line), median (black line), and minimum (lower gray line) values for the F-score, precision, and recall in each set of experiments. The last lower panels indicate the maximum (upper black line) and minimum (lower black line) with OC, and the maximum (upper gray dash-line) and minimum (lower grey dash-line) without OC.



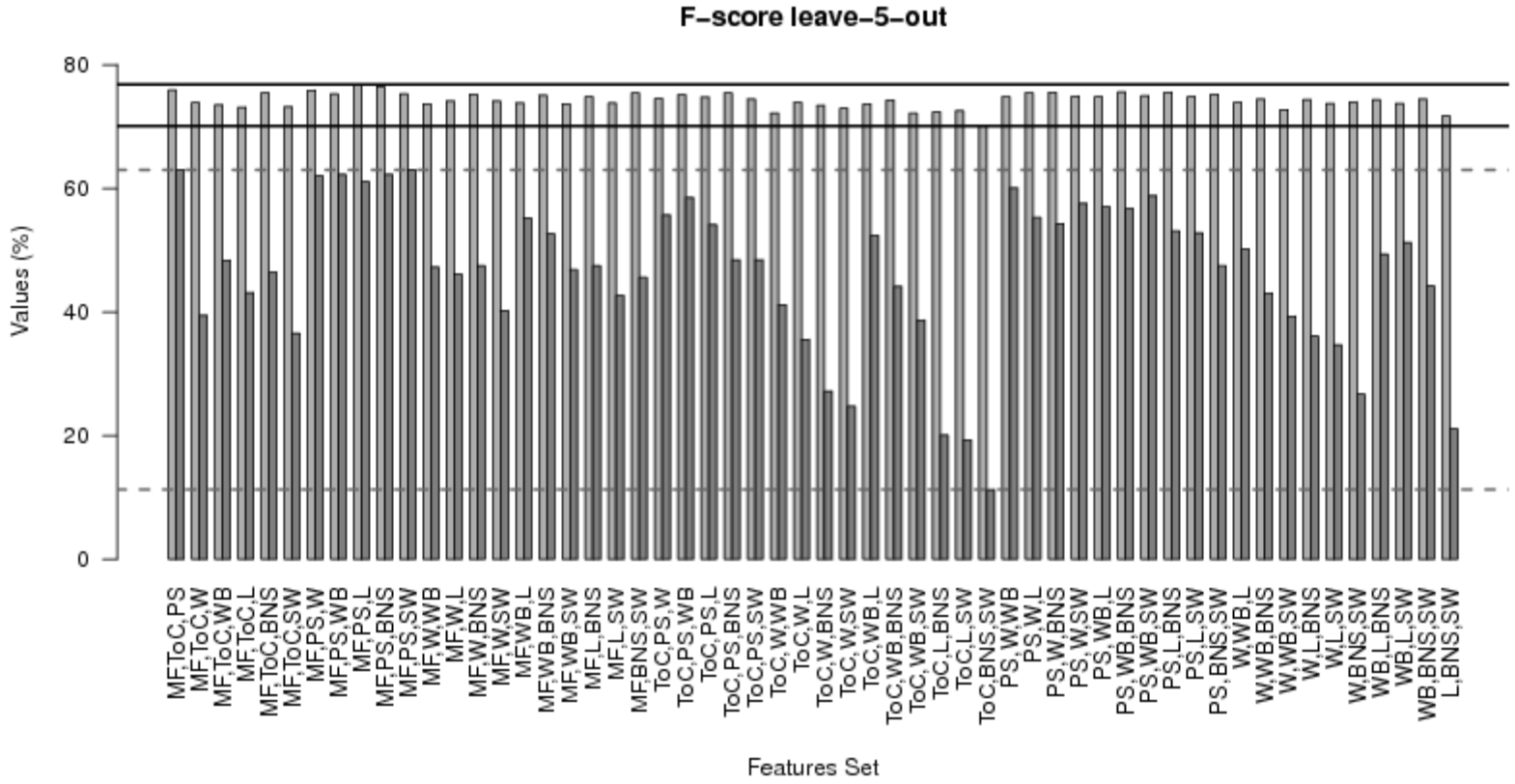
Supplementary Figure 6. Best CRF performers in the leave-n-out sets of experiments (1≤n≤7) over ***MedlineC***. The FA0 column corresponds to the performance of the CRF trained without considering any additional feature. The FA1-FA7 columns correspond to the CRF performance in the leave-one-out to leave-seven-out sets of experiments. The FA8 column corresponds to the CRF performance when considering all features. Each column presents the maximum F-score for each of the eight features. The width of the circle for each feature is proportional to the F-score. The color can go from white (low F-score) to black (high F-score). The upper three lower panels indicate maximum (upper gray line), median (black line), and minimum (lower gray line) values for the F-score, precision, and recall in each set of experiments. The last lower panels indicate the maximum (upper black line) and minimum (lower black line) with OC, and the maximum (upper gray dash-line) and minimum (lower grey dash-line) without OC.



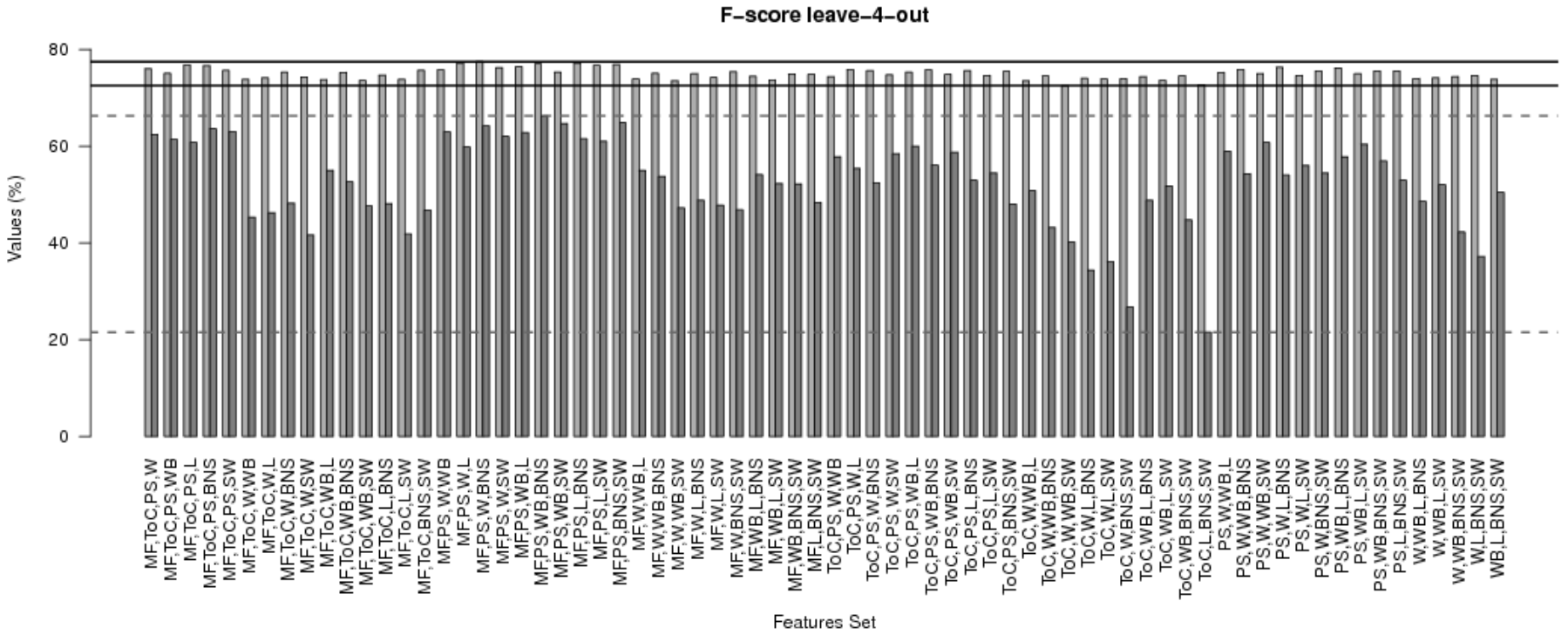
Supplementary Figure 7. Performance in the leave-one-out experiment set. Bars represents the F-scores obtained for both corpora, ***EvalC*** (light-gray bar) and ***MedlineC*** (dark-gray bar). Lines indicate the best and worst performance in each corpus. The upper black line and lower black line indicate the maximum and minimum, respectively, in ***EvalC***. The upper gray dash-line and lower gray dash-line indicate the maximum and minimum, respectively, in ***MedlineC***.



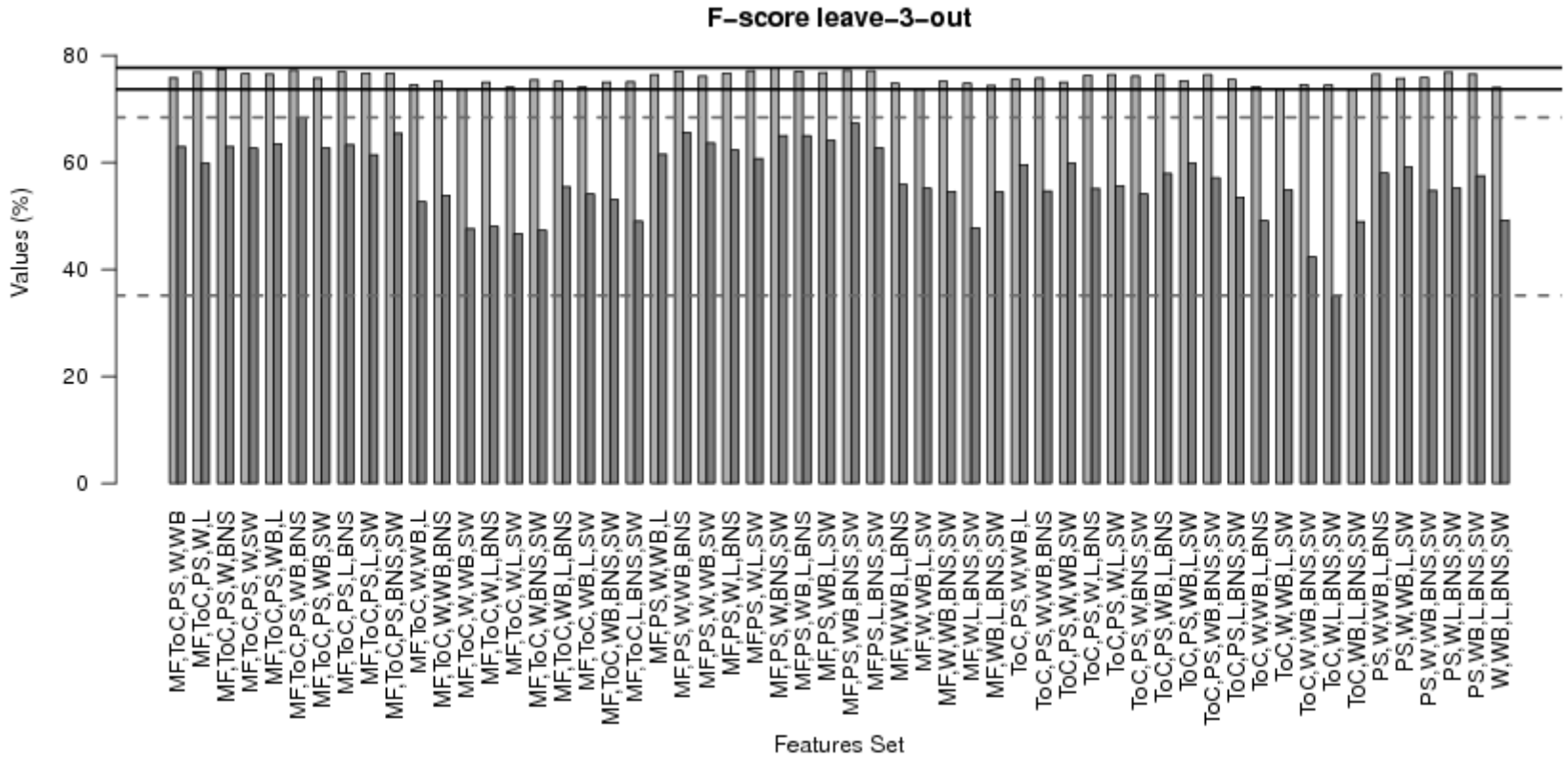
Supplementary Figure 8. Performance in the leave-two-out experiment set. Bars represents the F-scores obtained for both corpora, ***EvalC*** (light-gray bar) and ***MedlineC*** (dark-gray bar). Lines indicate the best and worst performance in each corpus. The upper black line and lower black line indicate the maximum and minimum, respectively, in ***EvalC***. The upper gray dash-line and lower gray dash-line indicate the maximum and minimum, respectively, in ***MedlineC***.



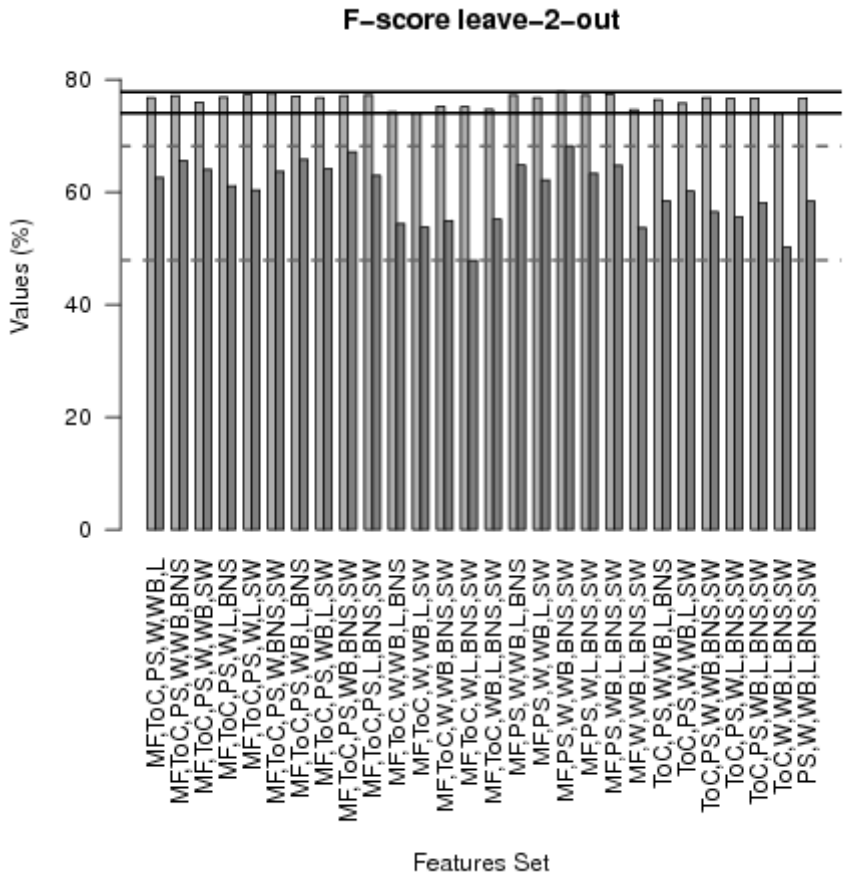
Supplementary Figure 9. Performance in the leave-three-out experiment set. Bars represents the F-scores obtained for both corpora, ***EvalC*** (light-gray bar) and ***MedlineC*** (dark-gray bar). Lines indicate the best and worst performance in each corpus. The upper black line and lower black line indicate the maximum and minimum, respectively, in ***EvalC***. The upper gray dash-line and lower gray dash-line indicate the maximum and minimum, respectively, in ***MedlineC***.



