Bioinformatics pre-selection of thioredoxin/glutaredoxin target proteins for the construction of cellular redox regulatory network

Dissertation

zur Erlangung des akademischen Grades eines
Doctors der Naturwissenschaften
der Universität Bielefeld.
Vorgelegt von

Hang-mao Lee

Bielefeld, im Mai 2012

MSc. Hang-mao Lee
AG Bioinformatik und Medizinische Informatik
Technische Fakultät
Universität Bielefeld
Email: hmlee@Techfak.Uni-Bielefeld.DE

Der Technischen Fakultät der Universität Bielefeld
am 30 Mai 2012 vorgelegt.

Gutachter:
Prof. Dr. Ralf Hofestädt, Universität Bielefeld
Prof. Dr. Karl-Josef Dietz, Universität Bielefeld

Prüfungsausschuß:
Prof. Dr. Ralf Hofestädt, Universität Bielefeld
Prof. Dr. Karl-Josef Dietz, Universität Bielefeld
Prof. Dr. Robert Giegerich, Universität Bielefeld
Dr. Ute von Wangenheim, Universität Bielefeld

Printed on non-aging paper according to DIN-ISO 9706

Abstract

The gradually accumulated knowledge of molecular interaction is assembled into biological network to show the global picture of biological system. The biological network construction is usually based on the data from biological databases or literature. Once a specialized or less investigated biological network is focused, the issue of data scarcity in the database and literature emerges.

Redox regulatory network sustains the redox homeostasis in the cell, and its capacity has impact on the functionality of its target protein. One critical step for extending the redox regulatory network is the identification of target protein of thioredoxin (Trx)/glutaredoxin (Grx). However, the redox regulatory network has been better explored in plants than in animal. When the specialized topic, such as the construction of redox regulatory network in human mitochondrion which this thesis is tackling, is focused, little information can be obtained through conventional methods of network construction, such as querying the biological databases or mining of literatures.

To overcome the data deficiency problem of the specialized topic, a bottom-up strategy is adopted to first identify the oxidation susceptible cysteine, which is an important feature for the chemical reaction mechanism between Trx/Grx and their target protein. In the first part of the thesis, a pre-selection tool for Trx/Grx target protein, termed ROCD, is implemented following a computational decision tree discovered from the study of physicochemical properties. ROCD pre-selected a group of proteins which contains the potential candidate and requires further validation. One of the validation methods for the computational prediction is through search for relevant literature. And again, owing to the same information deficiency issue from the specialized research topic, the directly relevant literature is missing most of the time. The second part of the thesis introduces a network-contexted document retrieval system, termed ncDocReSy, to assist the retrieval of indirectly relevant literature

based on the topology of biological network. ROCD is applied on the pre-selection of Trx/Grx target protein in the mitochondrion of human liver with the physicochemical values suggested from other study and results in 309 potential candidates. After the pre-selection step, ncDocReSy can be used in the process of manual curation of the pre-selection result by providing indirectly relevant literature.

In this thesis work, several bioinformatics facilities assisting resource integration were used, such as the ID mapping service and standard data exchange formats. These facilities help the communication and mutual understanding between different resources and are essential for the integrative usage of bioinformatics resources.

Table of Contents

List of Abbreviations

API	Application Programming Interface
ASA	Accessible Surface Area
ASN.1	Abstract Syntax Notation One
ATM	Automatic Term Mapping
BALOSCTdb	Balanced Susceptible Cysteine Thiol Database
BNDB	Biochemical Network Database
COPA	Cysteine Oxidation Prediction Algorithm
CORBA	Common Object Request Broker Architecture
CSML	Cell System Markup Language
EHMN	Edinburgh Human Metabolic Network
eUtils	Entrez Programming Utilities
GEO	Gene Expression Omnibus
GO	Gene Ontology
Grx	Glutaredoxin
GSH	Glutathione

HFPN	Hybrid Functional Petri Net
HFPNe	Hybrid Functional Petri Net with Extension
HPRD	Human Protein Reference Database
HTTP	Hypertext Transfer Protocol
IR	Information Retrieval
JAXB	Java Architecture XML Binding
MeSH	Medical Subject Headings
MIAPE	Minimum Information about a Proteomics Experiment
MIAME	Minimum Information about a Microarray Experiment
MIMIx	Minimum Information Required for Reporting a Molecular Interaction Experiment
MS	Mass Spectrometry
ncDocReSy	Network-Contexted Document Retrieval System
ncRLL	Network-Context Ranked Literature List
NER	Named-Entity Recognition
NLM	National Library of Medicine
NMR	Nuclear Magnetic Resonance
nOSC	Oxidation-non-Susceptible Cysteine
OBO	Open Biomedical Ontologies
ODE	Ordinary Differential Equation
OSC	Oxidation-Susceptible Cysteine
PBM	Protein Binding Microarray
pK_a	Acid Dissociation Constant
PMC	PubMed Central
PMCID	PMC Unique Identifier
PMID	PubMed Unique Identifier
PPI	Protein-Protein interaction

Prx	Peroxiredoxin
PSI-MI	Proteomics Standard Initiative Molecular Interaction
REST	Representational State Transfer
RNS	Redox Nitrogen Species
ROC	Reversibly Oxidized Cysteine
ROCD	Reversibly Oxidized Cysteine Detector
ROS	Reactive Oxygen Species
RRN	Redox Regulatory Network
SBGN	System Biology Graphical Notation
SBML	Systems Biology Markup Language
SELEX	Systematic Evolution of Ligands by Exponential Enrichment
SMR	SwissModel Repository
SOAP	Simple Object Access Protocol
SPACC	SwissProt Accession Number
SPID	SwissProt Identifier
TFBS	Transcription Factor Binding Site
TR	Thioredoxin Reductase
TTG	Target Protein of Thioredoxin and Glutaredoxin
Trx	Thioredoxin
UID	Unique Identifier
XML	Extensible Markup Language
YMD	Yale Microarray Database

Chapter 1

Introduction

1.1 Motivation

The chemical reaction of oxidation (withdrawal of electrons from a molecule) is linked to the process of reduction (uptake of electrons by a molecule) and is called redox reaction. Redox reactions are essential for the energy generation in the cell but also create chemical stress if they depart from the normal route. The redox regulatory system is the safeguard for the cellular redox status of the cell and counteracts the oxidative stress. The capacity of the redox regulatory systems has the impact on sustaining the normal function of proteins under oxidative stress.

The chloroplast in plants and mitochondria are organelles where vigorous electron transfer takes place and where reactive oxygen species are generated easily. Due to their central roles in energy metabolism of the cell, the redox homeostasis exerted by the redox regulatory systems in these organelles needs a deeper inspection.

Before an ultimately quantitative model of the redox regulatory network (RRN) can be build, the qualitative properties need to be defined first. One critical step in building a qualitative model of redox regulatory network is to identify the target proteins for the thioredoxin and glutaredoxin (TTG). Thioredoxins and glutaredoxins are the electron transmitters in the redox regulatory system and the reductants for the oxidized metabolic enzymes [Diet08].

The target proteins of thioredoxin and glutaredoxin have been investigated more in plants than in animals. Therefore, very few records could be found in the biological database or scientific literature when the database query targets animal species. However, the conventional method of biological network construction relies on the data from the biological databases or on the critical reading or computational analysis of scientific literature. The data deficiency in the database and literature poses an obstacle when the construction of a specialized and less investigated biological network, such as the redox regulatory network of human mitochondrion, is intended.

Despite little investigation of TTG in animals and humans, the underlining mechanism of chemical reaction between thioredoxin/glutaredoxin and the TTG is the same regardless of species. The prediction of TTG could be partially achieved with a bottom-up strategy by discriminating the existence of functional residue responsible for this specific chemical reaction (Fig. 1.1).

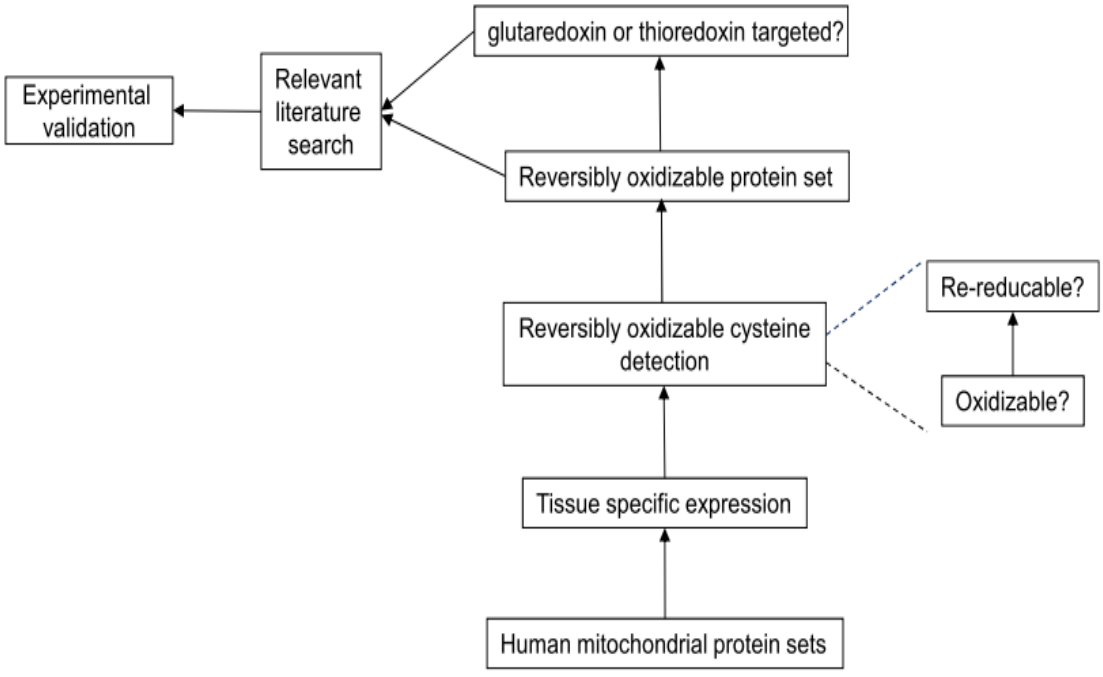


Figure 1.1 The bottom-up strategy of computational prediction of TTG in human mitochondrion. The TTG prediction workflow starts at identifying the tissue- and organelle-specific protein set which is followed by the identification of protein bearing cysteine that is oxidizable and re-reducable. The literature search can help the selection of promising candidate for experimental validation.

1.2 Aims

Due to incapability of simple database integration and text-mining of biological literature in answering the confronted biological question (explained later in chapter 3), a bottom-up strategy is adopted to preliminarily discriminate the functional residue essential for the chemical reaction between thioredoxin/glutaredoxin and its target. The discrimination method follows a decision rule generated from biochemical research and can be implemented into computer program which allows high-throughput processing. The implementation of such computer program carries out the pre-selection process for the final goal – identification of TTG, and the biologist can further filter the pre-selected result according to his/her domain knowledge or the published literature and then validate the pre-selected candidates by experiment. Resorting to literature is the common strategy for the interpretation of the experimental data and validation of the computational prediction. However, searching for the directly relevant literature concerning the less investigated or novel research topic usually fails. Thus searching for indirectly relevant literature will be beneficial for biologist's manual curation.

Due to the data and literature deficiency problem stated above, one aim of this dissertation work is to adopt a decision rule suggested from biochemical research on the proposed biological question to overcome the data deficiency issue in the biological network construction. The second aim was to assists biologist's manual curation of pre-selected result by providing relevant literature. Although this thesis focuses on the target protein of thioredoxin/glutaredoxin, the same strategy could be extended to other problems of target protein prediction, such as phosphorylation and ubiquitination, once the predicting rule for phosphorylated or ubiquitinated residue is available.

1.3 Structure

The thesis centers on two major topics as described in chapter 4 and 5 and uses the motivating biological question as an application case for these two work parts. The first three chapters prepare the essential knowledge to understand the two major works and pinpoint the biological question of redox regulation and redox regulatory networks motivating the thesis work. The two middle chapters describe the implementation of the major tools and are followed by an application of these tools to the initially posed biological question and conclusion.

Chapter 2 overviews the various bioinformatics resources for the cellular biology study. The heterogeneity in data structure and software architecture requires community-wide agreement on data structure and communication. Some bioinformatics facilities in help of integration process are thus introduced and will be mentioned in several places throughout the thesis. These facilities provide the knowledge of different biological themes as well as facilitate the communication and mutual understanding of the data content distributed in different knowledge bases. Chapter 3 presents the related works for the two major topics focused in the thesis. This chapter first introduces the biological background needed for the biological question addressed in this thesis. The biological molecules in redox regulatory network and the reaction mechanism of thiol-disulfide exchange, which is the essential reaction in the redox regulatory network, are mentioned here. This chapter also presents experimental and computational methods for the identification of members in redox regulatory network and points out the limitation of existing bioinformatics methodology on the proposed biological question. The limitation from the simple database integration and mining of unstructured text in the literature fosters the adoption of a bottom-up strategy by discriminating the amino acid which bears the desirable property. The discrimination rule is adopted directly from the published literature that endeavors to physical chemistry research. The second part of chapter 3 introduces the document retrieval systems in the biomedicine domain. The components of a document retrieval system are introduced, and the realization of these components is exemplified for PubMed. This second part serves as the background knowledge for chapter 5. Chapter 4 focuses on the work of a pre-selection tool, termed ROCD, for pre-selection of thioredoxin/glutaredoxin target protein. This pre-selection tool realizes the implement of a decision rule mentioned in chapter 3 to overcome the limitation from the existing bioinformatics tools. Chapter 5 is centered on a literature search tool, termed ncDocReSy, which can be utilized after the pre-selection process, so that the user can further decide promising candidates by reading the provided literature. ncDocReSy can be activated through JAVA Web Start from the result page of ROCD, so that the user can easily move from ROCD to ncDocReSy. The application of aforementioned tools on the pre-selection of target protein is provided in chapter 6. The results from the application of the pre-selection tool and network-contexted literature search are discussed in chapter 7. The work is concluded in chapter 8. The relationship between the chapter arrangement and the developed bioinformatics tools of this thesis work is shown in Fig. 1.2.

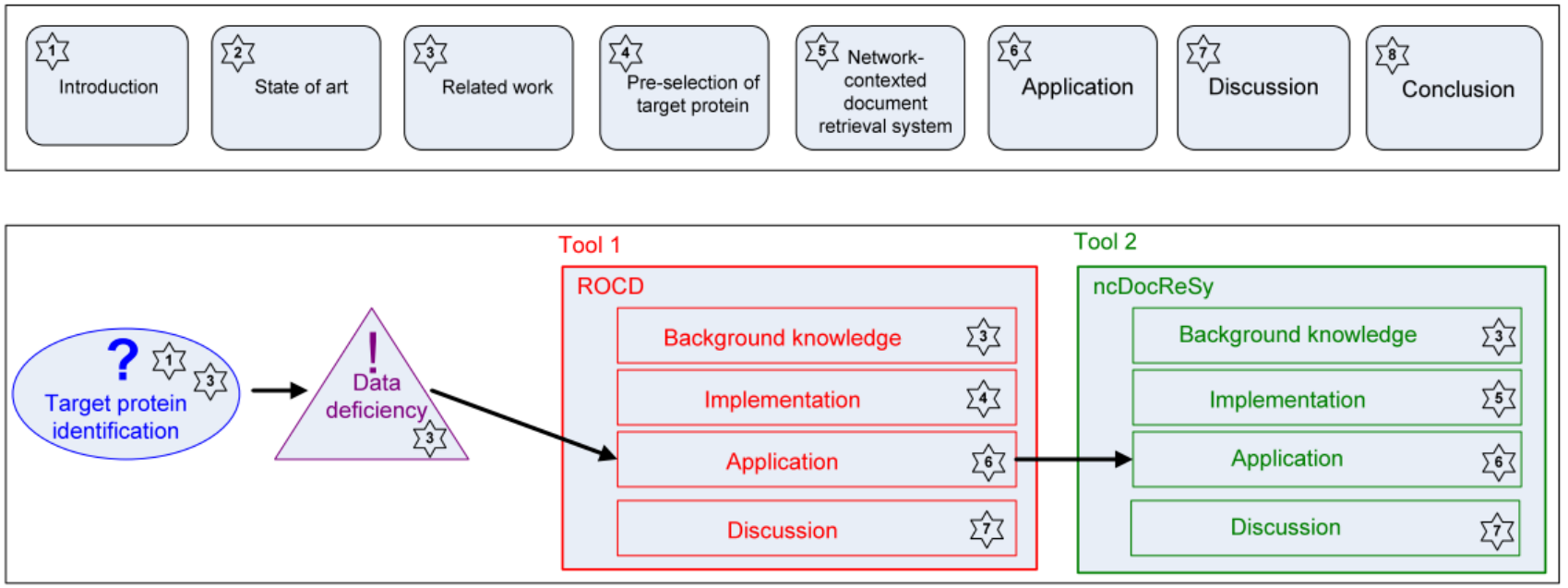


Figure 1.2 Chapter arrangement and the resolution strategy for the proposed biological question. The upper panel shows the chapter titles in this thesis. The lower panel shows the solution steps for the proposed biological question and the chapter of textual description.

Chapter 2

State of the art

With the continual development of experimental technologies for accumulating divergent molecular biology data, the development of bioinformatics is spurred by the diverse range of large-scale data which requires support from computer science for its storage and subsequent analysis [KB03]. With such abundant biological knowledge and the accumulated data from the experiments, bioinformatics resources were enriched by various databases and tools dealing with different types of data in the cellular biology. The various bioinformatics resources can be classified into four categories according to their implication in the different levels of functional hierarchy of the cell (Fig. 2.1). The resources in the bottom category deal with the atomic and residual property of macromolecule and come close to the discipline of physical chemistry, e.g. NetAcet [KBB05] for acetylated residue prediction and PHOSIDA [GGM11] for posttranslational modification. One category of resource has focused on the elementary building blocks of the cellular system–DNA/RNA, protein, and metabolite, e.g. HMDB [WKG+09] for human metabolites and Melina [OMM+07] for promoter analysis. Resources of this category deal with not only the static properties, such as functional region of the gene and three-dimensional structure of the protein, but also the dynamic properties which changes spatially and temporally, such as gene expression and post-translational modification. Another category of resource aims at the interaction between the elementary building blocks, such as protein-protein, protein-DNA, and protein-metabolite interactions, e.g. IntAct [KAB+12] for protein-protein interaction and STITCH [KSF+12] for protein-chemical interaction. The dataset included in this category is dynamic according to the condition of the cellular system. The fourth category of resources integrates the data from lower functional levels, establishes multidimensional

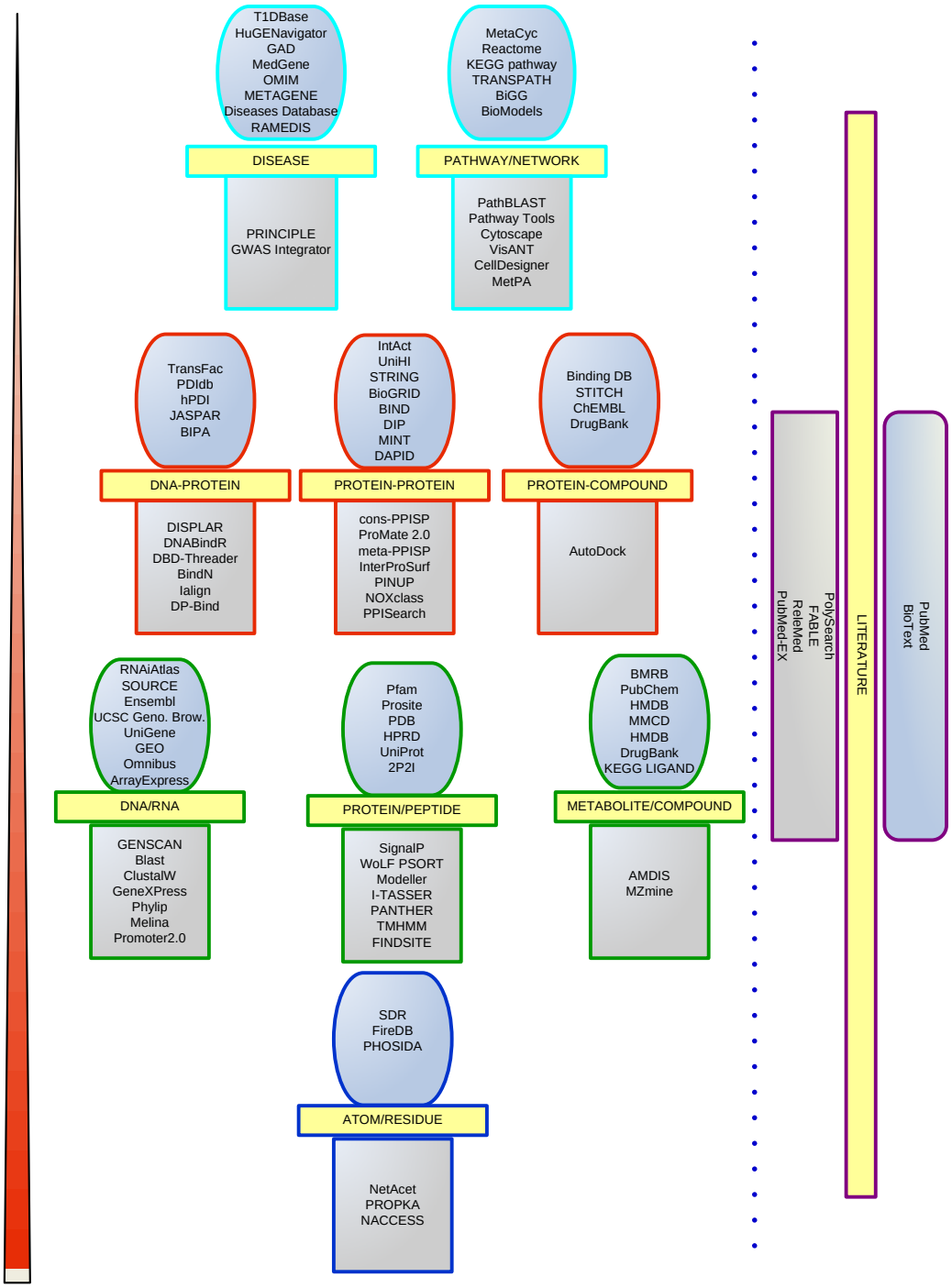


Figure 2.1 Bioinformatics resources categorized into four groups according to the functional hierarchy. The higher hierarchy is composed of the elements from the lower hierarchy. Parallel to the resources of the functional hierarchy is the unstructured knowledge embedded in the scientific literature. Different frame color has been applied to different level in the schematics.

networks and links them to phenotype. This category includes the resources describing metabolic pathway, gene regulatory network, and protein-protein interaction network, e.g. KEGG [KAG+08] for metabolic pathway and MetPA [XW10] for pathway analysis and visualization. Alongside the fore-mentioned knowledge stored in the structured biological databases, millions of journal article host the unstructured knowledge and have fostered the development of tools for information extraction.