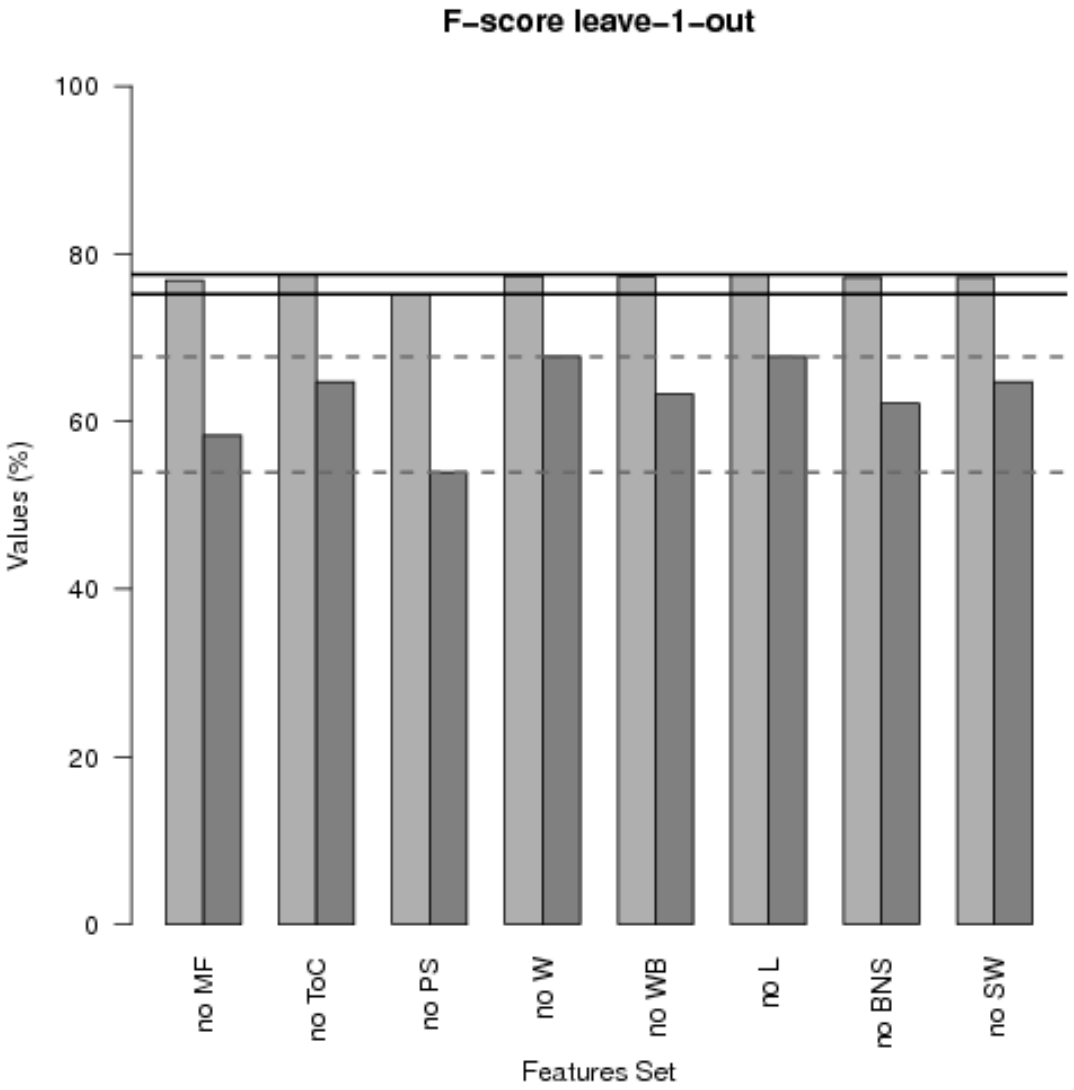
Supplementary Figure 10. Performance in the leave-four-out experiment set. Bars represents the F-scores obtained for both corpora, ***EvalC*** (light-gray bar) and ***MedlineC*** (dark-gray bar). Lines indicate the best and worst performance in each corpus. The upper black line and lower black line indicate the maximum and minimum, respectively, in ***EvalC***. The upper gray dash-line and lower gray dash-line indicate the maximum and minimum, respectively, in ***MedlineC***.



Supplementary Figure 11. Performance in the leave-five-out experiment set. Bars represents the F-scores obtained for both corpora, ***EvalC*** (light-gray bar) and ***MedlineC*** (dark-gray bar). Lines indicate the best and worst performance in each corpus. The upper black line and lower black line indicate the maximum and minimum, respectively, in ***EvalC***. The upper gray dash-line and lower gray dash-line indicate the maximum and minimum, respectively, in ***MedlineC***.



Supplementary Figure 12. Performance in the leave-six-out experiment set. Bars represents the F-scores obtained for both corpora, ***EvalC*** (light-gray bar) and ***MedlineC*** (dark-gray bar). Lines indicate the best and worst performance in each corpus. The upper black line and lower black line indicate the maximum and minimum, respectively, in ***EvalC***. The upper gray dash-line and lower gray dash-line indicate the maximum and minimum, respectively, in ***MedlineC***.



Supplementary Figure 13. Performance in the leave-seven-out experiment set. Bars represents the F-scores obtained for both corpora, ***EvalC*** (light-gray bar) and ***MedlineC*** (dark-gray bar). Lines indicate the best and worst performance in each corpus. The upper black line and lower black line indicate the maximum and minimum, respectively, in ***EvalC***. The upper gray dash-line and lower gray dash-line indicate the maximum and minimum, respectively, in ***MedlineC***.

Supplementary Tables

Supplementary Table 1. Types of features used

Feature	Description
Morphological Features*(MF)*	Identifies specific features of words. For example: is the word all caps? does it contain a number? dashes, slashes or other punctuation marks?
Prefixes/Suffixes *(PS)*	Identifies specific prefixes or suffixes of a given length (2, 3, 4 characters) that are common in chemical names.
Types of Characters*(ToC)*	Identifies specific types of characters that are more common in chemical names, such as Greek letters, roman numbers, etc.
Length *(L)*	Classifies tokens by length. If the length is less than 5, the token is Short. If length is between 5 and 15, the token is Medium, otherwise, the token is Large.
Word class *(W)*	Analyzes the structure of chemical names in terms of frequency of upper and lower case characters, digits and other types of characters.
Brief Word class *(WB)*	Same as W, collapsing consecutive identical types of character into one Examples of these two features: *i.e. 1-methyl: 0.aaaaaa and.1-methyl: 0.a*.
List *(BNS/SW)*	Matches token to basic name segments and to the stop words list.

Supplementary Table 2. Chemical entity types

Chemical types	***TrainC***	***EvalC***	***MedlineC***
IUPAC & PARTIUPAC	4033	483	151
MODIFIER	1039	104	14
FAMILY	0	99	0
ABB	0	161	0
SUM	0	49	0
TRIVIAL	0	414	0
Total	**5072**	**1310**	**165**

Supplementary Table 3. Modeling assumptions for the CRF model.

Training configurations	Tokenization type	Order CRF	OC
1	A	1	0
2	A	1	1
3	A	2	0
4	A	2	1
5	B	1	0
6	B	1	1
7	B	2	0
8	B	2	1

Supplementary Table 4. Summary of F-score values for each tool in each experiment

	Exp. 1	**Exp. 2**	**Exp. 3**	**Exp. 4**
CheNER	86.98	70.06	50.03	69.94
OSCAR4	40.24	4.53	59.68	4.52
ChemSpot	73.35	10.51	74.24	10.47

Exp1: identify IUPAC entities in EvalC; Exp2: identify IUPAC entities in MedlineC; Exp.3 identify all type of entities in EvalC (complete corpus); Exp.4: identify all type of entities in MedlineC (complete corpus).

Supplementary Table 5. Feature sets leading to best and worst performance in the leave-n-out set of experiments ($1 \leq n \leq 7$). Bold and italicized features correspond to both corpora simultaneously. Bold features correspond to the ***EvalC*** corpus alone. Italicized features correspond to the ***MedlineC*** corpus alone.

FR	Minimum F-score FS	Maximum F-score FS
Leave-1-out	***MF, ToC, W, WB, L, BNS, SW***	***MF, PS, WB, L, BNS, SW,*****W**, *ToC*
Leave-2-out	***MF, ToC, W, L, SW,* WB**, *BNS*	***MF, PS, W, WB, BNS, SW***
Leave-3-out	***ToC, L, BNS, SW,* WB**, *W*	***MF, PS, BNS,* W, SW,** *ToC, WB*
Leave-4-out	***ToC, SW,* W, WB**, *L, BNS*	***MF, PS, BNS,* W,** *WB*
Leave-5-out	***ToC, BSN, SW***	***MF, PS,* L,** *SW*
Leave-6-out	***ToC, SW***	***MF, PS***
Leave-7-out	***ToC***	***PS***

FR: Feature Removal

4.5 References

1. Smith L (2008) Overview of BioCreative II gene mention recognition. Genome Biol 9: S2. doi:10.1186/gb-2008-9-s2-s2.
2. Li Q, Cheng T, Wang Y, Bryant SH (2010) PubChem as a public resource for drug discovery. Drug Discov Today 15: 1052–1057. doi:10.1016/j.drudis.2010.10.003.
3. Wishart DS, Knox C, Guo AC, Cheng D, Shrivastava S, et al. (2007) DrugBank: a knowledgebase for drugs, drug actions and drug targets. Nucleic Acids Res 36: D901–D906. doi:10.1093/nar/gkm958.
4. A D McNaught, Wilkinson A (1997) IUPAC Compendium of Chemical Terminology. Gold Book. Second.
5. Vazquez M, Krallinger M, Leitner F, Valencia A (2011) Text Mining for Drugs and Chemical Compounds: Methods, Tools and Applications. Mol Informatics 30: 506–519. doi:10.1002/minf.201100005.
6. Lafferty J, McCallum A, Pereira F (2001) Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Dep Pap CIS. Available: http://repository.upenn.edu/cis_papers/159.
7. Klinger R, Kolářik C, Fluck J, Hofmann-Apitius M, Friedrich CM (2008) Detection of IUPAC and IUPAC-like chemical names. Bioinformatics 24: i268–i276. doi:10.1093/bioinformatics/btn181.
8. Kolářik C, Klinger R, Friedrich CM, Hofmann-apitius M, Fluck J (2008) Chemical Names: Terminological Resources and Corpora Annotation. Available: http://130.203.133.150/viewdoc/summary?doi=10.1.1.140.4078. Accessed 10 October 2012.
9. Rocktäschel T, Weidlich M, Leser U (2012) ChemSpot: A Hybrid System for Chemical Named Entity Recognition. Bioinformatics. Available: http://bioinformatics.oxfordjournals.org/content/early/2012/04/11/bioinformatics.bts183. Accessed 10 October 2012.
10. Corbett P, Murray-Rust P (2006) High-Throughput Identification of Chemistry in Life Science Texts. In: R. Berthold M, Glen RC, Fischer I, editors. Computational Life Sciences II. Berlin, Heidelberg: Springer Berlin Heidelberg, Vol. 4216. pp. 107–118. Available: http://link.springer.com/chapter/10.1007/11875741_11?null. Accessed 10 October 2012.
11. Jessop D, Adams S, Willighagen E, Hawizy L, Murray-Rust P (2011) OSCAR4: a flexible architecture for chemical text-mining. J Cheminformatics 3: 41. doi:10.1186/1758-2946-3-41.
12. NCI Thesaurus - Chemical Modifier - Terms | NCBO BioPortal (n.d.). Available: http://bioportal.bioontology.org/ontologies/46317?p=terms&conceptid=Chemical_Modifier. Accessed 10 October 2012.
13. Abbreviations and Symbols for Chemical Names of Special Interest in Biological Chemistry (1967). Eur J Biochem 1: 259–266. doi:10.1111/j.1432-1033.1967.tb00070.x.

Chapter 5. CheNER II

Proceedings of the Fourth BioCreative Challenge Evaluation Workshop vol.2 2013, **66**:69
ISBN:978-84-933255-8-9

A tool for the identification of chemical entities (CheNER-BioC)

Anabel Usié, Joaquim Cruz, Jorge Comas, Francesc Solsona and Rui Alves

Abstract

The CHEMDNER task is a Named Entity Recognition (NER) challenge proposed by the most recent BioCreAtIvE (IV) challenge. This task aims at automatically identifying and labeling different types of chemical names in biomedical text. There are two subtask to so: (1) CDI subtask which is base on listing the chemical entities found in each document, and (2) CEM subtask which is based on give the precise location of each entity found related to each document, differentiating between title or abstract.

We approach this challenge by proposing a hybrid approach that combines linear Conditional Random Fields (CRF) together with regular expression taggers and dictionary usage, followed by a post-processing step to tag those chemical names in a corpus of Medline abstracts.

Our system performs with an F-score of 72.34% and 73.54% on the development and sample sets, respectively, for the CDI subtask. For the CEM subtask the performance increases to 73.07% and 73.76% on the development and sample sets, respectively.

5.1 Introduction

The BioCreAtIvE (Critical Assessment of Information Extraction systems in Biology) challenge consists of a community-wide effort for the evaluation of how accurate and effective text mining (TM), information retrieval (IR) and information extraction (IE) systems are when they are applied to the biological domain.

BioCreAtIvE challenges became relevant owing to the growing importance of, and interest in, TM, IR and IE of biological texts. This importance results from the fact that biomedical literature accumulates at an ever increasing rate, as do other types of biological datasets, such as those derived from omics experiments. That accumulation makes the existent of reliable methods to automatically annotate and extract information from those datasets fundamental for researchers that want to extract as much usable information as they can from them [1].

Early interest in the area lead to the use of TM and IE techniques to implement systems that automatically identify genes, proteins, their interactions (PPIs), and functions within a text. However, there were no common standards or shared evaluation criteria to comparatively evaluate the various proposed approaches for doing that identification. Thus, developing common evaluation criteria and golden standard test data sets was crucial to evaluate the performance of those approaches [2,3].

The development of such criteria and data sets was informed by the methods used by the natural language processing (NLP) community in their Message Understanding Conferences (MUCs) [4] and Text Retrieval Conferences (TREC) [5]. While the development of evaluation criteria is fairly straightforward through the use of a variety of statistical approaches, the creation of golden standard test data sets is more difficult. This is due to the fact that such data sets have to be sufficiently large for their analysis to have statistical significance. Given that those data sets need to be assembled, examined, annotated, and curated by individual experts, a lot of time is required to create a truly useful golden standard data set.

The BioCreAtIvE challenge was set out to provide both evaluation methods and golden standard test sets that permitted the development of new applications, and the improvement of preexisting ones, for biological text mining.

The latest BioCreAtIvE challenge, BioCreAtIvE IV, addressed two main issues. The first regards the Named Entity Recognition (NER) of biological entities and concepts. NER uses different methods to identify biological significant entities in the literature and associate them to existing database entries, in a process also known as Named Entity Normalization (NEN). The second regards the identification of entity-fact associations. Examples of this are the automated annotation of protein-protein interactions or association of biological function to protein/gene mentions. This is done using different methods for the extraction of semantic concepts used by NLP in other areas.

BioCreAtIvE IV was built on the success of the previous BioCreAtIvE challenges [6–10]. It defined the following five tracks as priority areas of biological text mining, where progress should be fomented and evaluated:

1. **Interoperability (BioC):** Development of an interoperable BioNLP module that can be seamlessly coupled to all BioC compliant modules.
2. **Chemical and Drug Named Entity Recognition (CHEMDNER):** Detection of mentions of chemical compounds and drugs, in particular those chemical entity mentions that can subsequently be linked to a chemical structure.
3. **Comparative Toxicogenomics Database (CTD) Curation:** Provision of Web Services to identify gene, chemical, disease, and action term mentions supporting CTD curation in PubMed abstracts.
4. **Gene Ontology (GO) curation:** Development of automatic methods to aid GO curators in identifying articles with curatable GO information (triage) and extracting gene function terms and the associated evidence sentences in full-length articles.
5. **Interactive Curation (IAT):** Demonstration and evaluation of web-based systems addressing user-defined tasks, evaluated by curators on performance and usability.

Given that we had developed CheNER to identify IUPAC chemical names mentioned in biomedical documents, it was considered that track 2 of BioCreAtIvE IV presented a good opportunity to compare the performance of our software with that of other available tools. This led us to modify and improve CheNER in order to make it more general and participate in the CHEMDNER track. Hereafter, we refer to the version of CheNER that participated in BioCreAtIvE IV as CheNER-BioC.

5.2 Description of the CHEMDNER Track

The CHEMDNER track promotes the development and implementation of systems that identify mentions of chemical names and drugs. This identification is crucial to aid in subsequence text-processing strategies such as the identification of drug-protein interactions, the extraction of metabolic and pathways reaction relations, among others.

The CHEMDNER track is divided into two tasks:

1. Chemical Document Indexing (CDI) Task: In this task, given a set of documents, the evaluated software must identify the list of chemical entities mentioned in each document.
2. Chemical Entity Mention recognition (CEM) Task: In this task, given a set of documents, the evaluated software must identify the precise location (also differentiating between title and abstract) of chemical entities within each of the documents.

Several corpora were provided by the organizers as golden standard test sets to evaluate the performance of the tools in each subtask: a **sample corpus** (25 annotated Pubmed abstracts), a **training corpus** and a **development corpus** (3500 annotated Pubmed abstracts each), and a **test corpus** (20000 non annotated Pubmed abstracts).

The format of the three annotated corpora is described in Figure 1. The different types of chemical entities annotated in the corpora included

SYSTEMATIC names (IUPAC and IUPAC-like), common or TRIVIAL names, trade names, chemical IDENTIFIERS (from databases and companies), acronyms and ABBREVIATIONS, reference numbers, chemical structures (SMILES, InChI), FAMILY names, and FORMULAS. Names that are equal in various nomenclatures are also tagged as MULTIPLE, and names that are chemicals from unidentified nomenclatures are tagged as NO CLASS. Supplementary Materials section provides more information about the annotation of the corpora and about which entities are included in each classification.