Besides their essential function in structuring and exploiting –omics data, bioinformatics resources also adopt methodology from various disciplines, such as applied mathematics, statistics, biochemistry, chemistry, biophysics, to tackle the biological problem. The diverse resources of bioinformatics usually complement each other in terms of data coverage and capabilities. Observed from the current state of bioinformatics, the adaptation of an integrative strategy in bioinformatics could benefit from the merit of various available resources.

One application of integrative methodology on the biological question is exemplified by the work of Muehlberger *et al* [MMB+11]. Muehlberger *et al*. were interested in the molecular factors contributing to the bidirectional interplay between kidney and cardiovascular system and integrated the data from literature-mining, functional annotation of genes/proteins, network analysis, and identification of drug targets. Thus their analysis employed bioinformatics resources from different functional hierarchy (Fig. 2.2).

Bioinformatics databases and tools for cellular biology are very heterogeneous in term of the molecular type and property they concentrate on, such as UCSC genome browser [DKZ+12] for genomic sequence, UniProt [TUC10] for protein sequence and functional information, PDB [RBB+11] for the molecular structure. On the other hand, the same molecular type and its property are dealt by several analogous resources: gene expression data in ArrayExpress [PSK+11], Gene Expression Omnibus (GEO) [BTW+10] and Yale microarray database (YMD) [CWH+02]; pathway information in KEGG [KAG+08], BioCyc [KOM+05], and Reactome [VDS+07]; protein-protein interaction (PPI) data in HPRD [PGK+09], IntAct [AAA+10], and BioGrid [SBR+06]; the biological pathway visualization in Cytoscape [SMO+03] and CellDesigner [FMJ+08]. Each of these analogous resources contains a subset of information for the cellular system which at partly can complement each other. The data on human PPIs that come from six different primary databases show a small overlap, so that one way to increase coverage is to integrate data from different

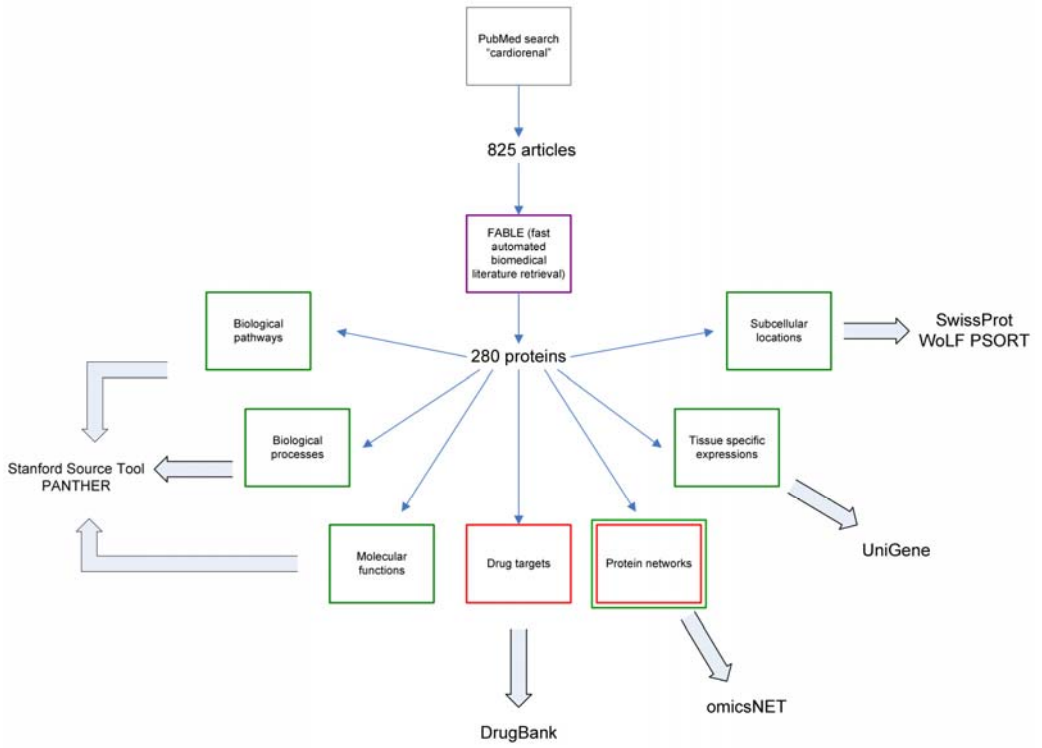


Figure 2.2 The analysis workflow and bioinformatics resources used in Muehlberger *et al*. (modified from Muehlberger *et al*.) This work combined bioinformatics resources devoted to gene, protein, drug, biological network, and literature mining. The colors applied to the square frame have the same implication as in Figure 2.1.

primary PPI databases [DF10]. Despite the similar type of information obtained from analogous resources, the data format, information granularity, and data annotation mostly are heterogeneous. Thus besides analogous data types, the heterogeneous data from different omics platform and higher level functional groupings (pathway, Gene Ontology, etc) must be integrated when a systematic analysis is conducted.

The integration of biological data has been concerned for long after the explosion of available biological data generated from genome sequence project and high-throughput experiment. The database integration allows the biologist to ask question that span several domains of knowledge through single portal interface [Stei03]. Several data integration approaches and applications have been developed to overcome the data scatter and redundancy issues in bioinformatics. Despite the data integration effort, the integrated database can only answer the question concerning the generally investigated research domains. Besides the structured knowledge from the database, the massive volume of biomedical literature also embeds unstructured knowledge as free text. The text-mining technique has been used to reveal the hidden knowledge from the literature. However, when a novel or specifically focused

question is asked, such as the motivating biological question of this thesis, the simple database integration and text-mining strategies will show little or no information.

During the process of integrated analysis with heterogenous data and tools, several facilities are helpful in enhancing the interoperability, unambiguous information representation, and programmatic efficiency. These facilities include the integrated databases, application programming interface, standard data exchange format, ontology, ID mapping tool, and the integration platform (Fig. 2.3).

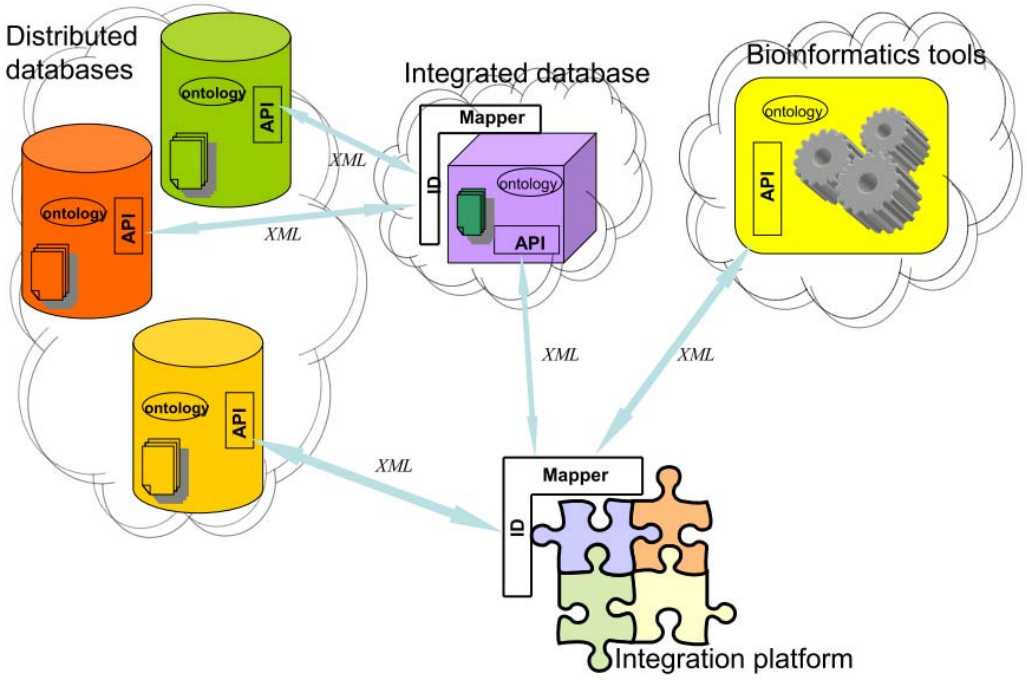


Figure 2.3 Facilities essential in integrative bioinformatics. The XML in this figure represents any standard data exchange format, such as SBML, and mzML.

2.1 Biological databases.

2.1.1 Primary databases for biological network construction

As shown in Fig. 2.1, various biological databases are available for the biomedical research. The most relevant databases for this thesis are the ones in the pathway/network category. Therefore, this section focuses on the biological databases for the work of biological network construction.

Since an organism is a complicated chemical factory, its pathways and networks become manageable only when we divide the whole system into several semi-isolated sub-systems and focus on one sub-system at a time. In the area of biological network

study, the biological network of cellular process is classified into three major sub-systems, gene regulatory network, protein-protein interaction network, and metabolic network.

There are several biological pathway/network and molecule databases which provide the building blocks of the biological network. The following paragraphs introduce some of the important biological pathway/network databases and classify them according to the type of database content.

2.1.1.1 Gene regulatory network and its resources

The genome is entire genetic information of an organism including the collection of genes in an organism. It contains static information which is mostly identical among organisms of the same species, although epigenetic modifications alter the use of the genome in dependence on previous life history. Although the genomic information is the same in each cell, the expression of its contained genes varies from tissue to tissue, and time to time. The spatial and temporal variations of gene expression enable the development and homeostatic episodes of the organism. The network which differentially controls the temporal and spatial expression is defined by the gene regulatory network. The gene regulatory network is composed of proteins known as transcription factors and the gene as regulated target of the transcription factors. Transcription factors (trans-regulatory element) recognize specific DNA motifs (cis-regulatory elements) in the gene promoter region and then recruit or intercept the transcription machinery, like RNA polymerase 2. The results would be the synthesis of mRNA or inhibition of the transcription according to the bonded transcription factors in the promoter.

The transcription of genes depends on the binding of adequate transcription factors to its cis-regulatory element. The information concerning the interaction between the transcription factor and its target cis-regulatory element is essential for the construction of gene regulatory network. The transcription factor binding site (TFBS) is usually represented as matrix-based binding site profile and sequence logo in the graphical form. This profile shows the relative frequency of each nucleotide for each base in the transcription factor/DNA binding region (Fig. 2.4). There are two main databases providing the information of transcription factor binding site profile.