A: CEM predictions

23964783	A:603:630	1	0.5
23964783	A:632:635	2	0.5
23964783	A:708:711	3	0.5

B: CDI predictions

23964783	TRP	1	0.5
23964783	calcium carbonate	2	0.5
23964783	3,3-dimethylpentane	3	0.5

Figure 1. Example output predictions for the CEM (A) and CDI (B) sub-task. Te first column corresponds to the article identifier (PMID). The second column corresponds .to the predictions, i.e., the mention offset set in case of the CEM subtask and the unique mention string in case of the CDI subtask. The third column corresponds to the actual rank for each prediction given an article and the last column to the corresponding confidence score

5.3 Challenges in the automatic identification of chemical names

The development of systems that automatically identify chemical entities in biomedical texts is challenging due to both, the diverse morphology of chemical entities and the various types of nomenclature that are used to describe them in those texts [11]. These factors make it difficult to develop a single approach that can successfully identify all types of chemical mentions with high accuracy.

For example, standard IUPAC nomenclature and the nomenclature based on brand or trivial names have significant morphology differences. Thus, the former nomenclature has a complex morphology and a set of rules that makes the number of possible chemical names virtually infinite, making it impractical to use a dictionary and requiring the use of more sophisticated techniques to identify those names. In contrast, the later nomenclature is very much finite and its mentions can be identified using a dictionary approach.

5.4 Proposed method to address the challenges in the automatic identification of chemical names

Here we present and test a set of hybrid approaches that combine dictionary matching, linear CRFs (conditional random fields) and regular expressions to tag chemical entities in the biomedical literature. The first approach uses CheNER [12], a tool which implements Conditional Random Fields [13] based on Mallet [14]. CheNER achieves good results in identifying IUAPC names on our previous work with the SCAI corpora [15,16]. However, the goal here is to identify all types of chemical names. Because of that we implemented a slightly different approach to that used in the original CheNER. First, new specific features were added to the CRF before training it to identify specific types of chemical names. Second, dictionary matching and a varied set of regular expression rules were also combined with the CRF and used to identify chemical names.

5.5 System Description

The systems we present are inspired by our previous work in developing CheNER a tool for the identification of IUPAC chemical names.

Linear 2nd order CRFs, with offset conjunction value of 1 and tokenization by spaces, are individually or collectively trained to identify chemical names of types SYSTEMATIC, TRIVIAL, FAMILY, FORMULA, ABBREVIATIONS and IDENTIFIERS. The training corpus used was the CHEMDNER training set. The features used in the training are shown in Table 1. In addition, a dictionary is used to assist in identifying TRIVIAL, FAMILY and ABBREVIATIONS name types. Finally, regular expressions are employed for the recognition of FORMULA and IDENTIFIERS name types.

Table 1. Some of the features used.

Name of feature	Description
Morphological features	Identifies specific features such as contains dashes?, is all cap?, has a greek letter?
Word class	Automatic generation of features in terms of frequency of upper and lower case characters, digits and other types of characters.
Autom. Prefixes/Suffixes	Automatic generation of suffix and prefix (length 2, 3 and 4)
List	Automatic generation for every token that match an element within the list. This list can be a list of basic names segments, a list of stop words, etc.

The output of the various methods is based on the IOB labeling scheme, which is then reformatted to the required specifications of the CDI and/or CEM output format.

Integrating the output of the various recognition approaches (CRF, dictionary matching, and regular expressions), requires a post-processing step to be implemented in CheNER-BioC. In this step we perform several clean up actions, such as correcting unequal numbers of closing or opening brackets or detagging "action words" that are often appended at the end of chemical mentions such as "-based", "-regulated", etc.

5.6 Results & Discussion

The initial performance of the proposed approaches was tested using the CHEMDNER sample and test sets. This allowed us to test the performance of the various combinations of CRFs/regular expression/dictionary matching. The various tests are described in Table 2. The performance of our systems on the sample and development sets for the CDI subtask are displayed in Table 3, while the results for the CEM subtask are shown in Table 4. The performances of the various systems are similar in the sample and in the test datasets, suggesting that those performances are likely to be close to the limits of the method.

Table 2. Runs description.

Run	Description
1	Combines a CRF for SYSTEMATIC with an individual Regular Expression tagger for TRIVIAL, FAMILY, ABBREVIATION, FORMULA and IDENTIFIER.
2	Combines an individual CRF for SYSTEMATIC and TRIVIAL with an individual Regular Expression tagger for FAMILY, ABBREVIATION, FORMULA and IDENTIFIER.
3	Combines an individual CRF for SYSTEMATIC, TRIVIAL, FAMILY, ABBREVIATION, FORMULA and IDENTIFIER.
4	Combines an individual CRF for SYSTEMATIC, TRIVIAL, FAMILY, ABBREVIATION, and FORMULA with an individual Regular Expression tagger for IDENTIFIER.
5	Combines an individual CRF with specific labels for SYSTEMATIC, TRIVIAL, FAMILY, ABBREVIATION, FORMULA and IDENTIFIER.

Table 3. CDI subtask results based on the Micro-average results. Also is shown the execution time. P:precision, R:recall, F:f-score, AP: average precision, Fs: FAP-s and E: execution time.

	CDI subtask									
	Sample set					*Development set*				
	Run 1	**Run 2**	**Run 3**	**Run 4**	**Run5**	**Run 1**	**Run 2**	**Run 3**	**Run 4**	**Run5**
P	79.26	79.86	83.94	84.27	77.19	76.31	78.48	82.39	82.62	75.94
R	63.03	63.30	55.58	59.40	70.21	65.36	66.64	54.53	61.11	69.07
F	70.22	70.62	66.88	69.98	73.54	70.41	72.08	65.62	70.26	72.34
AP	50.29	51.09	46.44	49.75	50.23	50.27	54.43	45.51	50.75	51.97
Fs	58.61	59.29	54.82	58.16	59.69	58.66	62.02	53.75	58.93	60.49
E	149s	419s	939s	811s	2214s	15451s	46064s	118094s	88611s	188270s

Table 4. CEM subtask results based on the Micro-average results. Also is shown the execution time. P:precision, R:recall, F:f-score, AP: average precision, Fs: FAP-s and E: execution time.

	CDI subtask									
	Sample set					*Development set*				
	Run 1	**Run 2**	**Run 3**	**Run 4**	**Run5**	**Run 1**	**Run 2**	**Run 3**	**Run 4**	**Run5**
P	81.12	81.68	85.35	85.29	81.29	77.58	80.49	85.17	85.15	81.49
R	67.50	66.07	51.91	61.07	67.50	65.71	66.13	48.72	59.45	66.23
F	73.68	73.05	61.46	71.12	73.76	71.15	72.61	61.98	70.02	73.07
AP	54.37	52.49	43.72	51.09	51.12	49.79	50.35	40.13	49.23	51.82
Fs	62.57	61.09	51.09	59.48	60.38	58.58	59.47	48.71	57.85	60.64
E	149s	419s	939s	811s	2214s	15451s	46064s	118094s	88611s	188270s

Specifically, and for both subtasks, the system with the best F-score performance uses a single CRF that simultaneously identifies each type of entity. In contrast, the second best system combines various identification methods. On the one hand it uses two different CRFs that independently identify SYSTEMATIC and TRIVIAL name types. On the other, it employs regular expression/dictionary matching for FORMULA, IDENTIFIER, ABBREVIATIONS and FAMILY name types.

It is worth noting that our best performing system is fourteen times slower than our fastest system and five times slower than our second best system. In contrast, its performance is only improved by a couple of percent point with respect to its slower alternatives. Taken together, these results could be used to argue that in some cases using a system that is not as accurate in tagging chemical names in biomedical texts might be worth it because of its speed.

5.7 Supplementary Materials

These materials contain the transcription of the annotation guideline provided by the BioCreAtIvE organizers for the CHEMDNER Track. The organizers have authorize this transcription:

- 1: Description of the CHEMDNER annotation guideline.
- 2: CHEMDNER data selection.
- 3: CHEMDNER chemical entities.
- 4: CHEMDNER entity mentions type description.
- 5: Ortograph/Grammar Rules.
- 6: Multiwords: single entities vs multiple entities.

5.7.1 Description of the CHEMDNER annotation guideline[1]

This document describes the data selection criteria and annotation guidelines used for the construction of the CHEMDNER task corpora. The annotation guidelines will be refined after iterative cycles of annotations of sample documents based on direct suggestions made by the curators as well as through the observation of inconsistencies detected when comparing the results provided by different curators. Some participating teams provided feedback to improve the documentation after the release of the first sample set prepared for the CHEMDNER task. These informal rounds of curation served to improve the guidelines in the sense of making them more intuitive and easy to follow for the annotators.

The manual annotation task basically consists in labeling or marking up manually the mention of chemical entities in text following a set of rules specified below. The text to be labeled consists mainly in PubMed abstracts (title and abstract text) in the first round of annotation followed by the annotation of a smaller set of full text scientific articles and patent abstracts.

[1] **Overview of the chemical compound and drug name recognition (CHEMDNER) task.** Krallinger et al. *Proceedings of the Fourth BioCreative Challenge Evaluation Workshop vol. 2 ,2-33* http://www.biocreative.org/tasks/biocreative-iv/chemdner/

When possible, the selected chemical entity mentions were classified into one of seven chemical entity mention (CEM) classes defined in more detail below. The color code corresponds to the color tags provided by the MyMiner and AnnotateIt annotation interfaces for each of the CEM classes, to make the manual labeling and visualization easier.

Supplementary Table 1. Chemical Entity Mention (CEM) classes defined for the CHEMDNER task. For each CEM a short description and illustrative example cases are provided.

Type of chemical	Description	Examples
SYSTEMATIC	Description Systematic names of chemical mentions, e.g. IUPAC and IUPAC-like names	2-Acetoxybenzoic acid 2-Acetoxybenzenecarboxylic acid 2-Acetoxybenzoic acid N-(4-hydroxyphenyl)acetamide 3,5,4'-trihydroxy-trans-stilbene
IDENTIFIERS	Database identifiers of chemicals: CAS numbers, PubChem identifiers, registry numbers and ChEBI and CHEMBL ids	CAS Registry Number: 501-36-0445154 CID 445154 CHEBI:28262 CHEMBL504
FORMULA	Mentions of molecular formula, SMILES, InChI, InChIKey	CC(=O)Oc1ccccc1C(=O)O InChI=1S/C9H8O4/c1-6(10)13-8-5-3-2-47(8)9(11)12/h2-5H,1H3,(H,11,12) C9H8O4 (CH3)2SO LUKBXSAWLPMMSZ-OWOJBTEDSA-N
TRIVIAL	Trivial, trade (brand), common or generic names of compounds. It includes International Nonproprietary Name (INN) as well as British Approved Name (BAN) and United States Adopted Name (USAN)	Aspirin Acylpyrin Paracetamol Acetaminophen Tylenol Panadol resveratrol
ABBREVIATION	Mentions of abbreviations and acronyms of chemicals compounds and drugs	DMSO GABA
FAMILY	Chemical families that can be associated to some chemical structure are also included. It involves: i-FAMILY- SYSTEMATIC: IUPAC (plurals) ii-FAMILY-FORMULA iii-FAMILY- TRIVIAL iv.-FAMILYABBREVIATION v-FAMILY - FAMILY (this fine grained subannotation will only be done initially for a subset of the data collection).	Iodopyridazines (FAMILY- SYSTEMATIC) diphenols (FAMILY- SYSTEMATIC) quinolines (FAMILY- SYSTEMATIC) terpenoids (FAMILY- TRIVIAL) ROH (FAMILY- FORMULA)
MULTIPLE	Mentions that do correspond to chemicals that are not described in a continuous string of characters. This is often the case of mentions of multiple chemicals joined by coordinated clauses.	thieno2,3-d and thieno3,2-d fused oxazin-4-ones

5.7.2. CHEMDNER data selection

One critical aspect when preparing annotated corpora is that the used documents should ideally be representative of the domain of interest. In order to select appropriate documents that mention different types of compounds, containing heterogeneous mention types, the organizers tried to detect in the first place what journals do actually often contain descriptions of chemical substances. Therefore existing substance annotations provided by the PubMed database were exploited.

Step 1 Selection based on subject categories from the ISI Web of Knowledge (The top 100 journals for each category were selected based on the journals impact factor):

- BIOCHEMISTRY & MOLECULAR BIOLOGY
- CHEMISTRY, APPLIED
- CHEMISTRY, MEDICINAL
- CHEMISTRY, MULTIDISCIPLINARY
- CHEMISTRY, ORGANIC
- CHEMISTRY, PHYSICAL
- ENDOCRINOLOGY & METABOLISM
- ENGINEERING, CHEMICAL
- POLYMER SCIENCE
- PHARMACOLOGY & PHARMACY TOXICOLOGY

Step 2 Then those journals were taken that had at least 100 articles

Step 3 From these journals, articles were selected that have been published in 2013 in English, with abstracts and links to full text articles (based on PubMed query).

Step 4 Sampling of the articles was carried out depending on the associated categories:

Supplementary Table 2. Number of articles used for each journal category.

Category	#Articles
BIOCHEMISTRY	1000
CHEMISTRY_APPLIED	1000
CHEMISTRY_MEDICINAL	2000
CHEMISTRY_MULTIDISCIPLINARY	1000
CHEMISTRY_ORGANIC	2000
CHEMISTRY_PHYSICAL	1000
ENDOCRINOLOGY	1000
ENGINEERING_CHEMICAL	4
PHARMACOLOGY	1000
POLYMER_SCIENCE	300
TOXYCOLOGY	2000

Step 5 A list of unique articles was generated, resulting in 10991 articles.

Step 6 A random set of 10,000 abstracts from the joined collection was selected.

Step 7 The following random subsets were generated:

- TRAINING SET (3500 abstracts) + SCAI EVAL corpus
- DEVELOPMENT SET (3500 abstracts)
- TEST SET (3000 abstracts)
- Also the SCAI EVAL corpus (Kolaric et al 2008) was added to the training set.

Step 8 After the selection of articles the next step was to label each article exhaustively for CEMs. All the above CEM classes were tagged in the text according to the provided annotation rules.

5.7.3 CHEMDNER chemical entities

The focus for defining the chemical entities annotated for the CHEMDNER task was primarily to capture those types of mentions that are of practical relevance. Therefore the covered chemical entities had to represent those kinds of mentions that can be exploited for linking articles to chemical structure information. The annotation carried out for the CHEMDNER task was only exhaustive for the types of chemical mentions that are described in more detail below. This implies that other types of mentions of chemicals and substances

were not labeled. The common characteristic among all the chemical mention types used for the CHEMDNER task was that they could be associated to chemical structure information to at least a certain degree of reliability. This implied that very general chemical concepts (nonstructural or non-specific chemical nouns), adjectives, verbs and other terms (reactions, enzymes) that cannot be associated directly to a chemical structure are excluded from the annotation.

The annotation process itself also relied heavily on the domain background knowledge of the annotators when labeling the chemical entity mentions. A requirement to carry out the manual annotation was that annotators should have a background in chemistry, chemoinformatics or biochemistry to make sure the annotations are correct. This also made it possible to provide a short and compact set of annotation rules rather than requiring very detailed guidelines for non-experts. In this sense we followed a similar strategy as done for the gene mention tasks of previous BioCreative efforts (Smith et al. 2008). The definition of the chemical entity mention types used for the CHEMDNER task were inspired by the annotation rules used by Kolaric et al. (2008) and by Corbett et al. (2007).

Chemical Entity Mentions (CEMs) for this task had to refer to names of specific chemicals, specific classes of chemicals or fragments of specific chemicals. General chemical concepts, proteins, lipids and macromolecular biochemicals are excluded from the annotation. Therefore genes, proteins and protein-like molecules (> 15 amino acids) were excluded from the annotation. Chemical concepts were annotated only if they provided structural information (e.g. FAMILY type detailed below).

In order to label chemical entity mentions a set of rules have been defined that are described below. Example cases are provided to aid in understanding the different rules. The correct CEM cases are marked in yellow.

As first general annotation guidelines consider:

Rule 1: Use of external knowledge sources

In case the curator is not sure if a mention corresponds to a compound or he does not know what kind of compound mention it is, he may consult external knowledge resources: Wikipedia, Chemspider, Chemical Suppliers Catalogues (Sigma Aldrich, Tocris...), Scifinder, http://global.britannica.com/ such as the web or chemical databases to resolve doubts. A list of useful external knowledge sources should be compiled. Ideally some aid here from the annotation system should be expected.

Rule 2: Not unclear mentions

Do not tag unclear cases. If the annotator is not sure about a given mention, even after consulting some external sources, the corresponding mention should remain unlabelled.

Alkaloid *stands for compounds with a basic nitrogen, but the boundary is not clear enough and the substructural pattern neither. However, chemists typically recognize them...*

Glucocorticoid *structurally similar, but without a strict group definition*

The following annotation rules define which chemicals are CEM

Positive Rules - CEM are:

P1. Chemical Nouns convertible to:

- -A single chemical structure diagram: single atoms, ions, isotopes, pure elements and molecules:

  Fluorine, Iron, Deuterium, Benzene, Pyridine

- A general Markush diagram with R groups. Typically, chemical functionalities, fragments and structural classes → assignable to the CEM = FAMILY class.

  Amides, Hydroxipyridines, ROH, Aminoacids, Methyl Group, O-H group

P2. General class names where the definition of the class includes information on some structural information or elemental composition, independently of their origin (synthetic small compounds or natural products) →CEM = FAMILY class.

Hydrocarbons, organochlorines, carbohydrates, organometallics, Lewis Acids, Grignard Reactants, polyketides, steroids, macrolides, terpenoids, fatty acids, nucleotides, nucleobases, Bronsted-Lowry acid, transition metal, halogen, Schiff base, Wittig Salt, Wittig Reagent, monosaccharide, sugars, saturated fatty acids, trans fatty acids, triglyceride, ...

P3. Small Biochemicals

- -Sacharids: monosaccharides, disaccharides and trisaccharides should be tagged:

  Glucose (monosaccharide)
  Fructose (monosaccharide)
  Ribose (monosaccharide)
  Sucrose (disaccharide)
  Streptomycin (an aminoglycoside trisaccharide)
  Gentamicin (an aminoglycoside trisaccharide)
  cyclodextrin (cyclic oligosaccharides) *not tagged*

- -Peptides and proteins: peptides and peptidomimetics should be tagged. By convention, a threshold of 15 aminoacids was chosen as cut-off. Thus, peptides with less than 15 aminoacids should be tagged as CEM (both, cyclic and non-cyclic peptides).

  Glutathione *(trimer)*
  Cyclosporin A *11 aminoacids*
  Degarelix
  Gonadotropin-releasing hormone (GnRH) *with 10 aminoacids*
  Luteinizing-hormone-releasing hormone (LHRH) *same as for GnRH*
  Azaline B *small peptide with < 15 aminoacids*

Angiotensin *<10 aminoacids*

As well as chemical modifications on these peptides:

[D-Ser-(But),6, des-Gly-NH210]LHRH ethylamide

But, for example, luteinizing hormone is a protein (92 aminoacids), so it should not be tagged. In the same way, chemically modified proteins with > 15 aminoacids should not be tagged.

Luteinizing hormone (LH) *untagged because it has 92 aminoacids*

- Nucleotides: Mentions of monomers, dimmers, trimers should be tagged.

  NADH

  NAD+

  Nicotine adenine dinucleotide

  ATP

  Adenosine Triphosphate

  Adenosine 5'-Triphosphate

  SAM S-Adenosyl methionine

  cAMP

- Lipids: Fatty acids and their derivatives (including tri-, di-, monoglycerides), sterol derivatives...excluding polymeric structures.

  Glycerol

  Prostaglandin A

  Leukotriene A4

  Cholesterol

  Lipopolysaccharides

  Eicosanoide

P4. Synthetic Polymers

Nylon

Polystyrene

Polyvinyl chloride (PVC)

Polyamides

P5. Special Cases

- Minerals:

  Calcite

  Silica

  Alumina

  Titania

- Laboratory Reagents: common synthetic chemistry laboratory reagents, but only if their chemical composition is well defined.

  Petroleum ether

  Silica gel

  Universal indicator

  Molecular Sieves

  Litmus

- Dye and indicator names:

  methyl red

  Coomassie Brilliant blue

  DAPI

Negative Rules - CEM are not:

N1. Other terms different from chemical nouns: adjectives (if isolated/outside from chemical nouns - see M3 and M4 below), pronouns, verbs, other terms (reactions and enzymes), chemical prefixes (if isolated/outside from chemical nouns), anaphors, referring expressions, compound numbers...

- Chemical Reactions:

  Deshidrogentation

  methylation

  hydrolysis

- Pronouns, anaphors:

"DAPI is a dye...**this** compound..." *do not tag "this"*

- Compound numbers in anaphors: Even if the numbers are combined with other word (generating anaphors), they should never be annotated:

  ...of 8-amino-2,6-methano-3-benzazocine (2)... *do not tag "2"*

  (S)-4-AHCDP (6) and (R)-4-AHCP (7) *do not tag "6" and "7"*

  cis-9, ortho-12 *do not tag these entities*

- Chemical Prefixes (outside chemical names):

  1,4-derivatives *do not tag "1,4-"*

N2. Chemical nouns named for a role or similar, that is, nonstructural concepts:

- **Generalities:** analogue, substituent, inhibitor, hit, agonist, antagonist, activator, effector, antioxidant, substrate, inactivator, pigment, agent, standard, pharmacophore, drug, promoter, exon, intron, gen, antifolate, food, compound,...
- **Biological Roles:** hormone, antibiotics, antigen, herbicides, antifungals, toxin, metabolite, antineoplastic agents, antiestrogens,...
- **Reactivity Role:** electrophile, nucleophile, michael acceptor, dienophile, chelator, alkylating reagent, oxidizer, cation, anion, lipophile,...
- **Laboratory Role:** solvent, reagent, starting materials, building blocks, buffer, catalyst,...
- **Elementary Particles:** neutron, proton, electron, helion,...
- **Plants** (and APIs from plants without a defined chemical structure): estragon
- **Oils, essences and general formulations of several compounds:** estragon

N3. Very nonspecific structural concepts:

- **General structural concepts:** atom, ion, molecule, polymer, stereoisomer, enantiomer, isomer, conformer, mesomer, conformation,

monomer, dimmer, trimer, tetramer, lipid, gen, protein, alkali, functional groups, carrier proteins, aglycone, oligosaccharide, glycoside, saturated fat,...

The stereoisomer 6, but not 7, activated cloned carminomycinone-aglycone (II) of carminomicin *not tagged*

Refer to M2 for the special case of conflictive words: acid, salt, metal

- **Vague topological descriptors:** macrocycle, catenane, rotaxane,...

N4. Context Criteria: Words are not CEM if they are not CEM in context, even if they are co-incidentally the same set of characters (synonyms and metaphors):

Lead compounds are often found in high-throughput screenings ("hits") or are secondary metabolites from natural sources → *not tagged*

Mutations in ICE genes disrupting mating-body formation lead to 5-fold decreased ICE transfer rates. → *not tagged*

Lead is a chemical element in the carbon group with symbol Pb.
The man without self-reliance and an iron will is the plaything of chance → *not tagged*

What the new gold standard will look like → *not tagged*

N5. Biomolecules/Macromolecular biochemicals: not large oligomeric and polymeric or established DNA/RNA/protein sequences:

Do not tag proteins, polypeptides (> 15aa), nucleic acid polymers, polysaccharides, oligosaccharides and other biochemicals. Exclude all large biopolymers regardless of how their structures are organized. *Chemical:* if it is best represented using a chemical structure. Biochemical: if it is more usually represented using a sequence or a block diagram.

ubiquitin, insulin, DNA, mRNA, keratin, collagen, starch, cellulose, glycogen, agarose, chitin, murein, peptidoglycans, glycoproteins, lipopolysaccharide,

interferon, human fibroblast interferon, Kozak sequence *(example of an established sequence of aminoacids)*, annexin, atrial natriuretic peptide *(28 aminoacids)*, peptide,

N6. General vague compositions

Pigments with a relatively varying mixture: melanin

N7. Special words not to be labeled by convention

Organic

Inorganic

Water and its physical states (Steam, Ice...) as well as adjectives (aqueous)

Proton, helion (proton for either the fundamental particle or the H+)

Lead → as it is a very common word in many chemical texts, meaning the "main" candidate compound from a chemical series or the verb "guide". As the expected chance of meaning the chemical element "lead" is much lower, we agreed in not including this word.

Gold *Same as for lead*

Note: In opposition to "lead" → the word "*iron*" should be tagged as within chemical texts it is much more probable to find this word referring to the chemical element than to the "cleaning" activity.

5.7.4. CHEMDNER entity mentions type description

The following CEM types were annotated for the CHEMDNER corpus. The following general guidelines should be applied when annotating the different CEM types:

Rule 3 → Each chemical mention can only be marked as a single CEM type

Rule 4 → Priority rules of CEM of various types

In case a CEM is comprised of a combination of different types or mentions, e.g. systematic, trivial, abbreviation, etc, the curator should label the mention according to the raking provided for the CEM, CEM1... CEM7. For example, if it contains at least a part that follows IUPAC rules, it should be tagged as SYSTEMATIC (even if the rest of the mentions correspond to trivial names, formula or identifiers and the IUPAC string is relatively short).

Asp-Glu-NSP	*FORMULA: where NSP is an abbreviation in the text*
Testosterone	*TRIVIAL*
3H-Testosterone	*SYSTEMATIC (as 3H is IUPAC)*
Sildenafil	*TRIVIAL*
N-methyl sildenafil	*SYSTEMATIC (as N-methyl is IUPAC)*
[N(gamma)-(IGly)Dab(8)]degarelix	*N(gamma) is IUPAC so it is composed of IUPAC + formula + trivial → results in SYSTEMATIC*
[(2-pyridyl)-methyl)d-Dap(3)]degarelix	*IUPAC + Formula + Trivial → results in SYSTEMATIC*
[IOrn(8)]degarelix	*composed of Formula + Trivial → results in FORMULA*
[Pra(7)]degarelix	*composed of Formula + Trivial → results in FORMULA*

CEM-1 (SYSTEMATIC): includes multi word systematic, CAS-style names and semi-systematic names such as mentions of chemical compounds following the IUPAC nomenclature guidelines (http://www.iupac.org/fileadmin/user_upload/publications/recommendations/Complet_eDraft.pdf). Also IUPAC-like mentions are included, as often the authors do not follow strictly the guidelines and sometimes authors combine chemical mentions that have both systematic and non-systematic mention elements.

1,2-dimethyl-3-hydroxypyridin-4-one

acetone semicarbazone

1-octanol

chloroacetyl chloride

iron

sodium

iron(III)

iron(3+)

acetylsalicylic acid

Polystyrene

Here we also include the mention of unique substances (not general family compounds) that are IUPAC or IUPAC-like next to non-essential parts of the chemical entity or name modifiers (see M1, M4 and M7):

2,3-Dihydrobenzofuran analogues

CEM-2 (IDENTIFIERS): corresponds to the following database identifiers of chemicals (strictly these databases): CAS registry numbers, PubChem, ChEBI and CHEMBL database identifiers and also company codes. These identifiers should only be labeled if the context provides enough information that these mentions correspond to chemical identifiers.

Its CAS Number is 28718-90-3...

Company codes: PD-0332991, FE200486

CEM-3 (FORMULA): this class corresponds to mentions of chemical formula, chemical line annotations, SMILES, InChI and 3-letter codes of nucleotides, amino acids and monossacharides:

$C_6H_{12}O_6$

EtOAc

Fe, Na, Fe(III), Li^+, Fe^{2+} *Atomic elements*

CC(=O)C *Chemical Line annotations*

D-Ala-D-Ala *3-letter codes of small peptides*

Glu-Cys-Gly *3-letter codes of small peptides*

GlcNAc *Oligosaccharides nomenclature: formula with abbreviation*

Asp-Glu-Fmoc *Formula (formula with abbreviation)*

InChI=1S/C22H15N/c1-3-8-16(9-4-1)21-19-13-7-12-18-14-15-20(23(18)19)22(21)17-10-5-2-6-11-17/h1-15H

t-BuOK

CEM-4 (TRIVIAL): this class included trivial and common names of compounds. It also includes trademark and commercial names of chemicals and drugs.

- **Drug Names**: aspirine, Viagra, Degarelix,...
- **Minerals:** calcite, silica, alumina, titania, zeolite,...
- **Metals (alloys):** bronze, steel,...
- **Allotropes:** Diamond, Graphite, monoclinic sulfur, ozone, ...
- **General names:** table salt, vinegar,...
- **Other common names (mainly for small biochemicals):** adenine, testosterone, mezerein, azalin B, mannitol, rosiglitazone, pyruvate kinase, xanthine oxidase, deferiprone,...

  Note that the name of the amino acids (serine, asparagine,...) is IUPAC, so they should be labeled as SYSTEMATIC.

CEM-5 (ABBREVIATION): this class covered the mentions of abbreviations and acronyms of chemical compounds and drugs. Only those abbreviations were annotated that could be clearly linked to chemical entities based on the annotators background knowledge or on descriptions provided in the article (ad-hoc abbreviations).
Annotate acronym, abbreviation and other definitions occurring before/after the chemical name separately:

H]-8-OH-DPAT [8-hydroxy-2-(N,N-di-n-propylamino)tetralin]
2,4-dinitrophenyl)sulfenyl (DNPS)
Gamma-aminobutyric acid (GABA)

Where:

[3H]-8-OH-DPAT	*Formula (formula + abbreviation)*
8-hydroxy-2-(N,N-di-n-propylamino)tetralin	*Systematic*
(2,4-dinitrophenyl)sulfenyl	*Systematic*
DNPS	*Abbreviation*

Gamma-aminobutyric acid *Systematic*

GABA *Abbreviation*

Include acronym and abbreviation definitions that occur inside chemical names:

H-Lys-Trp(NPS)-OMe *Formula (formula + abbreviation)*

CEM-6 (FAMILY): this mention type included well-defined chemical families that can be associated to some chemical structure. Pharmacological families were excluded from this class (refer to rule N2). This also included plural forms of systematic compound mentions that refer to a family of compounds. In this case name-internal cues can be a useful help. Initially the organizers planned to remove this class distributing the associated entities to other mention types. We finally decided to keep this as a separate CEM class because it involved chemical structural information and in some cases is of practical relevance.

In this particular case the mentions of type FAMILY involve the following subcategories as follows:

i. **CEM 6.1 FAMILY-SYSTEMATIC** CEM of type FAMILY that follows the systematic or semi-systematic nomenclature guidelines. Mainly plurals of IUPAC names

   Quinolines

   As well as the terms referring to general chemical groups (aldehyde, hydroxide, amino acid,...). In case of doubt, when the chemical entity may refer to either a single compound or a family of compounds (e.g. "urea"), the context should be considered to disambiguate.

ii. **CEM 6.2 FAMILY-FORMULA** CEM of type FAMILY that corresponds to a chemical formula (described in more detail in class FORMULA) If the formula encompasses > 1 compound:

   C-S-C bonds *Information on bonds/bridges (structural classes)*

   ROH

   CH stretching modes of DNP films

Note. Generic nomenclature is accepted within formulae only if the formula has more than 1 character:

MCl_2 where M is any metal

ROH stands for alcohols

M = Cu, Ag *M alone is not labeled*

R = amides, amines... *R alone is not labeled*

X = any halogen *X alone is not labeled*

iii. **CEM 6.3 FAMILY-TRIVIAL** CEM of type FAMILY that corresponds to a trivial name (described in more detail in class TRIVIAL structural class names)

Terpenoids

Sugars

Wittig Reagent

Lewis Acid

Synthetic polymers consisting of an undefined number of monomers (polyamide, polyvinylidene fluoride, PEG...) will be considered as FAMILY class members.

iv. **CEM 6.4 FAMILY-ABBREVIATION** CEM of type FAMILY that corresponds to an acronym or abbreviation (described in more detail in class ABBREVIATION)

v. **CEM 6.5 FAMILY-FAMILY** → other family names that do not match any of the other previous four classes. Are of the type family but one cannot clearly assign them to a more specific sub-class. For example, adjectives in M4:

Pyrazolic compounds

vi. **CEM-7 (MULTIPLE):** this class addressed mentions that did correspond to chemicals that are not described in a continuous string of characters. This is often the case of mentions of multiple chemicals joined by coordinated clauses or enumerations of chemical names (often used to avoid redundancies). Also parts of names divided by long text passages

fall into this class. The dependencies of the partial chemical compound mentions are not captured in this version of the task. Such MULTIPLE mentions could be decomposed later defining the dependencies, chaining rules or alternative allowed mentions in a second step if needed. They are only annotated if the corresponding joined mention (integrated form) would be one of the other chemical entity mentions defined for this task.