JASPAR

JASPAR [SAE+04] emerged in 2004. Its aim was to provide high-quality, experimentally verified, and non-redundant transcription factor binding site profiles.

It concentrates on the transcription factor information of multicellular eukaryotes and is open-accessible. It has a web interface and also develops an API for programming tools. In 2010 JASPAR [PTK+09] had its fourth major release. The latest release includes the TFBS from yeast and also computationally derived profiles. In the first version JASPAR contained only transcription factor binding site profile derived from SELEX (Systematic Evolution of Ligands by EXponential enrichment) [PT90] experiment. With the advance of high-throughput experiment, JASPAR also includes the data from ChIP-seq and ChIP-chip experiments. JASPAR now holds 457 non-redundant, curated profiles for species from fungi to vertebrates, and a collection of the other 840 entries of profiles of regulatory regions, profiles derived from computationally predicted method, and profiles bases on protein binding microarray (PBM) experiment.

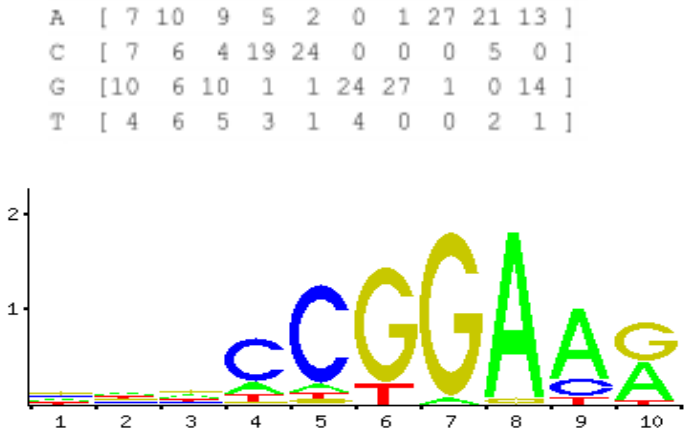


Figure 2.4 Transcription factor binding site profile and sequence logo of transcription factor Elk1. The transcription factor binding site profile shows the frequency of appearance of different nucleic acid in each position. The sequence logo depicts relative appearance frequency of nucleic acid of each position. Figure is extracted from JASPAR.

TRANSFAC

TRANSFAC® is a database on eukaryotic cis-acting regulatory DNA elements and trans-acting factors [MFG+03]. It covers the whole range of species from yeast to human. The binding site must be experimentally proven for their inclusion in the database. TRANSFAC® is maintained internally as a relational database with a web interface to the public. It provides information about the classification of the transcription factor, the cis-binding sequences, the gene regulated by the cis-element, the nucleotide weight matrices of cis-element, and the cell where the cis-element was discovered experimentally. The content of TRANSFAC® database is organized into six flat files. The web interface also provides the Match® tool, which searches for

putative transcription factor binding sites in DNA sequences based on weight matrices, and Patch®, whose function is similar to Match® but bases on single binding site sequence instead.

2.1.1.2 Protein-protein interaction network and its resources

Many biological functions are carried out by interaction of proteins. These functions might be transcription, post-translational modification, transport, complex formation and temporal and spatial regulation of diverse cellular processes. Signals from the environment are transferred to the responding machinery, such as the gene transcription, through protein-protein interactions. Proteins may interact to form protein complexes in order to exert their catalytic capability. The proteins involved in the same pathway may also form complexes to accelerate the metabolic turnover. Protein-protein interaction networks describe the pair-wise physical interaction between any two proteins. In the graphical representation of such a network, nodes represent the proteins, and edges represent the physical interaction and are non-directional. By studying this network, we can predict the known function of a protein, choose the experimental candidate, and understand the design principle of biological system [SWL+05].

In recent years, the development of high-throughput experimental techniques has generated substantial amounts of protein-protein interaction data deposited in several protein-protein interaction databases. However, protein-protein interaction is an ambiguous phrase which might infer direct physical contact or the association through some mediators between two proteins. Different experimental techniques can detect different kinds of “interactions”. There is the need to distinguish the protein-protein interactions based on different experimental strategies, such as yeast-2-hybrid, pull down assays or in vivo fluorescence energy transfer. Besides the experimentally derived data, predicted protein interactions are included in some protein-protein databases as well.

According to the approaches in the collection and presentation of interaction data, the protein-protein interaction databases could be categorized into three groups: (i) primary databases, which include experimentally proven protein interactions coming from either small-scale or large-scale published studies that have been manually curated; (ii) meta-databases, which include only experimentally proven PPIs obtained by consistent integration of several primary databases; (iii) prediction databases, which include mainly predicted PPIs derived using different approaches, combined with experimentally proven PPIs [DF10]. Some of the main primary

databases are introduced here.

IntAct

IntAct [AAA+10][KAF+07] is an open data molecular interaction database. Its main focus is protein-protein interaction data but also captures the interactions for DNA, RNA, and small molecule. The stored data are manually annotated by domain experts from published literature. The annotation of interactor is mapped to identifiers in UniProtKB, ChEBI, Ensembl, and the DDBJ/EMBL/GenBank. The binding sites are also cross-referenced to the InterPro database. It uses the controlled vocabulary developed by the Molecular Interaction group of Proteomics Standard Initiative (PSI-MI) [OH07] and Gene Ontology in its annotation. As of September 2011, it hosts over 275,000 curated binary interactions from 5000 publications [KAB+12]. IntAct can be accessed through web page and web service. The downloadable file is in PSI-MI or PSI-MITAB formats.

MINT

MINT [CAL+10] is a public repository for molecular interactions reported in peer-reviewed journals. All interactions are manually curated by professional curators. Each interaction record has the annotation of used detection method and the type of interaction (direct, association, co-localization, enzymatic reaction), and the interactors are cross-referenced to UniProtKB and RefSeq. It also developed a scoring system to evaluate the confidence of “direct physical interaction” between protein pair. These annotation features might be helpful to experimentalist while picking their candidates.

MINT has adopted the PSI-MI standards for the annotation and for the representation of molecular interactions. MINT is accessible through its web interface, web service, and the downloadable dataset in PSI-MI or tab-delimited format.

Human Protein Reference Database

Human Protein Reference Database (HPRD) [PGK+09] is set up as a resource for experimentally derived information about human proteome including protein-protein interactions, post-translational modifications and tissue expression. HPRD manually derives its content by a critical survey of published literature by expert biologists and through bioinfomatics analyses of the protein sequence. The extensive information contained in HPRD includes protein isoforms, domain architecture, protein functions, protein-protein interactions, post-translational modifications, enzyme-substrate relationships, subcellular localization, tissue expression, disease association of genes

[PNK+04]. PPIs are among the components of HPRD requested most by the users [PGK+09]. HPRD has the highest coverage of reported human PPIs among 6 primary PPI databases in 2009 (the other 5 PPI databases are: DIP, IntAct, MINT, BIND and BioGRID) [DF10]. HPRD data are available for download in XML as well as tab-delimited file formats.

2.1.1.3 Metabolic network and its resources

A living body is like a complex chemical factory. The cooperation of thousands of reactions constitutes the living phenomenon–growth, development, movement, response to stimuli, etc. The product from one reaction might be the substrate of succeeding reactions. A series of reactions carrying out a specific function is organized into pathways, and all the pathways of a living organism are integrated into the metabolic network. Metabolic network depicts all the biochemical reactions carried out in a living organism. These chemical reactions usually transform the substrates to products by a catalytic enzyme.

Metabolic pathway databases store the computationally accessible metabolic pathways which are dispersed in the textbook, scientific publication, and human mind. These databases improve the utility and circulation of metabolic pathway knowledge with the public access through internet. The construction of metabolic pathway databases starts with the collection of metabolic knowledge from the literature. After the completion of several genome projects and the improvement of sequencing technique, biochemical reactions of one species can be inferred from another species through ortholog identification. Like the protein-protein interaction data, metabolic pathway data also have variable confidence [RTD+10][DBJ+07] according to the construction methodology.

KEGG

KEGG is a database that integrates existing data from metabolites, enzymes, reactions, transcripts, and genes to facilitate data mining of biological information [Go10]. KEGG comprises 19 databases categorized into three categories: systems information, genomic information and chemical information [KAG+08]. The main database in systems information category is the KEGG PATHWAY, which contains organism-specific molecular interaction network in cells, such as pathway maps for metabolic, regulatory, signal transduction, cellular processes and human diseases [Go10] [KAG+08]. The genomics information category has two important databases: KEGG ORTHOLOGY and KEGG GENES. KEGG ORTHOLOGY is a database for pathway-based classification of orthologous genes, including orthologous

relationships of paralogous gene groups [KGH+06]. And KEGG GENES is a collection of genes compiled for all organisms with completed and partially sequenced genome from publicly available resources [Go10]. The chemical information category contains information on metabolites, drug molecules, glycans, and reactions, which are stored in the following six databases: ENZYME, COMPOUND, REACTION, GLYCAN, RPAIR, and DRUG [Go10]. KEGG offers several ways to access the database content through the KEGG website, a download at KEGG FTP, and web service.

Reactome

The aim of Reactome is to provide an integrated and qualitative views of human biological processes in a computationally accessible form [VDS+07]. It starts from the reaction level and organizes subsequent reactions into pathways. On the reaction level, the biological process is described as state transition of biomolecules so that the complexity of the many transformations in molecules, such as phosphorylation, transport, and isomerization, can be described in a computable form. The protein and small molecule entities are cross-referenced to popular external databases, such as UniProt and ChEBI, respectively. Every reaction entered in Reactome is backed up by evidence from biomedical literature. Two types of evidence are considered for incorporating data into Reactome - direct evidence, which comes from an experimental assay on human cells, and indirect evidence, which uses sequence similarities to infer from other species on human pathway [VDS+07]. All information in Reactome is expert-curated by PhD-level biologists. In addition to the human biological processes, Reactome holds information on the biological processes of non-human species which are computationally inferred from peer-reviewed curated human reactions. Reactome is accessible through its web interface, web services, or downloadable MySQL database dump.

MetaCyc and its derivatives

The MetaCyc database (MetaCyc.org) is a comprehensive and freely accessible resource for metabolic pathways and enzymes from all domains of life. With 1747 pathways from more than 2170 different organisms, MetaCyc is the largest collection of metabolic pathways currently available. Pathway reactions are linked to one or more well-characterized enzymes, and both pathways and enzymes are annotated with reviews, evidence codes, and literature citations [CAD+09].

The highly curated MetaCyc can serve as a reference for metabolic pathway construction for other organisms with an annotated genome. With the utilization of

MetaCyc, the annotated genomes, and the PathoLogic component of the Pathway Tools software [KPK+10], several computationally predicted metabolic networks are constructed and comprises the BioCyc database. BioCyc is a collection of more than 500 organism-specific pathway/genome databases for the sequenced and annotated genomes [CAD+09]. The computationally predicted metabolic networks utilizing the annotation of human genome from Ensembl, LocusLink, and GenBank is termed HumanCyc [RWG+04]. HumanCyc has been moderately curated by human review and literature-based curation. All data in MetaCyc, BioCyc, and HumanCyc is freely downloadable in standard file exchange formats (BioPAX, SBML) and other text-based formats (tabular, attribute-value). Programmatically accessing these databases is through web service, MySQL direct access, and APIs for Java and Perl programming languages.