> 7-[3-(fluoromethyl)piperazinyl]- and -(fluorohomopiperazinyl)quinolone antibacterials
>
> thieno2,3-d and thieno3,2-d fused oxazin-4-ones
>
> 4-(3-chloro-4-hydroxyphenyl)- and 4-(4-chloro-3-hydroxyphenyl)-1,2,3,4tetrahydroisoquinolines
>
> phenylsulfenyl or acyclic sulfenyl substituted dipeptides
>
> Hydroxy- and amino-substituted piperidinecarboxylic acids

Note1: if there are terms inside the sentence that do not form part of the chemical name â†’ they should not be tagged. Therefore, the potentially multiple entity will be splitted:

> elaidic (t-C18:1 delta9) and palmitic acid *two different entities*
>
> N-Substituted and unsubstituted 4-chlorobenzeneand- and 4-nitrobenzenesulfonamides *unsubstituted adds no positive chemical information and it should not be tagged. Then, N-substituted is outside the MULTIPLE CEM.*

Note2: on how to deal with the context. In the case of specific, isolated CEMs that, when isolated correspond to a specific chemical entity but that in the context refer to a class of compounds â†’ this CEM should be assigned to its non-family general class. Example:

> In general the synthetic route involved the coupling of diethyl N-[2-fluoro-4-(prop-2ynylamino)benzoyl]-L-glutamate with the appropriate 6-(bromomethyl)quinazoline followed by deprotection with mild alkali.

6-(bromomethyl)quinazoline *should be tagged as FAMILY*

5.7.5 Ortography/Grammar Rules

O1 Other languages

Names in other languages than English should be annotated regardless the language according to the general annotation rules and CEM classes.

(9E)-9-Octadecensäure	*German*
9 trans - ácido octadecanoico	*Spanish*
9-octadecenoic acid, (9E)-Acide (9E)-9-octadécénoïque	*French*

O2 Mis-spellings & conversion errors

Mentions of chemicals (as long as they follow some of the other mention rules) that are misspelled should be tagged. This also includes mentions suffering from automatic conversion errors generated by text conversion programs.

ch1oro	*where 1 is "one"□ not "l"□*
1. 1 equiv. Br2in dioxane, ...	*where it should be "Br2 in dioxane"□*

O3 "A B" wrong space

White space-separated words that should properly be a single word â†’ should be marked up as single entity.

... the acetoxy ethyl group was ...

O4 Chemicals named after people

Mentions of chemicals named after people should be tagged if they do refer to very clear chemical structures. These mentions correspond generally to "trivial"□ or "family"□ names widely used.

Tröger's base	*Trivial*
Schiff base	*Family-Trivial*
Grignard reagents	*Family-Trivial*

But this only applies for chemical entities (not chemical reactions):

Gewald thiophene synthesis *only tag thiophene*

O5 Sentence boundary

Chemical entity mentions cannot span multiple sentences.

O6 Not short mentions

Do not tag acronyms that are of 1 or 2 letters in length. 1-letter code of aminoacids/nucleotides or biochemical mutation mentions should be excluded. 1letter code of chemical elements should be annotated (as FORMULA)

A T R Arg176Met

1154C>T (A385V) and 1193T>C (M398T) in the coding exons *untagged*

Pd/C *these are tagged because they are of CEM FORMULA*

N-terminal *N (nitrogen should be tagged as CEM FORMULA)*

O7 Not flanking white space characters

Not tag white space characters flanking the CEM. Annotators should try to define the mentions precisely, and not include flanking whitespace or other spacing characters.

O8 Not Commas, full stops, brackets

Do not include as part of the CEM: off commas, full stops, brackets, and references to papers etc. that aren't a part of the name itself. Do include as part of CEM the square brackets around inorganic complexes and ionic liquids only if the bracket appears within the name.

[Co(CN)53I]

but:

[Cu(H2O)6]$^{2+}$

Acetate, bromine, the new compounds (aspirin and (carboxyalkyl)hydroxypyridinone)

Deferiprone (1,2-dimethyl-3-hydroxypyridin-4-one)

O9 Include prefixes for stereochemistry

Include in the CEM label prefixes that denote stereochemistry or regiochemistry of the compound.

cis-methanoglutamate

cis-platin

(S)-Alanine

(3RS,4SR)-4-acetamidopiperidine-3-carboxylic acid

cis-isomer 22 *nothing tagged (no anaphors o general terms)*

O10 Not Trademarks

Do not include trademark symbols as part of CEM

Aspirin ®

Mesupron ®

O11 Not trailing hyphen/apostrophe

Do not tag trailing hyphens or the apostrophe-s in possessives. Exception: keep them in CAS names, keep them in case of FAMILY mentions.

Methyl-group

Kainite-preferring subunits GluR6 *(GluR6 is a protein receptor)*

Chloroform-induced ventricular tachycardia

Benzoic acid, 4-[[6-[[3'-(aminomethyl)[1,1'-biphenyl]-3-yl]oxy]-3,5-difluoro2-pyridinyl]oxy]-

Benzene's activity

Acyl-enzyme inhibitors

O12 Do not break up words to get at the CEM inside

Methylating *Not to be tagged (chemical reaction)*

Dienophile *Not to be tagged (reactivity role)*

Carbonium *To be tagged as ion (CEM), but not decomposed*

Acetyltransferase *Not to be tagged (enzyme)*

exo-**ATP**-site-directed reagents *ATP Not to be tagged inside the word*

m**Glu**R1alpha, m**Glu**R2 *Glu not to be tagged inside the receptors*

but:

ATP-site-directed inactivations

anti-dopamine beta-hydroxylase

non-N-methyl-d-aspartate(non-NMDA) glutamate (Glu)

O13 Numbers in formula and numbers as part of the name

Include numbers on the front of formulae that indicate stoichiometry.

C6H8O3.2H2O	FORMULA
$2H_2$ + O_2 -> $2H_2O$	FORMULA

Include numbers that specify positions of a molecule only if they are part of the name:

C-2 carbon	*only carbon is annotated*
C-2 and C-3 positions	*nothing is annotated*
N-1 position "standard" substitution	*nothing is annotated*
...possessing a [4-hydroxy-3-(hydroxymethyl)-1-butyl] substituent at **N-1** exhibited an activity...	
Ser473	*only Ser is annotated*
Thr-384	*only Thr is annotated*

If the positions identify general positions in compounds → these general positions should also be annotated

4-bromo derivative	*tag the 4- position*
5-vinyl substituent	
5-[2-(1-azirinyl)]uracil analogues	
5-vinyluracils	
5-vinyl substituent of the respective 5-vinyluracils	
2'-fluoro analogues	
N-methyl derivative	

5-[2-(1-azirinyl)]uracil analogues

with 5 -- 19 spacer atoms between N6 or C-8 and iodine have been evaluated *do not tag the N6 and the C-8 positions*

This rule on general positions applies for both numeric and string-defined (ortho, meta, para, o-...) positions in the molecule:

o-nitrophenyl-modified analogues

O14 State/charge/surface symbols

Include in the CEM oxidation state symbols, charge symbols, state symbols and surface symbols that occur on the end of names

Cu^{2+}

Cu(II)

CuSO4(aq)

Au(111) surface

(14)C *isotope*

5.7.6 Multiwords: single entities vs multiple entities

M1 The longest CEM should always be tagged, but only including those words that are actually part of the chemical name. Non-essential parts of the chemical entity and name modifiers should NOT be tagged:

sodium ion

hydroxyl radical

nitrogen gas

gold nanoparticules

methyl group

phenyl ring

caffeine analogue

carbon atom

cocaine addiction

Krebs citric acid cycle

Pyridine derivatives

Perovskite structure

but **substituted modifier should be tagged if inside a chemical entity (meaning R**):

N-**substituted**-2-alkyl-3-hydroxy-4(1H)-pyridinones

chloro-**substituted** phenyls

6-fluoro-7-**substituted**-1,4-dihydro-4-oxoquinoline-3-carboxylic acids

2,4-diamino-5-(2',5'-**substituted** benzyl)pyrimidines

N-methyl-**substituted** sulfonamides

but not if the word substituted (or similar words) do not provide specific information on the substitution (i.e., "isolated" words):

disubstituted naphtalenes

substituted 1,4-dihydronapthoquinones, hydroindoloquinones

amide alkyl substituents

14-substituted derivatives of carminomycinone

5-substituted acyclic pyrimidine nucleosides

N-Substituted and unsubstituted **4-chlorobenze- and 4-nitrobenzenesulfonamides**

M2 Conflictive words: CEM or Modifiers? "Acid" "Base" "Salt" "Metal"

Do only mark these words if they are part of a longer specific chemical name or if they refer to explicit classes of compounds (e.g. transition metal). Alone, these words should not be tagged (except for the case of the word "salt" meaning "sodium chloride").

Strong acid

Organic acid

lysergic acid

carboxylic acid

table salt *incluso de esta se podría hacer una exception como* water

organic salt

citric acid trisodium salt

transition metal

metal oxide

heavy metal

the sodium salt

in treatment with aqueous alkali or acid *do not tag alkali/acid*

M3 Adjectives with valid CEMs

Adjectives are only to be annotated if i) precede/follow a valid chemical entity and ii) add more precise structural information to this chemical entity. The whole concept (adjective + chemical noun) should be tagged as a unique chemical entity assignable to the chemical class of the chemical entity alone. This is independent on the origin of the root name of the adjective (i.e. systematic names or common names: pyrazolic vs nicotinic) and on the adjective ending ("-ed"□, "-ing"□,"-olic").

polychlorinated biphenyl

disubstituted naphtalenes

acetylated phenoles

dry ether

ethanolic KOH

allylic alcohol

colloidal silver

dry ice *which is CO2, not H2O*

fuming sulphuric acid *which is H2S2O7,not H2SO4*

warm HCl

aqueous sodium carbonate

molecular nitrogen

primary alcohols *specifies the precise type of alcohols*

secondary hydroxy groups *specifies the precise type of hydroxyl groups*

stainless steel

tertiary 2-(3-hydroxyphenyl)-2-phenethylamine

ionotropic glutamate receptors *do not tag "ionotropic"*

M4 Adjectives with general classes

Adjectives are only to be annotated if i) precede/follow a general compound class (compound(s), hit, analogue(s), derivative(s), series(s)...) and ii) add more precise structural information to this chemical entity (chemical class). Typically, these adjectives end as "oic", "-oid", "al", "-ois".

In contrast to M3, here only the adjective is tagged as a chemical noun of type FAMILY-FAMILY:

Pirazolic compounds	*Family-Family*
Terpenoids analogues	*Family-Family*

But not if they still result in very wide compound families (commonly, adjectives finished in -ed correspond add less specific (R-group related) information than the others (-oic adjectives):

Methoxylated analogues	*nothing tagged*
Fluorinated compounds	*nothing tagged*

But not when found in different contexts:

glycemic control	*nothing tagged*
noradrenergic areas	*nothing tagged*

M5 Negative adjectives

"Negative" concepts that discard specific chemical structures but that do not explicit define a chemical structure should not be tagged.

2-desamino, 2-desamino-2-hydroxymethyl, and 2-desamino-2-methoxy analogues *desamino meaning "replace the amino group by hydrogen"*

Similarly, the prefix non- should not be included:

non-steroidal	*tag only steroidal*
non-fluorinated parent compounds	*do not tag fluorinated as stated in M4.*

But it the term is to be tagged → then tag the corresponding adjective:

non fluorinated quinazolines	*tag the adjective*

non-fluorinated quinazolines *tag the adjective*

M6 Enumerations and list of compounds vs multiple entities:

If full names are enumerated, tag separately each individual CEM:

citric acid and acetic acid

lithium carbonate, sodium carbonate

hexane-ethyl acetate, pyrane, aspirin/ibuprofen

aspirin, sugar, 4-methoxy phenol, and R-OH

If chemicals or class names of compounds are not described in a continuous string of characters → tag the whole string (including words such as "and", "or" and commas) as a single entity of class type multiple. Avoid the generation of "half truth".

citric and acetic acid

lithium,sodium and potassium carbonate

pyrimidine derivatives and pyridine analogues *(as "pyridimide derivatives" is not a CM)*

M7 CEM Overlapping with Enzymes

Mentions of CEM that are part of mentions of enzymes should be tagged.

i. Two independent words where we only analyze the CEM:

   K+ ATPase

   Pyruvate kinase

   phosphatidylinositol 3-kinase

   metabotropic Glu receptors

ii. In the cases of hyphens we always split the words, and then they are independently analyzed:

   Pyruvate-kinase

   K+-ATPase

iii. Enzyme compound transformation "A B -ase", meaning "the -ase enzyme that catalyses the transformation of A to B", should be marked up as separate entities.

Squalene hopene cyclase (SHC) catalyzes the complex

Quinazoline antifolate thymidylate synthase inhibitors

M8 CEM Overlapping with other non-chemical entities

Tag the corresponding chemical entity. For example, chemical formulae that appear inside mathematical formulae or equations (gradient, concentration):

^{1}H NMR

$d[Na^{+}]/dt = x$

[caffeine]=10 mM

M9 CEM1 CEM2 → A single CEM or two CEMS?

If there are two continuous words of type CEM: "CEM1" and "CEM2", each of which would individually be of class CEM:

- if they denote a single entity → label as a unique single CEM
- if they denote different chemical entities → label as independent CEM's

Use of adenine nucleotide derivatives â†’ conceptually are a single entity (tagged as trivial)

NOTE!!! This criterion is not in agreement with rules defined by Corbett et al.2007, as we found that the strict classification of these rules (with interpretation) would be really time expensive and a potential source of discordance if no extra careful reading...

- **Generic terms that mirror IUPAC formation → a single entity**

  Alkyl acetates

  Isopropyl halides

- **Complexes and host-guest compounds defined by two continuous words →a single entity**

  $Cu2^{+} \cdot 2H20$

  Hydroxypropil-beta-cyclodextrin-itraconazole

- **Mixtures defined as "CEM1/CEM2" or "CEM1-CEM2" → separate entities**

  hexane-ethyl acetate *hexane =CEM1 and ethyl acetate = CEM2*

  Pd/C preparation *Pd = CEM1 and C = CEM2*

  isosteric benzene-thiophene replacement

- "**the CEM2 that is part of the CEM1" → a single entity**

  carbonyl carbon

  acetoxy methyl signal

  acetoxy methyl group

- **"the CEM1 that is a CEM2" or "the CEM2 that is an CEM1" → a single entity**

  S-propionylthiolactyl-D-Glu-L-Lys thioester → *Difficult to differentiate that "thioester" is already implicitly mentioned in the previous CEM; by default, from practical perspective, we will annotate as unique CEM.*

  terpenoid limonene → *We will separate them; since in this case "terpenoid" is an adjective and does NOT provide additional structural information to its corresponding name (as explained above). Therefore, in this case terpenoid will not be annotated*

  pyrimidine nucleosides → *tagged together as a single entity*

- **"the CEM2 that contains an CEM1 group/moiety" → single entity**

  Methyl ether

  Tripeptide thioester

- Terms ending in "glycoside" → single entity

  Limonoid glycosides

  Nominilic acid glycoside

Note: all these examples apply for the case of mentions next to each other. If the words are separated by other words, annotate them separately.

A complex of hydroxypropyl-bet-cyclodextrin and itraconazole

5.8 References

1. Altman RB, Bergman CM, Blake J, Blaschke C, Cohen A, et al. (2008) Text mining for biology - the way forward: opinions from leading scientists. Genome Biol 9: S7. doi:10.1186/gb-2008-9-s2-s7.
2. Blaschke C (2002) Information extraction in molecular biology. Brief Bioinform 3: 154–165. doi:10.1093/bib/3.2.154.
3. Hersh W (2005) Evaluation of biomedical text-mining systems: lessons learned from information retrieval. Brief Bioinform 6: 344–356.
4. Message Understanding Conferences: MUC (n.d.). Available: http://www-nlpir.nist.gov/related_projects/muc/.
5. Text Retrieval Conferences: TREC (n.d.). Available: http://trec.nist.gov/.
6. Hirschman L, Yeh A, Blaschke C, Valencia A (2005) Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC Bioinformatics 6: S1. doi:10.1186/1471-2105-6-S1-S1.
7. Krallinger M, Morgan A, Smith L, Leitner F, Tanabe L, et al. (2008) Evaluation of text-mining systems for biology: overview of the Second BioCreative community challenge. Genome Biol 9 Suppl 2: S1. doi:10.1186/gb-2008-9-s2-s1.
8. Leitner F, Mardis SA, Krallinger M, Cesareni G, Hirschman LA, et al. (2010) An Overview of BioCreative II.5. IEEE/ACM Trans Comput Biol Bioinform 7: 385–399. doi:10.1109/TCBB.2010.61.
9. Arighi C, Lu Z, Krallinger M, Cohen K, Wilbur W, et al. (2011) Overview of the BioCreative III Workshop. BMC Bioinformatics 12: S1. doi:10.1186/1471-2105-12-S8-S1.
10. Krallinger M, Leitner F, Rodriguez-Penagos C, Valencia A (2008) Overview of the protein-protein interaction annotation extraction task of BioCreative II. Genome Biol 9: S4. doi:10.1186/gb-2008-9-s2-s4.
11. Vazquez M, Krallinger M, Leitner F, Valencia A (2011) Text Mining for Drugs and Chemical Compounds: Methods, Tools and Applications. Mol Informatics 30: 506–519. doi:10.1002/minf.201100005.
12. CheNER: Chemical Named Entity Recognizer (under revision) (2013).
13. Lafferty J, McCallum A, Pereira F (2001) Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. Dep Pap CIS. Available: http://repository.upenn.edu/cis_papers/159.
14. McCallum A (2002) Mallet: A machine learning for language toolkit. Available: http://mallet.cs.umass.edu/about.phpmallet.cs.umass.edu. Accessed 10 October 2012.
15. Klinger R, Kolářik C, Fluck J, Hofmann-Apitius M, Friedrich CM (2008) Detection of IUPAC and IUPAC-like chemical names. Bioinformatics 24: i268–i276. doi:10.1093/bioinformatics/btn181.
16. Kolářik C, Klinger R, Friedrich CM, Hofmann-apitius M, Fluck J (2008) Chemical Names: Terminological Resources and Corpora Annotation. Available: http://130.203.133.150/viewdoc/summary?doi=10.1.1.140.4078. Accessed 10 October 2012.