Edinburgh human metabolic network (EHMN)

EHMN is a high-quality human metabolic network manually reconstructed by integrating genome annotation information from different databases (KEGG, UniPort, HGNC, EntrezGene, Ensembl, Genecard) and metabolic reaction information from literature. EHMN reorganized the metabolic pathways in order to reduce the reaction overlapping between pathways and integrate functionally related small pathways. After the reorganization, EHMN included 2823 reactions in 66 pathways [MSM+07]. The updated version integrated sub-cellular location information and added transport reactions [HMZG10].

2.1.2 Integrated data repository

The number of biological databases has experienced an explosive growth following the fast generation of biological data. Each biological database contains different subset of biological knowledge which fits the developer’s interest. When questions that span several domains are asked, researchers have to transverse different databases to assemble the knowledge [Stei03]. It’s also common that the relationship between different biological domains is unclear due to the incomplete or missing links. In addition, the different information system and software solution of the dispersed databases are incompatible to each other, which is problematic for programmatic employment [HKT+10]. Database integration aims to provide a single portal for accessing originally dispersed data and a unified interface for programmatic employment.

In the area of inter-disciplinary study, such as systems biology, several integrated databases are present. The following paragraphs introduce two integrated biological

databases, which have the interface for programmatic access or provide data dump and are still accessible (Dec 2011), and one toolkit which facilitates the construction of local data warehouse. These resources allow the easy data integration into other systems.

DAWIS-M.D.

DAWIS-M.D. [HKT+10] integrates 11 databases–BRENDA[CSG+09], EMBL [KAA+06], HPRD [PGK+09], KEGG [KAG+08], OMIM [HSA+05], SCOP [MBHC95], TRANSFAC® [WCH+00], TRANSPATH® [KPV+06], ENZYME [Bai00], GO [TGOC00] and UniProt [ABW+04]–into an unified relational database model and provides the access through the web page, web service, and a network editor–VANESA [JKT+10]. User can inquire the relationships and interactions between 12 biomedical domains–compound, disease, drug, transcription factor, enzyme, gene, glycan, gene ontology, pathway, protein, reaction and reaction pair domain. Apart from aforementioned 11 databases, DAWIS-M.D. is working on the inclusion of protein-protein interaction data from IntAct and MINT. Although the unofficial inclusion of these two databases in the current DAWIS-M.D., IntAct and MINT already can be queried through VENESA.

BNDB

Biochemical Network Database (BNDB) is a data warehouse hosting the data in the domains of sequence of biological object (SwissProt[WAB+06], RefSeq [PTM05], InterPro [MAA+05]), pathway (KEGG [KGH+06], BioCyc [KZM+04], TRANSPATH® [KPV+06]), molecule interaction (DIP [SMS+04], MINT [ZMQ+02], IntAct [HML+04], HPRD [PNA+03], BIND [AAA+05], BioGRID [SBR+06], TRANSFAC® [MKF+06]), and gene expression (GEO [BTW+07]). BNDB is implemented as a relational database using MySQL. Users can access BNDB through the web interface or the standalone Java-based viewer. A programming interface called Biochemical network library BN++ is designed for the C++ language [KBB+07][Url17].

BioWarehouse

BioWarehouse [LPW+06] is a toolkit for constructing bioinformatics database warehouse in user's local MySQL or Oracle relational database managers. It differs from the two above-mentioned data warehouses in that it provides only the framework instead of the data content of a data warehouse. The BioWarehouse toolkit is a collection of programs for creating the BioWarehouse schema and for loading data into BioWarehouse. The program for data loading is termed "loader". The

BioWarehouse toolkit has implemented several loader programs each of which parses the flat file of a source database and inserts the data into the data warehouse. BioWarehouse contains loaders for the following source database and data formats [Url18]: BioCyc [KOM+05], BioPax [DCP+10], CMR [PUD+01], eco2dbase [VSC+92], Enzyme DB [Bai00], GenBank [BKL+03], Gene Ontology [GOC12], KEGG, MAGE (MicroArray Gene Expression), MetaCyc Ontology, NCBI Taxonomy [WCL+00], and UniProt.

2.2 Web-based access

Since the life science databases and bioinformatics tools are globally decentralized, the interoperability is one of the concerns when those services are utilized integratively. To allow automated access to these bioinformatics services by software program, the Application Programming Interface (API) has to be provided for each service. Owing to the easy access and prevalence of internet, numerous bioinformatics services are accessible through the web. Besides the web-page access for the human researcher, the web service technology is a popular choice for the web-based programmatic access [SAC+08][KAN+10].

There are two web service models that are chosen by the bioinformatics community. One model is the traditional SOAP (Simple Object Access Protocol)-based web service which encapsulates the client request and the server response in XML following SOAP specification and is transmitted through Hypertext Transfer Protocol (HTTP). The SOAP specification allows the client program to include complex parameters in the request to the service. The use of HTTP protocol allows the SOAP message unrestricted by firewalls. The SOAP-based web service is a suitable API for an analysis service which usually requires complex input with multiple numbers of parameters.

The other model uses Representational State Transfer (REST) by which the required operation and its parameters are encoded as standard HTTP GET or POST request [SAC+08]. The REST-based service is typically rather easy to use and suitable for data-retrieval service [KNT10] from the database.

2.3 Standards for data exchange

Biological data are generated from different sources. The data might be retrieved from the biological databases, the output file of a program, or in any arbitrary format that is defined by any individual. The various formats have made the interpretation of data difficult and error-prone. A unified data exchange format upon the agreement of

research community is therefore necessary.

The eXtensible Markup Language (XML) is a popular form chosen by several standard data exchange formats over the other formats, such as flat file, Abstract Syntax Notation One (ASN.1), and the Common Object Request Broker Architecture (CORBA) [AVB01].

The standard data exchange format is usually designed to be domain-specific due to different data requirement in different research fields. The following section is devoted to the introduction of some standard data exchange formats for experimental raw data and biological model.

2.3.1 Biological experiment raw data exchange format

The omics-wide techniques can generate huge amounts of data in a single experiment. Besides the measurement acquired from the experiment, metadata of the experiment, such as experimental conditions, is needed to allow the faithful reproduction of the experimental result by others. Several textual guidance has suggested the minimum information to be revealed about the microarray, proteomics, and molecular interaction experiments: MIAME (minimum information about a microarray experiment), MIAPE (minimum information about a proteomics experiment), MIMIx (minimum information required for reporting a molecular interaction experiment) [OH07].

Complying with the minimum information requirement, several data exchange formats have been proposed to assist the digital dissemination of the omics raw data: MAGE-ML [SMS+02] for the microarray experiment, GelML [GHM+10] for the gel-based proteomics experiment, mzML [MCS+11] for the MS (mass spectrometry)-based proteomics experiment, PSI-MI for molecular interaction experiment [OH07].

2.3.2 Biological model exchange format

2.3.2.1 SBML

Systems Biology Markup Language (SBML) is the effort from a broad community and orientated towards describing biological processes including metabolic pathways, cell signaling pathways, and many others. SBML is continuously evolved, and a version tailing is usually specified when SBML is referred. The versioning system of SBML comprises a *level* number representing substantial changes and a *version* number representing the minor revision of the preceding level. The current release

(Oct 2010) is Level 3 Version 1. A SBML document starts with a model declaration and contains the following components: function definitions, unit definitions, compartments, species, parameters, initial assignments, rules, constraints, reactions, and events. The reaction component is where the transformation, transportation and binding process of biological entities are specified. This component also contains the information of reaction reversibility and mathematical formula describing the rate of the reaction [HBH+10][Url3]. Since its first release in 2001, SBML has been supported by around 230 software programs and is the most successful standard model exchange format [LDR+10].

2.3.2.2 CSML

CSML (Cell System Markup Language) had its first release (CSML1.0) in 2005 and supports the Petri net based network representation. The latest release (CSML3.0) has made CSML an integrated data exchange format which covers widely used data formats and applications, e.g. CellML1.1, SBML Level2, BioPAX, and Cytoscape [NSJ+10]. CSML3.0 supports HFPNe (Hybrid Functional Petri net with extension) [NDMM04] architecture where generic entity and process are introduced to handle any type of objects and relations. The inclusion of biological meanings, simulation model, and layout information makes CSML a language with high expressive power. The tools assisting the transformation of other standard exchange formats, such as SBML and CellML, to CSML are available. Therefore the models deposited in other repositories in SBML and CellML formats could be easily transformed to CSML [NSJ+10][Url7].

2.3.2.3 BioPAX

BioPAX is a community standard for pathway data sharing. The goal of the BioPAX project is to provide a data exchange format for pathway data and support the data models from a wide range of popular pathway databases. The BioPAX ontology could represent the data models from a number of existing pathway databases, such as BioCyc, BIND [BBH03], PATIKA [DBD+02], Reactome, aMAZE [LAC+04], KEGG, INOH [YSN+11], NCI/Nature PID [SAK+09], and PANTHER pathways [MDM+10]. BioPAX supports the description of several types of pathways which are typically represented in databases, literatures and textbooks: metabolic and signaling pathway, gene regulatory networks and genetic and molecular interactions.

The BioPAX language uses a discrete representation of biological pathway. Dynamic and quantitative aspects of biological processes are addressed through the coordination with other mathematical modeling language communities. The visual

properties of the pathway is not included in BioPAX but through the coordination with the System Biology Graphical Notation (SBGN) community.

The BioPAX community has developed the Paxtools Java programming library to support the import, export and validation of BioPAX-formatted data. BioPAX has been supported by several databases and software [DCP+10][Url5].

2.3.2.4 CellML

The specification of CellML was motivated by The International Union of Physiological Sciences Physiome Project, which was aimed to provide a framework for the modeling of the human body. CellML differs from other biological markup languages in scopes of biological systems it covers. The flexibility of the CellML enables the description of electrophysiological and mechanical models as well as biochemical pathway models. In the current release of CellML 1.1, each model encoded in CellML has a unique identifier. When the model identifier is used together with the model's Uniform Resource Locator, each model can be referred by a model's Universal Resource Identifier. The Universal Resource Identifier of model allows the reuse of previously published models in a new model.

CellML model repository (http://www.cellml.org/models/) has stored more than 500 models including electrophysiological and mechanical models, as well as biochemical pathways models. CellML is currently supported by numerous tools and techniques for editing (both visual and textual), validation, sharing and curation (through an online repository), generation of code (for external use), and execution of CellML models [GNC+08][CLN+03].

2.4 Domain-specific ontologies and thesaurus

Ontology is the specifications of the entities, their attributes and relationships among the entities in a domain of discourse. Sometimes it is also referred by the term "thesaurus" or "controlled vocabulary" [RSN07]. Because the ontology defines a common vocabulary in a domain, researchers can use the same term when a specific entity is referred. The common vocabulary allows the unbiased transmission of knowledge among the domain research groups. A large ontology can simply reuse several existing ontologies describing portions of the large domain [NM01].