Chapter 6. Protein-MetReS

Protein-MetReS

Anabel Usié, Hiren Karathia, Ivan Teixidó, Francesc Solsona and Rui Alves

Abstract

Protein-MetReS is a tool developed to unify structural analysis of individual proteins and docking techniques to predict how those proteins form complexes.

The structural analysis is achieved in two ways. First, the tool provides access to repositories of experimentally determined and computationally predicted protein structures. Second, in the absence of available structural information, the program permits using various types of modeling methods to predict that structure.

Both theoretical models and experimental structures can be used by the various docking tools integrated in Protein-MetReS for protein-protein docking.

Overall, Protein-MetReS provides a unique interface to analyze both theoretical models and experimental structures of individual proteins and protein-protein complexes.

6.1 Introduction

Knowing the three-dimensional (3D) structure of a protein is crucial to understand its biological function and investigate biologically relevant interactions in molecular detail. Experimentally determined structures of individual proteins and protein complexes are deposited in a centralized repository, the **Protein Data Bank** (**PDB**), where the information is freely available to the scientific community [1].

In this post genomics era, millions of new genes and proteins are identified and sequenced each year. In contrast, only hundreds to tens of thousands protein structures are experimentally determined in the same period. Only a very small fraction of these pertains to protein complexes. Because of this, predicting both protein structures and protein-protein interactions (PPIs) are important and challenging problems in bioinformatics and computational biology.

Most of the computational methods for protein structure prediction from sequence emerged from the CASP[1] (Critical Assessment of Structure Prediction) competition [2]. The primary goal of this competition, which is in its tenth round, is to help advance the methods for predicting protein 3D structure from its amino acid sequence. The next round of CASP will start in May 2014.

Similar to this, most of the computational methods for PPI prediction are evaluated in the CAPRI[2] (Critical Assessment of PRediction of Interactions) competition. The primary goal of this competition is to assess methods predicting PPIs, based on *in silico* protein-protein docking (PPD) of protein structures [3,4]. Since the beginning of the competition in 2001, there are typically between two and four prediction rounds per year.

6.1.1 Protein structure prediction

Currently, the most accurate method for **structure prediction** is **homology modeling** (**HM**), which is based on the use of an experimentally determined

[1] http://predictioncenter.org/
[2] http://www.ebi.ac.uk/msd-srv/capri/

structure from a sequence homologue as a template to build the model of the target sequence. The accuracy of HM is strongly conditioned by the sequence identity between the target sequence and the template and by the quality of the alignment between the two sequences. In general, good models can be obtained from templates with more than 75% sequence identity.

When the sequence identity between target and template is below 30%, homology modeling may not provide reasonable models. In some cases this is due to the difficulty in identifying appropriate templates and creating a high quality alignment between the target sequence and the template. If this is the case, the use of **fold-recognition modeling** (**FRM**) methods may improve the accuracy of the modeling predictions. FRM combines sequence comparisons with prediction of the fold for the target sequence. By combining the two, FRM identifies the best templates on which to base the prediction and improves the alignment of the target sequence to that of the template(s).

When no suitable template is identified through HM or FRM, ***ab initio* modeling** (**AIM**) approaches provide an alternative for predicting protein structures "from scratch". The most successful AIM approach splits the sequence whose structure one wants to predict into smaller subsequences of few amino acids. Then, for each subsequence, it finds homologues with experimentally determined structures, creating a set of alternative models for each of these subsequences. The models are then assembled from the N- to the C-terminal of the sequence, mimicking the protein synthesis and folding process. Physically impossible folding conformations are eliminated in this process. Fully assembled models are further optimized and ranked, based on energy calculations. This approach is slow, very intensive, and requires vast computational resources.

The single largest publicly accessible repository of protein structural models is the **SWISS-MODEL Repository** (**SMR**) [5]. In this database one finds all theoretical models that were built using its own model building server, SWISS-MODEL [6–8]. Unlike experimentally determined structures, structural models do not have a central repository where they are deposited. This is due to

various factors. One of these factors is the fact that there are tens of servers to create models of protein structures and unifying results coming from all of them is complicated. Another factor is the widely varying accuracy of structural models and modeling servers.

Unifying access to protein modeling data is a difficult and important goal of this research community. The **Protein Model Portal** (**PMP**) [9] was developed to address this issue. To our knowledge, it is the available tool that integrates the largest set of concurrent tools for building structural models using HM of proteins. It also integrates some repositories with protein structures and protein homology models.

6.1.2 Protein interaction prediction (Docking)

Protein–protein docking (**PPD**) is defined as "*the prediction of the structure of two proteins in a complex, given only the structure of the interacting proteins*". The prediction of **PPIs** plays a central role in biological and medical sciences, because the physiological and pathological effect of proteins is often mediated by, and effected through, these interactions.

There are two main PPD approaches: (1) methods based on maximizing the shape complementary between two proteins [10–12]and (2) methods based on simulating the actual docking process calculating the pairwise interaction energies and minimizing the energy of the complex formed in different ways [13]. There are also hybrid methods that combine both approaches in different ways.

Three key steps are decisive for the accuracy and effectiveness of all major PPD approaches: (1) an adequate representation of the structural system, (2) an appropriate and efficient search of conformational space of the individual proteins and their possible complexes, and (3) an adequate scoring function that can be used for the ranking of potential solutions. On one hand, the speed and effectiveness are the two critical elements of the search procedure. On the other hand, the scoring function should differentiate between as many potential solutions as possible to allow for an effective discrimination between native and

non-native docked conformations (see [14] for more detailed information). Due to the difference between the number of determined protein structures and sequences, a potentially important application of PPD is to predict PPIs between models of proteins structures. In principle, there is no difference between the methods applied to dock experimentally determined or computationally predicted structures. However, most docking algorithms rely on atom-level representations of the structures. Therefore, the higher resolution of experimentally determined protein structures leads to more accurate PPD predictions.

6.1.3 Objective

It should by now be obvious that a better integration of HM, FRM and AIM methods is still forthcoming, as is integrating these methods with those for PPD. Taking this into account, it was our goal to develop a meta-tool, Protein-MetReS, that integrates HM, FRM, AIM, and PPD approaches.

6.2 Implementation

Protein-MetReS is an application that allows users to perform structure analysis, prediction and docking of up to the full proteome of organisms with fully sequenced genomes. The motor and user interface of the application was implemented using JAVA-FX technology. The database containing the full proteomes of more than 1200 organism with completely sequenced genomes is described in more detail in Chapter 2. This database also contains information about previously determined protein structures. It was implemented using MySQL technology. The program is user-centric and organism centric. It requires a user authentication to start running. Such authentication allows the system to link the user's results to the appropriate organisms of interest and to the right user.

Protein-MetReS implements three main functionalities: (1) Identification of previously preexisting protein structures, (2) sequence-based structural modeling of proteins, and (3) protein docking.

To identify preexisting protein structures, either experimentally determined or predicted, Protein-MetReS integrates searches to the PDB and SMR repositories (see Table 1). To build theoretical models for the sequences of interest, Protein-MetReS allows the users to submit their sequences to various HM and FRM servers (Table 2). In parallel it also enables the users to submit their structures of interest to several PPD servers (Table 3), in order to predict possible modes of interaction between their proteins.

Table 1. Protein Structure/Model Repositories.

Repository	Type of data
PDB[3] [1]	Protein structure
SWISS-MODEL Repository[4] [5]	Protein model

Table 2. Structural modeling prediction tools.

Tool	Method	Execution
MODELLER[5] [15,16]	HM	Local
SWISS-MODEL[6] [6–8]	HM	Remote (via web)
Phyre2 [17] [7]	FRM	Remote (via web)

Table 3. Protein docking tools.

Tool	Method	Execution
Gramm-X[8] [18,19]	Simulation, also called *Ab initio*	Local
HEX[9] [20–22]	*Ab initio*	Local
RosettaDocking[10] [23,24]	*Ab initio*	Remote (via web)

A summary workflow chart of how the program works is shown in Figure 1.

[3] http://www.rcsb.org/pdb/home/home.do
[4] http://swissmodel.expasy.org/repository/
[5] http://salilab.org/modeller/
[6] http://swissmodel.expasy.org/workspace/index.php?func=modelling_simple1
[7] http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index
[8] http://vakser.bioinformatics.ku.edu/main/resources_gramm.php
[9] http://hex.loria.fr/
[10] http://rosie.rosettacommons.org/docking/submit

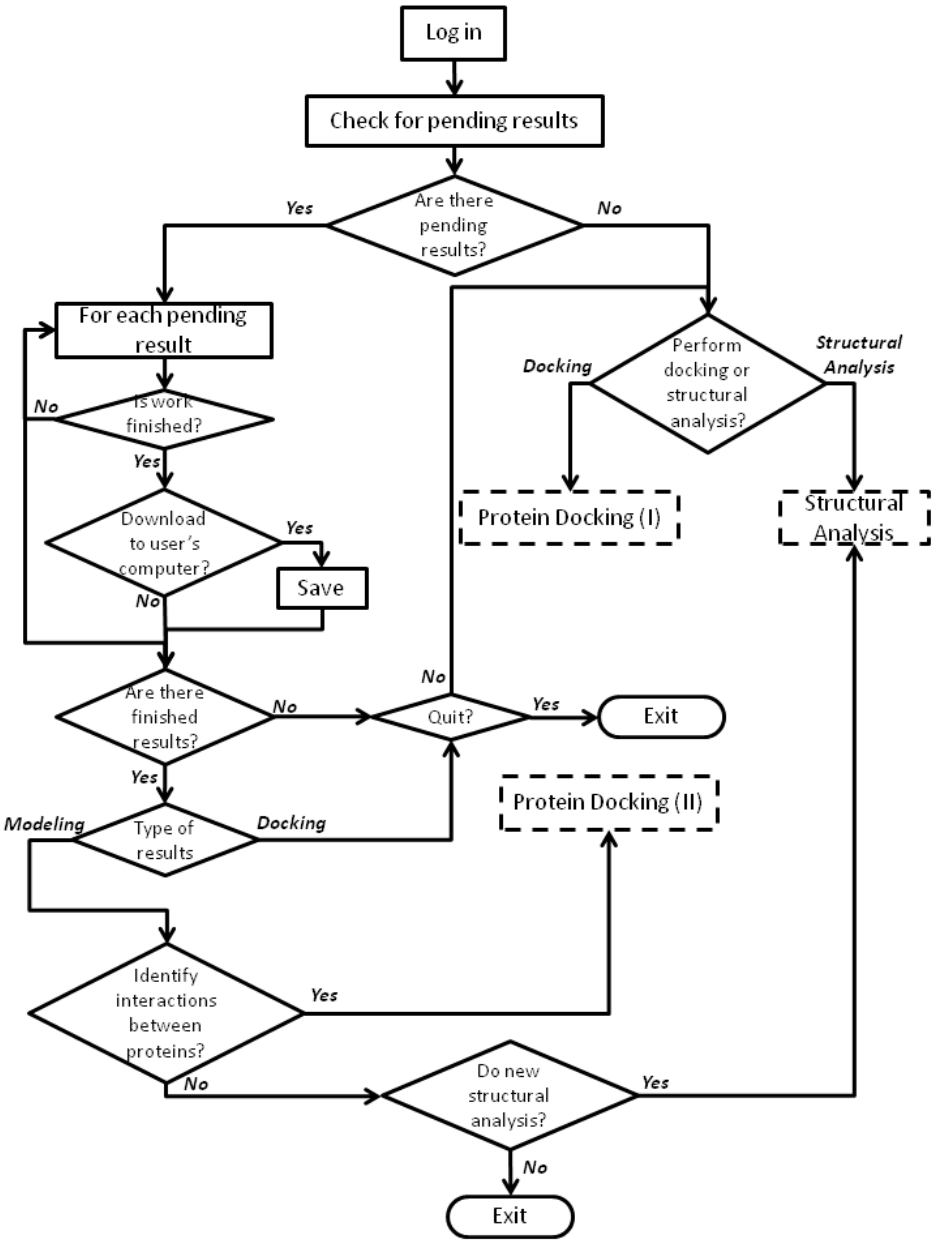


Figure 1. Workflow chart for Protein-MetReS. After the user logs in, the program verifies the completion state of any previous calculations done by that user. If no previous calculations that have yet to be visualized are available, users must chose to start either structure analysis/prediction or PPD. If new structural results are available, the user must decide what to do afterwards. Possibilities are either finding/predicting structures for new sequences or perform docking of the structures that are available. Users can also directly perform PPD by providing the program with structures for that docking. Both, structural results and docking results can be downloaded and saved by the users to their local computers.

6.3 Results

Protein-MetReS can be used either to obtain protein structures from a set of repositories or to build protein models. It can also be used for PPD in order to predict the most likely modes of interaction between proteins. Users can

download the application from http://metres.udl.cat and run it locally. An internet connection and a standalone version of JAVA (version 7 or later) are required. Upon starting the program, users log into the central Protein-MetReS database and a checkup of pending results is made. If the user has no pending calculations, s/he can start working by choosing the organism of interest. Once this is done, all proteins from this organism are loaded to the user interface. At this stage a choice of proteins of interest needs to be made. After this, the program allows the user to identify preexisting structures or create new structural models for those proteins. It also allows PPD of structures that are available to the user. If the users have previous calculations whose results have not been completed yet, they must wait for an e-mail announcing that completion before proceeding with the analysis of those results. If the users have previous calculations that have finished and whose results have not been visualized yet, they are provided with options about what actions can be taken. If these results are structural models, users can analyze the structures and perform PPD with Protein-MetReS. If these results pertain to PPD calculations, the users can analyze and download them. The flow of this process is shown in Figure 1.

6.3.1 Finding or predicting protein structures

When the user chooses to focus on protein structure analysis of a set of proteins, the natural flow of the program is as follows. First, the system shows the list of repositories and modeling tools that Protein-MetReS integrates (see Table 1 and Table 2). By default, all repositories and modeling tools are selected. The target sequences of the proteins of interest are used to check if existing structures or models are found in the selected repositories, and also in the local database. This last checkup is done because the system stores the models built by the users, to avoid re-building the same model.

The sequences of all proteins for which no preexisting structures or models are available are then submitted to the structural modeling servers selected by the users. The system will notify the user by e-mail upon conclusion of the

calculations. Note that modeling results are stored in the server's database. Figure 2 provides a more detailed workflow chart for the protein structure identification/modeling part of the application.