2.4.1 Gene ontology

Gene Ontology (GO) [GOC12] provides a unified terminology describing the biological processes, molecular functions, and cellular components of genes across

species. The terms in GO are organized in a hierarchical structure, in which the children terms is a specialization of parent term via is-a relations. Each term in the hierarchy is assigned a unique zero-padded seven digit identifier (often called the term accession or term accession number) prefixed by GO. The textual definition and the references of such information are provided for each term [Url4]. GO has been adopted by different model organism databases to unambiguously describe the biological processes, molecular functions, and cellular components of gene products. Besides the database annotation, analysis of GO codes is a common analysis process for high throughput experimental data [RSN07].

2.4.2 MeSH

The resource Medical Subject Headings (MeSH) provides a controlled vocabulary developed by the National Library of Medicine (NLM) to index, categorize, and search the data collection in NLM. MeSH terms are organized in a hierarchical structure where the broader headings are at the more general level. The top level of MeSH has 16 categories: anatomy; organisms; diseases; chemicals and drugs; analytical, diagnostic and therapeutic techniques and equipment; psychiatry and psychology; phenomena and processes; disciplines and occupations; anthropology, education, sociology and social phenomena; technology, industry, agriculture; humanities; information science; named groups; health care; publication characteristics; geographicals. Each of these categories is further divided into subcategories.

Each entity in the MeSH hierarchy is termed descriptor, which is used to index citations in NLM's MEDLINE database. Each descriptor is affiliated with several entry terms, which are synonyms, alternate forms, and other closely related terms to the descriptor. MeSH contains 26142 descriptors and over 177000 entry terms in 2011. MeSH is continuously updated according to the emergence of new terms and new research area [Url2].

2.4.3 Open biomedical ontology foundry

Since the utilization of ontology is beneficial in information integration and computer reasoning with data [RSN07], several bodies have devoted to the development of ontology with the application on specific domains. The various ontologies developed by different bodies sometimes have their scale of subject overlapping with each other. One instance of such problem is the ontology about cell type, which had ever had at least four versions from different research groups. Open Biomedical Ontologies (OBO) foundry is an umbrella body for the coordinated development of life-science

ontologies. After the reformation by OBO Foundry, three of the previous uncoordinated cell type ontologies have been merged into a single cell type ontology, which is reused by the fourth ontology. OBO now comprises over 90 ontologies for the knowledge domains in anatomy, phenotype, biochemistry, genomics, molecular biology, taxonomy, medicine, etc [SAR+07] [Url6].

2.5 ID mapping

The current biological knowledge is dispersed in many independent molecular biology databases. If the record entry in a database needs to be referenced, the unique identifier (UID) of the database is always an unambiguous indicator of the data entry other than some text-based descriptor, such as protein/gene name or gene symbol. However, different molecule database has adopted its own unique identifier referring to the same bioentity: HPRD uses HPRD ID, and UniProt uses UniProt ID for protein reference.

For each biological molecule, there are several databases each of which is devoted to certain properties of that molecule: the structure information of protein is stored in PDB, and the interacting protein counterpart is stored in IntAct, HPRD, and others. Moreover the same biological molecule is referred by different UID from varied databases. The ID mapping tool is necessary when diverse datasets need to be compared and integrated. Several ID mapping tools have been developed and dedicated to either gene-centric or protein-centric view [HMS+11]. Here are two examples of ID mapping tools for each view (Table 2.1).

iProClass [WHN+04][HMS+11] is a protein-centered database providing links to over 50 databases of protein family, function, pathway, interaction, modification, structure, genome, ontology, literature, and taxonomy. A database ID mapping service is implemented by the iProClass. It uses SwissProt accession number (SPACC) and identifier (SPID) as the entity identifier for each protein. Through querying iProClass by SwissProt accession number/ID, the IDs or accession numbers of the other databases can be easily retrieved for the queried protein. iProClass is a useful resource for easy ID mapping, such as finding the related PDB ID for a protein, especially when the SwissProt accession number/ID is chosen as the protein entity identifier in the user's application. iProClass is freely accessible from the PIR web site, and the whole data set is downloadable from its FTP site in XML and tab-delimited formats. Besides, a REST-based web service for programmatic access has newly been implemented [HMS+11].

			iProClass ID mapping	DAVID ID conversion
ID type	***Category*:Biological object**	Protein sequence	UniProtKB, UniRef100, UniParc, RefSeq, GePept, NR	UniProtKB, RefSeq_PROTEIN, UniRef100, PIR_ID, PIR_ACCESSION, GENPEPT_ACCESSION, PIR_NREF_ID*
		Gene and genome	GenBank/EMBL/DDBJ, UniGene, FlyBase, MGD, SGD, WormBase, TAIR, TIGR	ENSEMBL_ID, ENTREZ_GENE_ID, GenBank, GENE_SYMBOL, UniGene, RefSeq_GENOMIC, RefSeq_MRNA, RefSeq_RNA
		Proteomic peptide ID databases	GPMDB, PRIDE, PeptideAtlas	
	***Category*:Annotation of biological object**	Taxonomy	NCBI taxon, NEWT	
		Gene expression	GEO, CleanEx, SOURCE	
		Protein expression	Swiss-2DPAGE, PMG	
		Function and pathway	EC-IUBMB, KEGG, BioCyc	
		Genetic variation and disease	OMIM, HapMap	
		Ontology	GO	
		Interaction	IntAct, DIP	
		Modification	RESID, phosphosite	
		Structure	PDB, SCOP, CATH, MMDB, PDBsum, ModeBase	
		Classification	PIRSF, UniRef50, UniRef90, Pfam, InterPro, PANTHER, COG, SMART, TIGRFAMs	UniRef100
		Microarray		AFFY_ID
Mixed type search			no	yes
Access method			web page, FTP, REST	web page, FTP
Usability for work			proteomics	genomics

Table 2.1 Protein-centric and gene-centric ID mapping tools (*: obsolete)

In contrast to iProClass ID mapping, DAVID gene ID conversion tool [HSS+08] is a gene-centric ID mapping service, which is more suitable for the genomic- and microarray-related work. It integrates around 20 ID types and allows mixture of ID types in one single query. Users can access DAVID ID conversion tool by its web interface or download the entire data file.

2.6 Integration platform

The various scattered biological databases and bioinformatics tools make the integrative analysis of heterogeneous data difficult. Researchers have to spend time on shuttling between databases, manipulating data format, formatting the query, extracting intended information from the result file, and comparing the data sets from different resources. An integration platform which provides a consistent data structure, friendly user interface, and function extensibility would ease the integration process.

For the integrative analysis of cellular systems, an integration platform would be valuable due to the scale and types of data to be analyzed, such as the ones from the high-throughput omics techniques and enzyme kinetics parameters. Several popular integrative platforms have been developed and are supported in a community effort. Common features of these integration platforms include the network-based graphical representation, data import from standard data exchange formats, and database integration which simplifies the data retrieval from different databases. Some of them also allow the network editing, experimental data mapping, network topology property analysis, simulation, and the function extension via the plug-in architecture (Table 2.2).

2.6.1 Cytoscape

Cytoscape [SOR+11] is an open source platform for network analysis and visualization and is supported by the development teams from private and public sectors. It was initially designed for biological research but can be applied to analyze the network in the other areas. The core distribution of Cytoscape provides the function of data integration and visualization. It accepts some standard file exchange formats, such as GML and SBML. The graphical properties of node and edge, like node shape, edge thickness, node color, arrow shape, can easily be changed by the user.

The most attractive feature of Cytoscape is its extensibility. It has a well-documented API which allows the development of Cytoscape plugin by anyone. With the support of various plugin, the function of Cytoscape is largely extended. These additional functions cover: network query and download services, network data integration and filtering, attribute-directed network layout, gene ontology enrichment analysis, detection of network motifs, functional module, protein complex, and domain interaction [CSC+07]. The numbers of published Cytoscape plugin in the year of 2007, 2008, 2009, 2010 are 9, 29, 30, and 55, respectively (statistics from the Cytoscape Plugin page in Feb. 2011).

2.6.2 CellDesigner

CellDesigner [FMJ+08] is a model editor and simulator of gene regulatory and biochemical networks. Besides editing and modeling, it has an ODE (ordinary differential equation)-based simulator for symbolic and numerical analysis of chemical reaction networks. It supports input/output in SBML and BioPax formats and represents biological network as process diagram where different states of a molecule are drawn as separate nodes, and the state transition is drawn as arrow. The process diagram will make the network interpretation semantically and visually unambiguous. Since its release as version 4.0 in 2008, CellDesigner also developed a plugin framework. The network data could be obtained from the connected databases: BioModels, PANTHER pathway database, MetaCyc, or may be drawn by users.

2.6.3 VisANT

The special feature of VisANT [HNY+07] is its support of metagraph development. Metagraph allows the implementation of metanodes which contain subnetworks inside the node. The content of each metanode can be contracted or expanded. The network is created by file importing or retrieved from KEGG or Predictome databases [MYC+02]. Predictome was developed by the VisANT team and contains the predicted protein association of 44 genomes bases on three computational methods and large-scale protein-protein interaction data. The nodes and edges in the loaded network can be filtered and removed but not added. User can map gene expression data onto the network in VisANT. The portable file format for the input is KGML, and the output format can be VisML, tab-delimited, or SVG image.

2.6.4 Cell Illustrator

Cell Illustrator is a biopathway modeling and simulation platform which uses hybrid functional Petri net with extension (HFPNe) as the mathematical model, which enables the simulation in discrete or continuous modes. The full function of Cell Illustrator requires a commercial license. It supports the file import and export from SBML, CellML, and BioPAX formats. The CSML models deposited in the CSML pathway library could be directly loaded into Cell Illustrator [NSJ+10].

	VisANT	CellDesigner	Cell Illustrator online	Cytoscape
Version	3.91	4.2	5	2.8
Backend database or connectivity to public databases	Predictome(experimental and predicted data by integrating IntAct, HPRD, BioGrid,MINT,BIND,MIPS), KEGG, GenBank, SwissProt	PANTHER Pathways, BioModels, SABIO-RK, SGD, iHOP, DBGET, PubMed, Entrez Gene, UniProt, MetaCyc, Gene Wiki	CSMLDB	Pathway Commons
Standard data exchange format compliance	PSI-MI, GML	SBML, BioPAX	CSML, SBML, CellML, BioPAX	SBML, BioPAX, PSI-MI
Other supported data format	KGML, visML, tab-delimited, SVG		SVG	SIF, NNF, GML, XGMML
Plugin architecture	Y	Y	N	Y
Graphical editing	N	Y	Y	Y
Simulation	N	ODE	HFPNe	Plugin-dependent
Experimental data mapping	Y	N	N	Plugin-dependent
Run mode	On-line java applet, stand-alone application	Stand-alone application	Java Web Start	Stand-alone application
License	Free	Free	Commercial	Free

Table 2.2 Overview of integration platforms

2.7 Summary

This chapter presents various bioinformatics resources and tools for cellular biology and classifies them according to the functional hierarchy of cellular system. The capability of different resource complements each other, and the integrative utilization of bioinformatics resources is intended in the systematic study. This chapter also presents the bioinformatics facilities which aid the integrated study in cellular biology. These facilities represent the biological knowledge repository and facilities which aid resource communication and mutual understanding, such as the ID mapping service, standard data exchange format, application programming interface, ontology and integration platform. These facilities are crucial when an integrative and systematic methodology covering multi-omics and dynamic data is adopted [NBG+06], and the tools in each category of facilities keeps evolving with time. Some of these facilities will be used in the work carried out in this dissertation.

Chapter 3

Related work

This chapter presents the related works for the two topics included in the thesis. The first part is devoted to the problem of TTG identification, and the second to the document retrieval system in the biomedical research domain.

3.1 Redox regulatory network and its target protein

Changes in redox balance and development of oxidative stress are associated with many cell functions and life processes including aging, diseases, loss of fitness, and yield [MHMF96] [BG91] [Ferr09] [Diet08]. On the molecular scale, oxidation will change the structure of biomolecules and often switches protein activity or causes protein malfunction. To keep the cellular environment in a proper redox state, cells contain several antioxidants, such as vitamin C, vitamin E, and ubiquinol and also a set of antioxidant enzymes [Diet03]. By decomposing reactive oxygen species (ROS) and reactive nitrogen species (RNS) these antioxidants constitute the first line of defense to avoid damage to macromolecules by uncontrolled oxidation. Once ROS or RNS escape from the first defense line, lipids, nucleic acids and also proteins may get oxidized. A major oxidation reaction of proteins is the reversible dithiol-disulfide transition. Cells have developed two rescue systems that involve thioredoxins and glutaredoxins [MBVR09] [HJB+05] to re-reduce the oxidized proteins and sustain the normal protein structure and function in an oxidizing environment (Fig. 3.1). Thus this system is suitable for reversible regulation of protein functions, e.g. enzyme activities.

The thioredoxin and the glutaredoxin disulfide reduction systems are two main modules within the redox regulatory network (RRN) and prevalent in almost all species. Within the RRN, involved proteins can be classified into several functional elements: redox input elements feed electrons into the network, such as NADPH and glutathione (GSH); redox transmitters distribute the electrons to redox target, such as thioredoxin (Trx) and glutaredoxin (Grx); redox sensors transmit information on the cellular concentration of ROS to the redox network, such as peroxiredoxin (Prx); redox target proteins are enzymes catalyzing metabolic reactions [Diet08]. All the involved proteins bear reversible oxidation prone cysteine residues which allow the protein to switch between its reduced, in which the protein is in dithiol state, and oxidized forms, in which the protein is in disulfide state. The reducing equivalents are transferred from NADPH to the redox target proteins through a set of specific proteins (Fig. 3.1). Thioredoxin and glutaredoxin are the electron donors and recover the activity for the oxidized target proteins.

The relationship between the redox regulatory network and metabolic network is depicted in Fig. 3.1. The RRN functions as the regulatory system of metabolic enzyme but also of transcription factors, translation machinery and many other cell processes [BB05]. In the redox regulatory network, only the redox state of involved proteins is concerned. The redox state of the target proteins of the redox regulatory network will then influence the mass flow of metabolites.

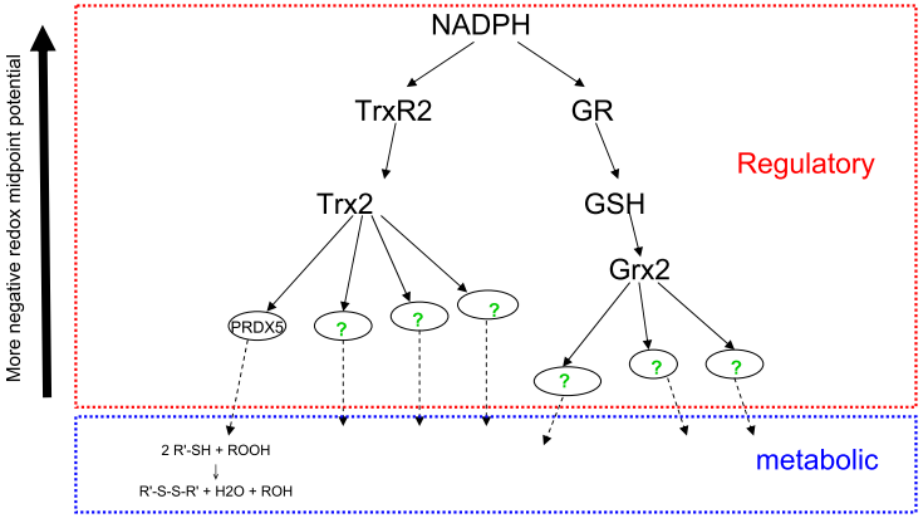


Figure 3.1 Redox regulatory network of the human mitochondrion and exemplarily affected metabolic network. Abbreviations: Trx2: thioredoxin-2, TrxR: thioredoxin reductase, GR: glutathione reductase, GSH: glutathione, Grx2: glutaredoxin-2.

3.1.1 Thioredoxins (Trx) and glutaredoxins (Grx)- as redox transmitters

Thioredoxin and glutaredoxin are two protein families both of which possess the Trx-fold, a partial structure of the whole thioredoxin composed of 3 α-helix and 3 β-sheets (N′- β_1- α_1 - β_2 - α_2 - β_3 - β_4- α_3 - C′). The proper members of both families have the CxxC motif (C: cysteine; x: any one of the 20 amino acids) at their active site with some exception with the CxxS (S: serine) motif. The active site CxxC motif is always located at the N-terminal side of the α_1 helix. In spite of the common fold and motif, Trxs and Grxs show little similarity in amino acid sequence [MBVR09] [CM10] (Fig 3.2).

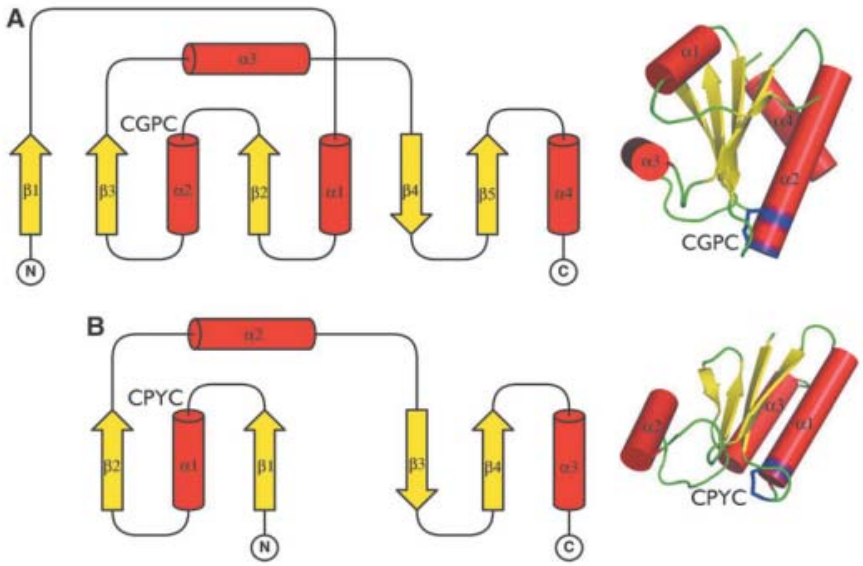


Figure 3.2 Structure dissection of Trx and definition of the Trx-fold. (A) The secondary structure of the thioredoxin protein. (B) The Trx-fold which is only the partial structure of Trx. The active site CxxC motif is located at the N-terminal of an α helix. (Source: [CM10])

Thioredoxin and glutaredoxin exist prevalently in all organisms. Prokaryotic thioredoxins show about 50% sequence homology. Thioredoxins in mammalian cells (rabbit and calf) are 90% similar and have 27% overall similarity to the *E. coli* protein. Glutaredoxins from *E. coli* and calf show about 30% identical residues [Hol89].

Organisms have developed their specialized subset of Trxs and Grxs, which are located in various cellular compartments. For instance, 19 different thioredoxins have been identified in the genome of *Arabidopsis thaliana* that can be grouped in 6 sub-families by their protein sequences. Each of the Trx sub-family has different distribution among organelles [BB05][MBVR09]. In addition, there are a large number of proteins

with Trx-like domains in plants. The high number of Grx genes is also seen in *Arabidopsis*. The Grxs are classified in three sub-groups according to protein sequence, and there is extensive heterogeneity within the sub-group in terms of protein size and subcellular location [MBVR09]. The human genome encodes two thioredoxins (Trx1, Trx2), two dicysteinic glutaredoxins (Grx1, Grx2), and several Grx2 variants. Trx1 and Grx1 are cytosolic proteins, and Trx2 and Grx2 locate in mitochondria.

Trx and Grx function as the redox transmitter in the cell. They are not engaged in metabolite turnover but couple redox input elements to the redox state of target proteins and thereby modify the activity of metabolic enzymes, although it should be noted that some Grxs are involved in deglutathionylation and FeS-cluster formation in plants [DP11].

3.1.2 Target protein of redox regulatory network

The proteins connecting the upstream regulatory network and the downstream metabolic network are the target proteins of Trx/Grx. The critical step in expanding the redox regulatory network is to identify the Trx-/Grx- target proteins in order to complete the network. The Trx-/Grx- target proteins are the ones which carry out the thiol-disulfide exchange reaction with Trx/Grx and therefore contain reversibly oxidized cysteines.

Proper Trx/Grx contain two redox-active cysteines which undergo the thiol-disulfide exchange with its target protein. The redox active cysteine can exist either in the oxidized or the reduced state according to the redox potential of the environment. When the cysteine is in its reduced state, the cysteine is in the free thiol form. Once the cysteine is oxidized, it forms a disulfide bond with another redox-active cysteine, and this disulfide bond can be reduced again by an enzyme (thioredoxin reductase (TR) or the tripeptide glutathione (GSH)). Owing to the two reversibly oxidized cysteines (ROCs), Trx/Grx carries out the role as the redox transmitter and the reductant for its target proteins.

When Trx/Grx reduce their target proteins which bear the disulfide bond, one of its redox-active cysteines (the one which is closer to the surface and to the N-terminus) attacks the disulfide bond on the target protein and forms an intermolecular disulfide bond with the target protein. This intermolecular disulfide bond is then attacked by the other redox-active cysteine, and an intramolecular disulfide bond is formed between two ROCs of Trx and Grx, and the release of the target protein is followed [KH80] (Fig. 3.3).

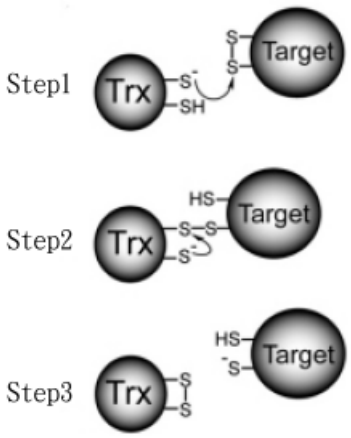


Figure 3.3 Thiol-disulfide exchange mechanism. The intermolecular disulfide bond is formed between Trx and its target in Step 2. [source: MHF+06]

The net reaction can be seen as the exchange of free thiols and disulfide bond between Trx/Grx and the target protein.