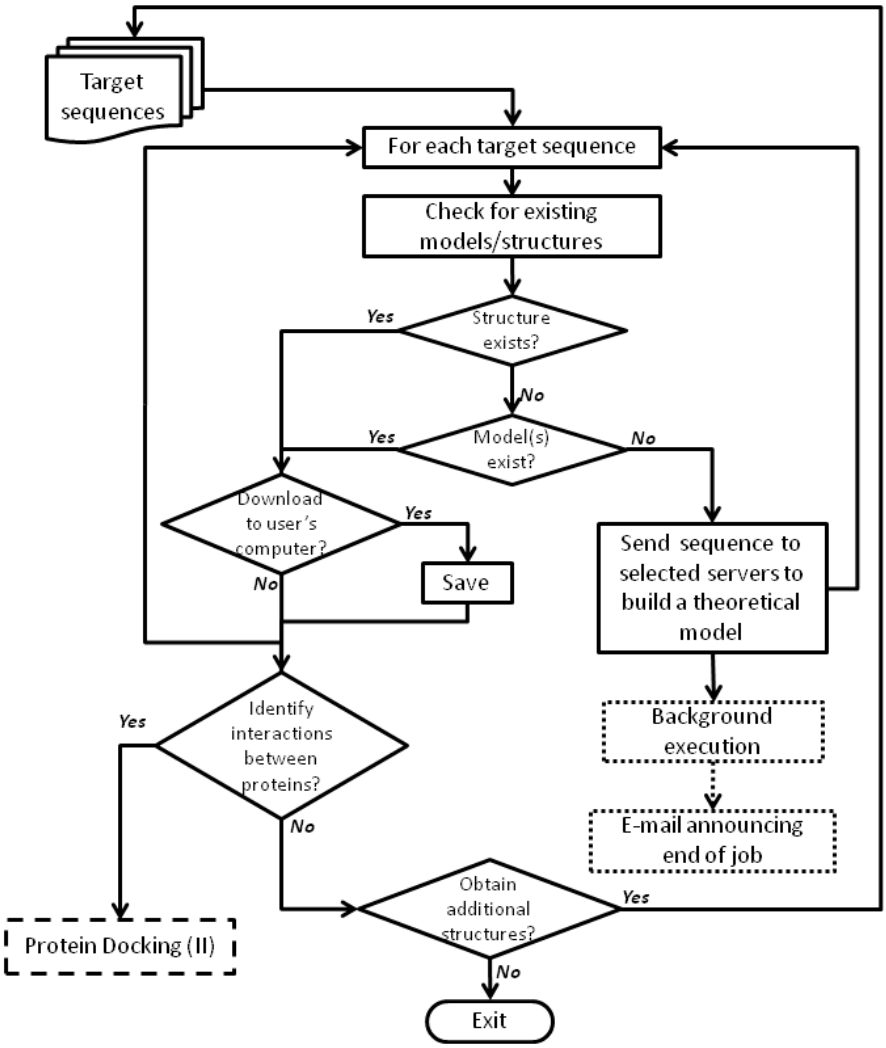


Figure 2. Workflow for the structural analysis in Protein-MetReS. In structural identification/prediction mode, the program uses homology modeling to check if the target sequence has an existing structure, if not checks if there is any model in the available repositories or in our database (where models created by Protein-MetReS are stored). If so, users can download the structures or models to their computers. If not, users can send their target sequences to the modeling servers. The subsequent modeling runs in the background, and the system will notice the user via email upon completion. Note that the user must select at least one repository and one server.

6.3.2 Visualizing structural analysis results

When models or experimental determinations of structures for the sequences that interest the users become available, the following information is displayed by default: the PDB accession number for the template or structure (including

the chain) and its resolution, the percentage of sequence identity, the e-value, the provider, the length of the amino acid sequence aligned, and the part of the sequence for which a structure is available.

In addition, a "model details" view is available, where the template-target sequence alignment and model building date are displayed, and an image of the structure is shown and linked to the interactive Jmol 3D molecular viewer [25]. If the model has not been updated for more than 3 months, a warning message is displayed, providing the user with the option of re-submitting the target sequence and re-building the model.

6.3.3 Predicting and analyzing protein- protein docking

To perform protein docking, users must have structure files available, whether they result from experimental determinations or from structural modeling. There are two ways to provide such files to the program. First, users can upload structure files directly located in their computers. Second, they can use the relevant structures available in the Protein-MetReS database, as the software provides a list of proteins of the organism of interest with available models or structures.

Once structures are provided or selected, the users can setup the pairs of proteins for which they are interested in performing PPD. By default, the system selects all possible pairs of proteins to dock. If more than one protein is selected, self-docking is not considered by default but can be selected by the user. Then, the system shows the list of docking tools that Protein-MetReS integrates (see Table 3). By default, all the tools are selected. Each tool is configured with default parameters. The users are allowed to change these parameters at will.

At this stage, the system checks the database to identify any previous docking studies, to avoid repeating calculations. All pairs of proteins whose docking has not been done before are then performed by the system, which will notify the user by e-mail upon conclusion of the calculations. Note that docking results

are stored in the server's database. Figure 3 provides a more detailed workflow chart for the PPD part of the application.

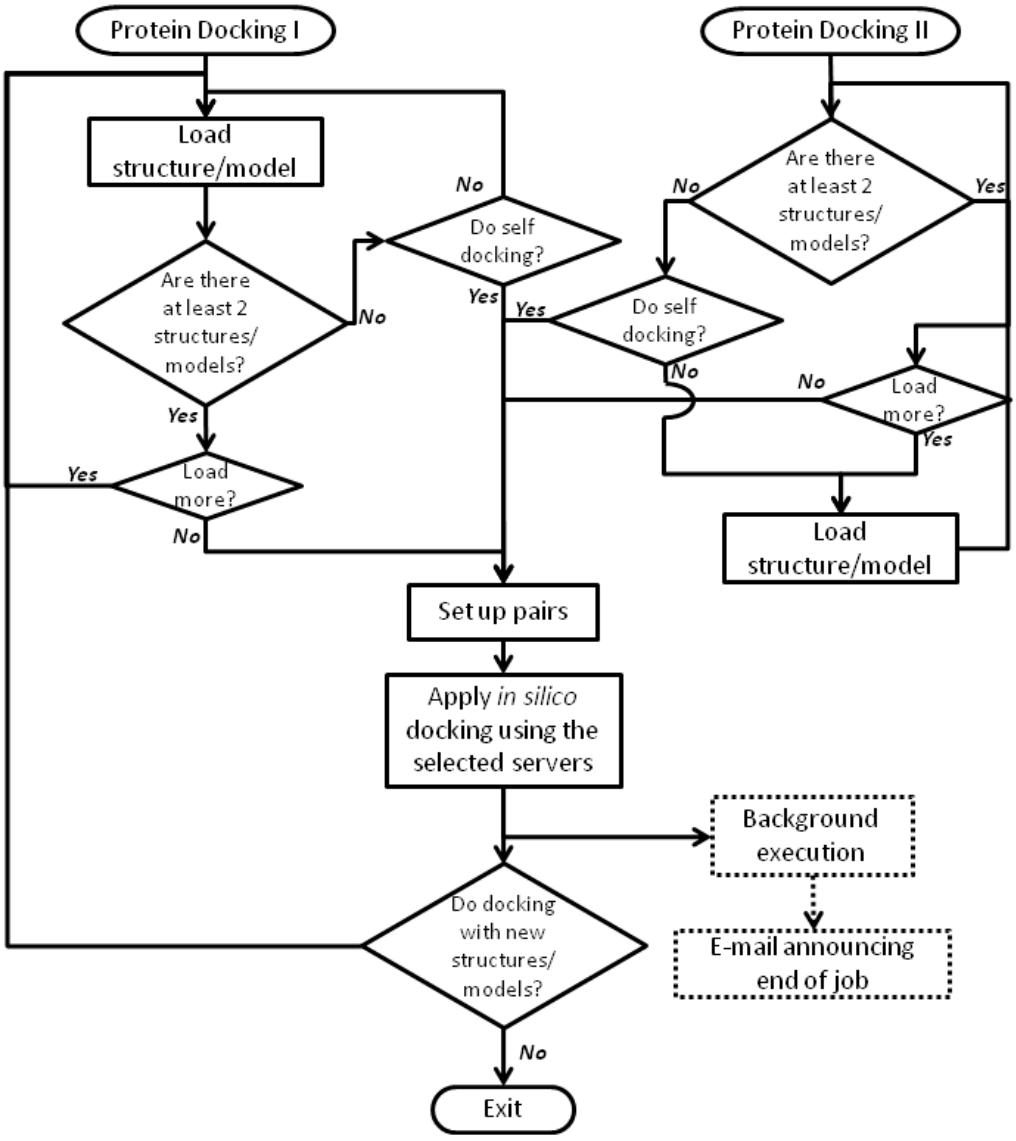


Figure 3. Workflow chart for protein PPD. On the one hand, if Protein-MetReS has no previous structural results, it prompts the user to load structures or models for subsequent docking (Protein Docking I). On the other hand, if there are previous structural results for the user, s/he is led directly to the PPD functionality (Protein Docking II). At this stage, Protein-MetReS asks the user to decide which protein to dock and which docking servers to use. Once this is settled, the program runs the docking process in the background. Self docking is also allowed by the program.

6.3.4 Visualizing protein docking results

Protein-MetReS displays the following information for the docking of a pair of proteins in tabular format: docking method, model and/or structures that were used in the docking, as well as the ten best docking results from each tool. A

"complex details" view is also available for each complex, where the complex image linked to the interactive Jmol 3D molecular viewer is displayed.

6.3.5 Protein-MetReS vs. Protein Model Portal

Given that Protein-MetReS and PMP have partially overlapping audiences, we need to compare how the set of modeling tools used by both systems differs. This is done in Table 4. In addition, we also need to compare the repositories used by both. This is done in Table 5.

Table 4. Modeling tools used by both, Protein-MetReS and Protein Model Portal.

Protein-MetReS	Protein Model Portal
SWISS-MODEL[11]	SWISS-MODEL[11]
MODELLER[12]	ModWeb[13]
PHYRE[14]	M4T[15]
	HHPred[16]
	I-TASSER[17]

In summary, PMP integrates more modeling tools than Protein-MetReS, but the coverage of methods and results are similar. ModWeb [26] is based on the MODELLER tool, and M4T [27] use similar techniques as MODELLER. By using the standalone version of MODELLER Protein-MetReS covers practically the same modeling space and characteristics than both, ModWeb and M4T.

HHPred [28] is a modeling tool that identifies homology and predicts protein-structure by using **Hide Markov Models** (**HMM**) via HHsearch [29]. Phyre2 also uses HHsearch and incorporates a new *ab initio* method called Poing [30] to model regions of proteins with no detectable homology to known structures. I-TASSER [31] is a service for protein structure prediction where the 3D models are built by using FRM and threading techniques. It also uses AIM to model regions of the target sequence that are not aligned to a template (mainly loops)

[11] http://swissmodel.expasy.org/workspace/index.php?func=modelling_simple1
[12] http://salilab.org/modeller/
[13] https://modbase.compbio.ucsf.edu/scgi/modweb.cgi
[14] http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index
[15] http://manaslu.aecom.yu.edu/M4T/
[16] http://toolkit.lmb.uni-muenchen.de/hhpred
[17] http://zhanglab.ccmb.med.umich.edu/I-TASSER/

or when no appropriate template is identified. I-TASSER follows a similar procedure as Phyre2 to build the 3D structures, but uses a different methodology in doing so.

Taking all these factors into account, we decided that the first prototype of Protein-MetReS would only include Phyre2. Future versions might include additional FRM servers, if this is deemed to improved the predictive capabilities of the application. In addition, ROSETTA will be locally installed to allow for AIM by the application.

Table 5. Repositories used by both, Protein-MetReS and Protein Model Portal.

Protein-MetReS	Protein Model Portal
SMR[18]	SMR[16]
PDB[19]	PDB[17]
	ModBase[20]
	CSMP[21], NESG[22], NYSGRC[23]
	GPCRDB[24]

Regarding structural repositories, PMP uses the repositories shown in Table 5. Such repositories are the CSMP (Center for Structures of Membrane Proteins) [32,33], the NESG (NorthEast Structural Genomics consortium) [34], the NYSGRC (New York Structural Genomics Research Consortium) [35] and the GPCRDB (G Protein-Coupled-Receptor DataBase) [36]. This information is stored in the centralized PDB and that is why we chose to include only this repository in Protein-MetReS. Regarding model repositories, PMP uses the ModBase repository[26], which contains the models created by ModWeb, and the SMR, which contains models built by SWISS-MODEL. We, in turn, chose to use the SMR and use MODELLER to build our own database of MODELLER-based models.

[18] http://swissmodel.expasy.org/repository/
[19] http://www.rcsb.org/pdb/home/home.do
[20] http://modbase.compbio.ucsf.edu/modbase-cgi/index.cgi
[21] http://csmp.ucsf.edu/index.htm
[22] http://www.nesg.org/
[23] http://www.nysgrc.org/psi3-cgi/index.cgi
[24] http://www.gpcr.org/7tm/

6.4 Discussion

Here we present Protein-MetReS, a user friendly tool to unify access to different modeling tools and to different docking tools. The modeling tools integrated by Protein-MetReS cover most of the result space covered by those integrated in the PMP. Protein-MetReS clearly differentiates itself from PMP by integrating various PPD tools.

There are two aspects in which future development of the structure analysis and prediction functionality of Protein-MetReS will focus. First, future implementations of Protein-MetReS will invest in implementing an AIM to cover situations where HM and FRM are not sensitive enough to create modeling structure. This will be done by installing a local copy of ROSETTA, the most successful AIM to date. Second, future versions of Protein-MetReS will include more sophisticated methods to evaluate the quality of 3D models. At this stage, we either use straightforward evaluations implemented locally, when the structural model is created by MODELLER, or rely on the evaluations that SWISS-MODEL and PHYRE2 perform. Having high quality structural models is very important if one wants to increase the reliability of PPD analysis of these models.

There are also aspects in which future development of the PPD functionality of Protein-MetReS will focus. Chief among these is the identification of the "real" complex from the list of those predicted by PPD methods. While this complex is typically among the top 10 solutions generated by PPD docking programs, it is not always easily identified. Future versions of Protein-MetReS will implement methods to identify correlated mutations in the sequence of proteins being docked. These correlated mutations can then be used to filter out spurious docking solutions, based on the assumption that the co-evolution of the sequence for the individual members of a protein-protein complex in different organisms keeps traces of the evolution of the complex [37].

6.5 References

1. Berman HM, Bhat TN, Bourne PE, Feng Z, Gilliland G, et al. (2000) The Protein Data Bank and the challenge of structural genomics. Nat Struct Biol 7 Suppl: 957–959. doi:10.1038/80734.
2. Moult J, Fidelis K, Kryshtafovych A, Tramontano A (2011) Critical assessment of methods of protein structure prediction (CASP)-round IX. Proteins Struct Funct Bioinforma 79: 1–5. doi:10.1002/prot.23200.
3. Fernández-Recio J, Sternberg MJE (2010) The 4th meeting on the Critical Assessment of Predicted Interaction (CAPRI) held at the Mare Nostrum, Barcelona. Proteins Struct Funct Bioinforma 78: 3065–3066. doi:10.1002/prot.22801.
4. Janin J, Henrick K, Moult J, Eyck LT, Sternberg MJE, et al. (2003) CAPRI: a Critical Assessment of PRedicted Interactions. Proteins 52: 2–9. doi:10.1002/prot.10381.
5. Kiefer F, Arnold K, Künzli M, Bordoli L, Schwede T (2009) The SWISS-MODEL Repository and associated resources. Nucleic Acids Res 37: D387–392. doi:10.1093/nar/gkn750.
6. Bordoli L, Kiefer F, Arnold K, Benkert P, Battey J, et al. (2009) Protein structure homology modeling using SWISS-MODEL workspace. Nat Protoc 4: 1–13. doi:10.1038/nprot.2008.197.
7. Arnold K, Bordoli L, Kopp J, Schwede T (2006) The SWISS-MODEL workspace: a web-based environment for protein structure homology modelling. Bioinforma Oxf Engl 22: 195–201. doi:10.1093/bioinformatics/bti770.
8. Peitsch MC (1995) Protein Modeling by E-mail. Bio/Technology 13: 658–660. doi:10.1038/nbt0795-658.
9. Haas J, Roth S, Arnold K, Kiefer F, Schmidt T, et al. (2013) The Protein Model Portal--a comprehensive resource for protein structure and model information. Database 2013: bat031–bat031. doi:10.1093/database/bat031.
10. Goldman BB, Wipke WT (2000) QSD quadratic shape descriptors. 2. Molecular docking using quadratic shape descriptors (QSDock). Proteins 38: 79–94.
11. Meng EC, Shoichet BK, Kuntz ID (1992) Automated docking with grid-based energy evaluation. J Comput Chem 13: 505–524. doi:10.1002/jcc.540130412.
12. Morris GM, Goodsell DS, Halliday RS, Huey R, Hart WE, et al. (1998) Automated docking using a Lamarckian genetic algorithm and an empirical binding free energy function. J Comput Chem 19: 1639–1662. doi:10.1002/(SICI)1096-987X(19981115)19:14<1639::AID-JCC10>3.0.CO;2-B.
13. Feig M, Onufriev A, Lee MS, Im W, Case DA, et al. (2004) Performance comparison of generalized born and Poisson methods in the calculation of electrostatic solvation energies for protein structures. J Comput Chem 25: 265–284. doi:10.1002/jcc.10378.

14. Halperin I, Ma B, Wolfson H, Nussinov R (2002) Principles of docking: An overview of search algorithms and a guide to scoring functions. Proteins 47: 409–443. doi:10.1002/prot.10115.
15. Eswar N, Webb B, Marti-Renom MA, Madhusudhan MS, Eramian D, et al. (2007) Comparative protein structure modeling using MODELLER. Curr Protoc Protein Sci Editor Board John E Coligan Al Chapter 2: Unit 2.9. doi:10.1002/0471140864.ps0209s50.
16. Sánchez R, Sali A (1997) Evaluation of comparative protein structure modeling by MODELLER-3. Proteins Suppl 1: 50–58.
17. Kelley LA, Sternberg MJE (2009) Protein structure prediction on the Web: a case study using the Phyre server. Nat Protoc 4: 363–371. doi:10.1038/nprot.2009.2.
18. Tovchigrechko A, Vakser IA (2006) GRAMM-X public web server for protein-protein docking. Nucleic Acids Res 34: W310–314. doi:10.1093/nar/gkl206.
19. Tovchigrechko A, Vakser IA (2005) Development and testing of an automated approach to protein docking. Proteins 60: 296–301. doi:10.1002/prot.20573.
20. Ritchie DW (2005) High-order analytic translation matrix elements for real-space six-dimensional polar Fourier correlations. J Appl Crystallogr 38: 808–818. doi:10.1107/S002188980502474X.
21. Mustard D, Ritchie DW (2005) Docking essential dynamics eigenstructures. Proteins 60: 269–274. doi:10.1002/prot.20569.
22. Ritchie DW, Kemp GJ (2000) Protein docking using spherical polar Fourier correlations. Proteins 39: 178–194.
23. Lyskov S, Gray JJ (2008) The RosettaDock server for local protein-protein docking. Nucleic Acids Res 36: W233–238. doi:10.1093/nar/gkn216.
24. Lyskov S, Chou F-C, Conchúir SÓ, Der BS, Drew K, et al. (2013) Serverification of molecular modeling applications: the Rosetta Online Server that Includes Everyone (ROSIE). PloS One 8: e63906. doi:10.1371/journal.pone.0063906.
25. Jmol: an open-source Java viewer for chemical structures in 3D. (n.d.). Available: http://jmol.sourceforge.net/.
26. Pieper U, Eswar N, Braberg H, Madhusudhan MS, Davis FP, et al. (2004) MODBASE, a database of annotated comparative protein structure models, and associated resources. Nucleic Acids Res 32: D217–222. doi:10.1093/nar/gkh095.
27. Rykunov D, Steinberger E, Madrid-Aliste CJ, Fiser A (2009) Improved scoring function for comparative modeling using the M4T method. J Struct Funct Genomics 10: 95–99. doi:10.1007/s10969-008-9044-9.
28. Söding J, Biegert A, Lupas AN (2005) The HHpred interactive server for protein homology detection and structure prediction. Nucleic Acids Res 33: W244–248. doi:10.1093/nar/gki408.
29. Söding J (2005) Protein homology detection by HMM-HMM comparison. Bioinforma Oxf Engl 21: 951–960. doi:10.1093/bioinformatics/bti125.

30. Jefferys BR, Kelley LA, Sternberg MJE (2010) Protein folding requires crowd control in a simulated cell. J Mol Biol 397: 1329–1338. doi:10.1016/j.jmb.2010.01.074.
31. Roy A, Kucukural A, Zhang Y (2010) I-TASSER: a unified platform for automated protein structure and function prediction. Nat Protoc 5: 725–738. doi:10.1038/nprot.2010.5.
32. Stroud RM, Choe S, Holton J, Kaback HR, Kwiatkowski W, et al. (2009) 2007 annual progress report synopsis of the Center for Structures of Membrane Proteins. J Struct Funct Genomics 10: 193–208. doi:10.1007/s10969-008-9058-3.
33. Phillips GN Jr, Fox BG, Markley JL, Volkman BF, Bae E, et al. (2007) Structures of proteins of biomedical interest from the Center for Eukaryotic Structural Genomics. J Struct Funct Genomics 8: 73–84. doi:10.1007/s10969-007-9023-6.
34. Xiao R, Anderson S, Aramini J, Belote R, Buchwald WA, et al. (2010) The high-throughput protein sample production platform of the Northeast Structural Genomics Consortium. J Struct Biol 172: 21–33. doi:10.1016/j.jsb.2010.07.011.
35. Sauder MJ, Rutter ME, Bain K, Rooney I, Gheyi T, et al. (2008) High throughput protein production and crystallization at NYSGXRC. Methods Mol Biol Clifton NJ 426: 561–575. doi:10.1007/978-1-60327-058-8_37.
36. Horn F, Bettler E, Oliveira L, Campagne F, Cohen FE, et al. (2003) GPCRDB information system for G protein-coupled receptors. Nucleic Acids Res 31: 294–297.
37. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A (1997) Correlated mutations contain information about protein-protein interaction. J Mol Biol 271: 511–523. doi:10.1006/jmbi.1997.1198.

Chapter 7. Discussion

7.1 General Remarks

This thesis describes the development of a set of tools that automatically analyze information contained in large datasets, with the ultimate goal of reconstructing biological molecular circuits. The work reported here focuses mostly on the identification of biologically relevant protein interactions. Protein interactions are basic for life processes and the study of such interactions helps in the study and interpretation of the information encoded in genomes and facilitates the understanding of the cell systems. A wide array of experimental techniques can be used to study protein interaction. These include yeast two-hybrid [1], tandem affinity purification [2], mass spectrometry [3], immunoprecipitation [4], pulldown assay [5], phage display [6], and protein chips [7], among others. Each of these methods is very resource intensive, requiring the use of many man-hours and/or expensive equipment and reactives.

Because of this intensive use of resources and time, computational methods provide an indispensable alternative to the prediction of protein-protein interactions (PPIs) that can potentially speed up and lower the cost of PPI studies. The only computational approaches that directly model physical interactions between proteins are docking [8,9] and binding simulations [10]. For high-throughput structural analysis on a genome scale, docking approaches are potentially suitable. This is because such approaches are focused on the final configuration(s) of the complex rather than the modeling of real binding pathways as binding simulations approaches. Over the past few years, owing to the CASP and CAPRI competitions, there has been a gradual increase in the accuracy of predicting protein structures and complexes. However, this is still not enough and further developments are still required to increase the accuracy and extend the capabilities of *in silico* protein docking. Improvements in accuracy could focus, for example, and as discussed in Chapter 6, on the problem that the correct complex configuration is often not the top ranked solution in a docking experiment. So, better discriminating functions need to be

developed for the docking experiments to discriminate the correct complex among all solutions. These functions can be coupled to other methods, such as the identification of correlated mutation in complementary protein surfaces [11], or the usage of well-known false-positive solutions, denoted as decoys [12]. Extending the capabilities of docking methods could focus on the fact that, in the presence of two protein structures, these methods will always produce a set of possible complexes between those proteins. Incorporating methods that can filter out impossible docking pairs as a first step of the docking would thus extend the capabilities of docking experiments. Some possible alternatives for this filtering process are now described.

One alternative is by using bibliomic analysis and automatically extracting relevant biological information from scientific documents [13–18]. Another alternative is by using bioinformatics methods to predict cellular localization of proteins [19,20]. A third example would be to analyze when proteins are expressed during cell cycle. Evolutionary information could also be used for this. For example, phylogenetic profiling can be used to identify pairs of proteins with a high probability of functional interaction. This method uses homology information to describe a gene's context in fully sequenced genomes relying on the following idea: if two or more genes or proteins are simultaneously present or absent in the same set of genomes, such genes have a high probability of being involved in a common function because their presence or absence may indicate simultaneous co evolution of proteins [21]. Homology transfer of experimentally determined interactions can also be used [22]. By comparing the proteins that are known to physically interact in an organism and transferring that information through homology and analogy to other organisms, one can also prioritize proteins pairs for docking. This information is available in databases such as DIP[1] (Database of Interacting Proteins), IntAct[2], MINT[3] (Molecular INTeraction) and MIPS[4] (Mammalian Protein-Protein Interaction).

[1] http://dip.doe-mbi.ucla.edu/dip/Main.cgi
[2] http://www.ebi.ac.uk/intact/
[3] http://mint.bio.uniroma2.it/mint/Welcome.do

The use of these methods for filtering pairs of proteins to dock would be straight forward: Protein pairs that are localized in different cellular compartments, have failed to interact in experiments in other organisms, have incompatible phylogenetic conservation profiles, or are not present at the same time in the cell can be discarded.

In general, identifying PPI networks enables the reconstruction of biological circuits and the identification of variations in the design of those circuits. The information extracted for any single dataset or derived by specific methods is, at best, partial. Thus, it is the integration of many parallel datasets and reconstruction methods that can hope to increase the accuracy of computational predictions of circuit variants. We hope to have contributed for this increase with the set of computational tools that use different data sets. Such tools are Biblio-MetReS, Protein-MetReS, CheNER and CheNER-BioC. Biblio- and Protein-MetReS have immediate applicability in circuit reconstruction. In contrast CheNER is an initial step that permits identifying chemical entities. This identification can, in the future, be used for the development of tools that identify chemical regulation of biological circuits. As was briefly discussed in the Introduction, this is fundamental to help discovering the design principles in molecular and cellular circuits.

[4] http://mips.helmholtz-muenchen.de/proj/ppi/

7.2 Future directions

The text-mining tools developed during the course of this thesis achieved a good performance in comparison with other available tools. To keep this performance it is important to regularly update these tools to include relevant improvements in text-mining techniques and enhance their performance. It is also important to incorporate new functionalities that facilitate the interpretation of the results. For instance, detection of interactions in PPI networks would be enhanced by including functionalities that detect the context of that interaction. For example, a study of action words associated to the interaction, such as regulate, activate, etc., would be very helpful for the causal reconstruction of protein structure. Immediate improvements to Protein-MetReS should prioritize the inclusion of a new modeling server implementing an *ab initio* modeling method, as well as quality estimators for the protein predicted structures. Another priority should be the incorporation of functionalities to discriminate the correct complex generated by docking among all solutions, for example by using decoys.

Tools that analyze other types of large scale data should also be included in the MetReS project. For example, a set of tools that enables the usage of phylogenetic and other evolutionary information to reconstruct protein-protein interactions would be very useful. Likewise, a tool integrating information from databases of experimentally determined physical protein-protein interactions that would allow for a transfer of this information among organisms homology should be developed. Moreover, tools that integrate and analyze metabolomic information and gene expression data to aid in the reconstruction of protein-protein interactions would greatly enhance the scope of the project.

Collecting all this information in a complementary and integrated set of tools would greatly contribute for the automated reconstruction of molecular circuits and facilitate the discovery and analysis of design principles in those circuits.

7.3 References

1. Bartel PL, Fields S (1995) Analyzing protein-protein interactions using two-hybrid system. Methods Enzymol 254: 241–263.
2. Puig O, Caspary F, Rigaut G, Rutz B, Bouveret E, et al. (2001) The tandem affinity purification (TAP) method: a general procedure of protein complex purification. Methods San Diego Calif 24: 218–229. doi:10.1006/meth.2001.1183.
3. Ho Y, Gruhler A, Heilbut A, Bader GD, Moore L, et al. (2002) Systematic identification of protein complexes in Saccharomyces cerevisiae by mass spectrometry. Nature 415: 180–183. doi:10.1038/415180a.
4. Klupp BG, Böttcher S, Granzow H, Kopp M, Mettenleiter TC (2005) Complex formation between the UL16 and UL21 tegument proteins of pseudorabies virus. J Virol 79: 1510–1522. doi:10.1128/JVI.79.3.1510-1522.2005.
5. Dellis S, Strickland KC, McCrary WJ, Patel A, Stocum E, et al. (2004) Protein interactions among the vaccinia virus late transcription factors. Virology 329: 328–336. doi:10.1016/j.virol.2004.08.017.
6. Mullaney BP, Pallavicini MG (2001) Protein-protein interactions in hematology and phage display. Exp Hematol 29: 1136–1146.
7. Zhu H, Bilgin M, Bangham R, Hall D, Casamayor A, et al. (2001) Global analysis of protein activities using proteome chips. Science 293: 2101–2105. doi:10.1126/science.1062191.
8. Vajda S, Sippl M, Novotny J (1997) Empirical potentials and functions for protein folding and binding. Curr Opin Struct Biol 7: 222–228.
9. Sternberg MJ, Gabb HA, Jackson RM (1998) Predictive docking of protein-protein and protein-DNA complexes. Curr Opin Struct Biol 8: 250–256.
10. McCammon JA (1998) Theory of biomolecular recognition. Curr Opin Struct Biol 8: 245–249.
11. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A (1997) Correlated mutations contain information about protein-protein interaction. J Mol Biol 271: 511–523. doi:10.1006/jmbi.1997.1198.
12. Graves AP, Brenk R, Shoichet BK (2005) Decoys for Docking. J Med Chem 48: 3714–3728. doi:10.1021/jm0491187.
13. Usié A, Karathia H, Teixidó I, Valls J, Faus X, et al. (2011) Biblio-MetReS: A bibliometric network reconstruction application and server. BMC Bioinformatics 12: 387. doi:10.1186/1471-2105-12-387.
14. Von Mering C, Jensen LJ, Kuhn M, Chaffron S, Doerks T, et al. (2007) STRING 7—recent developments in the integration and prediction of protein interactions. Nucleic Acids Res 35: D358–D362. doi:10.1093/nar/gkl825.
15. Snel B, Lehmann G, Bork P, Huynen MA (2000) STRING: a web-server to retrieve and display the repeatedly occurring neighbourhood of a gene. Nucleic Acids Res 28: 3442–3444. doi:10.1093/nar/28.18.3442.
16. Szklarczyk D, Franceschini A, Kuhn M, Simonovic M, Roth A, et al. (2011) The STRING database in 2011: functional interaction networks of proteins,

globally integrated and scored. Nucleic Acids Res 39: D561–568. doi:10.1093/nar/gkq973.
17. Fernández JM, Hoffmann R, Valencia A (2007) iHOP web services. Nucleic Acids Res 35: W21–26. doi:10.1093/nar/gkm298.
18. Hoffmann R, Valencia A (2005) Implementing the iHOP concept for navigation of biomedical literature. Bioinformatics 21: ii252–ii258. doi:10.1093/bioinformatics/bti1142.
19. Lin H-N, Chen C-T, Sung T-Y, Ho S-Y, Hsu W-L (2009) Protein subcellular localization prediction of eukaryotes using a knowledge-based approach. BMC Bioinformatics 10: S8. doi:10.1186/1471-2105-10-S15-S8.
20. Chou K-C, Shen H-B (2010) A New Method for Predicting the Subcellular Localization of Eukaryotic Proteins with Both Single and Multiple Sites: Euk-mPLoc 2.0. PLoS ONE 5: e9931. doi:10.1371/journal.pone.0009931.
21. Lin T-W, Wu J-W, Chang DT-H (2013) Combining Phylogenetic Profiling-Based and Machine Learning-Based Techniques to Predict Functional Related Proteins. PLoS ONE 8: e75940. doi:10.1371/journal.pone.0075940.
22. Chen C-C, Lin C-Y, Lo Y-S, Yang J-M (2009) PPISearch: a web server for searching homologous protein-protein interactions across multiple species. Nucleic Acids Res 37: W369–375. doi:10.1093/nar/gkp309.

Chapter 8. Conclusions

1. The development of Biblio-MetReS to reconstruct molecular networks through automated text-mining of scientific documents led to the following conclusions:
   - The software has a performance in identifying gene co-occurrences that is similar to that of other comparable tools (iHOP, STRING).
   - The software generates gene-gene co-occurrence networks that are more up to date with the scientific knowledge than those generated by comparable tools. This comes at the cost of being slower than iHOP or STRING.
   - The software generates two new co-occurrence networks: (1) proteins with biological process/pathways and (2) biological process/pathways with themselves, which is a functionality that is unique to this program. The program provides clear statistical estimators to evaluate the significance of each interaction via calculating mutual information and p-value.
2. The development of CheNER to identify different types of chemical compounds through automated text-mining of scientific documents led to the following conclusions:
   a. The program has the best performance in identifying standard IUPAC names alone.
   b. The BioCreAtIvE challenge showed that our program has the best performance of any single application in identifying chemical names in general. However, this performance is easily surpassed by methods that combine different tools and generate a unified output.
3. The development of Protein-MetReS to integrate different resources for protein structure, analysis, modeling, and docking led to the following conclusions:
   a. Protein-MetReS is the only tool of its kind that integrates structural analysis, prediction and docking functionalities.

   b. In terms of structural analysis, Protein-MetReS is not yet as complete as the Protein Model Portal in terms of integrating modeling tools and quality evaluation methods for structural models.

4. I provided a set of tools to the community that can facilitate the initial steps of circuit reconstruction to aid in the discovery of biological design principles for those circuits.