3.1.2.1 Target proteins from the literature

The chloroplast is the compartment in plant cells where photosynthesis and vigorous light-driven electron transfer take place. The function of Trx and its target proteins in chloroplast has been largely investigated owing to its redox regulation of photosynthetic enzyme activity. The putative and established Trx targets in chloroplast span in several functional categories: Calvin cycle, photosynthetic electron transfer, amino acid biosynthesis, etc [LMZ+07]. The mitochondrion represents another organelle in eukaryotic cells which is the site of respiration including the respiratory electron transfer chain. The pre-mature leak of electron from the electron transfer chains to oxygen generates superoxide which is an active oxidant and can react with lipids and proteins. The target proteins of Trx in the mitochondrion have been explored in plants. The identified target proteins of thioredoxin in plant mitochondria also function in several different cellular processes, such as photorespiration, citric acid cycle, lipid metabolism, and electron transport (Appendix A)[BVT+04].

Fu *et al*. has performed a proteomics identification of Trx1 reduction target proteins from the hearts of the transgenic mice which over-expressed Trx1 (cytosolic thioredoxin). They have identified 78 putative Trx1 reductive sites in 55 proteins which have diverse functions [FWL+09].

3.1.2.2 Interacting proteins from the databases

Several databases host the protein-protein interaction data generated by different technologies. These technologies can only detect the interacting protein pair mediated by non-covalent bond interaction, such as van der Waals', ionic forces, and hydrogen bond. However, the target proteins in the reduction cycle interact with Trx/Grx through intermediate covalent disulfide bond formation which is inaccessible by non-covalent bond-based technique of protein-protein interaction detection. Therefore the interacting counterparts found in most protein-protein interaction database are rarely real target proteins of Trx/Grx within the thiol-/disulfide redox regulatory network of the cell.

Two protein-protein interaction databases were searched for the interacting proteins of Trx1, Trx2, Grx1, Grx2 in human, mouse, and rat. The SwissProt accession numbers of these 12 proteins were used as the query for IntAct and BioGRID, and the returned IDs were mapped to SwissProt accession number through database cross-referencing. Table 3.1 lists the interacting proteins retrieved from the protein-protein interaction database. There were no or few interacting proteins found for Trx and Grx in most cases. Even if some interacting proteins can be retrieved from the protein-protein interaction database, the type of interaction remained unclear.

Besides exploring the individual biomolecule database, search in the integrative database is also conducted. The integrated database DAWIS-M.D. is queried through the network editor VENESA, and there is only one interacting protein retrieved for human glutaredoxin-2 from HPRD (Fig. 3.4). As for human thioredoxin-2, there is no interacting protein retrieved from either HPRD, or IntAct, or MINT.

3.1.2.3 Associated proteins from text-mining

The electronic availability of publications in bibliographic database, such as PubMed, enables the application of text mining and information extraction to biomedical literature. Text mining is the use of automated methods for exploiting the enormous amount of knowledge available in the biomedical literature [CH08]. Owing to the automated information extraction, text-mining tool can reveal the possible association between biological entities which is not collected in the structured database. ANDVisio is the visualization tool for ANDCell database which is constructed through computer linguistic text analysis on texts of scientific publication and database content [PYD+11]. If ANDVisio was used on the proposed biological question, namely the search for human mitochondrial Trx and Grx targets, 12 proteins were found to be associated with human

	IntAct	BioGRID
P10599 THIO_HUMAN	P30480 Q96BK5 Q92905 Q92905 Q53HC9 Q9HC24 Q9H3M7 Q9H3M7 P11171 P25942 Q86XK2 Q8IXH7 Q9Y4K3 O60739 P78330 P04406 O95990 P40337	P02452 P19883 Q15080 P08670 Q96BK5 Q9H3M7 P40337 Q99683 P25942 P19838 P04406 P04637 P04150
Q99757 THIOM_HUMAN	P40692	None
P35754 GLRX1_HUMAN	None	Q04656 P35670
Q9NS18 GLRX2_HUMAN	None	Q16881
P10639 THIO_MOUSE	None	None
P97493 THIOM_MOUSE	None	None
Q9QUH0 GLRX1_MOUSE	None	None
Q923X4 GLRX2_MOUSE	None	None
P11232 THIO_RAT	P19357	None
P97615 THIOM_RAT	None	None
Q9ESH6 GLRX1_RAT	None	None
Q6AXW1 GLRX2_RAT	None	None

Table 3.1 The interacting proteins for thioredoxins and glutaredoxins retrieved from IntAct and BioGrid

Trx-2 and 4 with human Grx-2 (Fig. 3.5). The relationships between human thioredoxin-2 and its associated proteins are annotated as activity down-regulation, activity regulation, association, and transport regulation and the ones between human Grx-2 and its associated proteins as association and transport regulation. A closer investigation of the identified proteins reveals that they do not necessarily reflect direct redox interactions but indirect relationships, e.g. the release of cytochrome c in cell death regulation in response to oxidative stimuli.

3.1.3 Strategies in thioredoxin/glutaredoxin target protein identification

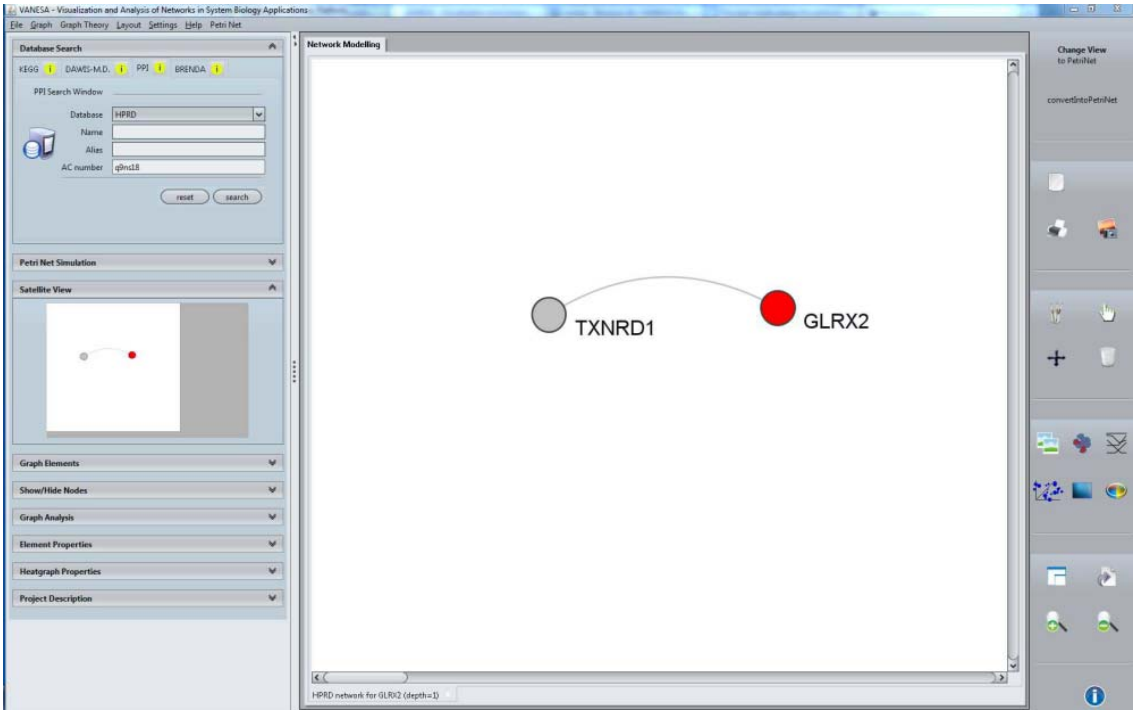


Figure 3.4 The interacting protein retrieved from DAWIS-M.D. for human glutaredoxin-2

The proteins connecting the upstream regulatory network and the downstream metabolic or other functional responses are the target proteins of Trx/Grx. The critical step in expanding the redox regulatory network is to identify the Trx/Grx target proteins in order to complete the network. The Trx/Grx target proteins are the ones which carry out the thiol-disulfide exchange reaction with Trx/Grx and therefore contain reversibly oxidized cysteines.

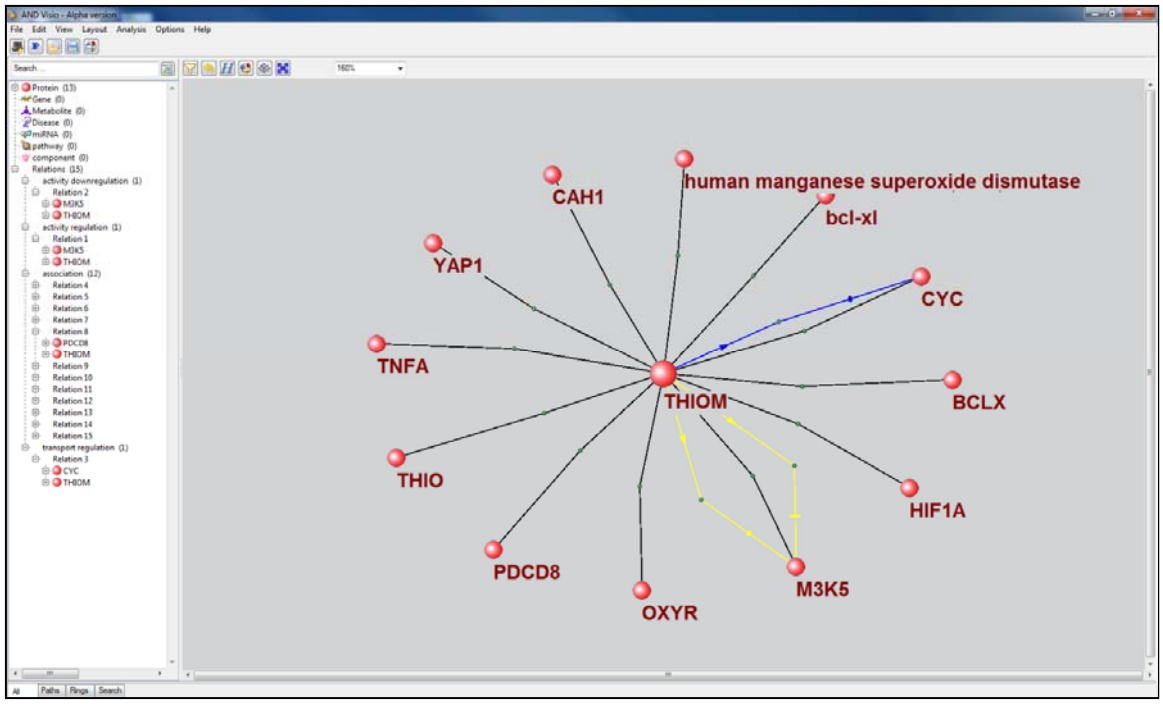
3.1.3.1 Experimental method

The experimental technique used to discover the Trx/Grx target protein from a pool of proteins is affinity chromatography, diagonal redox SDS polyacrylamide gel electrophoresis or other redox proteomics approaches.

3.1.3.1.1 Affinity chromatography

In this experiment the native Trx/Grx is replaced by a mutant protein whose second active cysteine is mutated to serine. This mutant is immobilized on the resin before the complex protein sample is applied to the column. The mixed disulfide bond between Trx/Grx and their target protein cannot be broken by the introduced serine which replaces the resolving cysteine. Following washing of the column to remove non-specifically bound protein, the covalently trapped proteins are eluted under reducing conditions. The

A



B

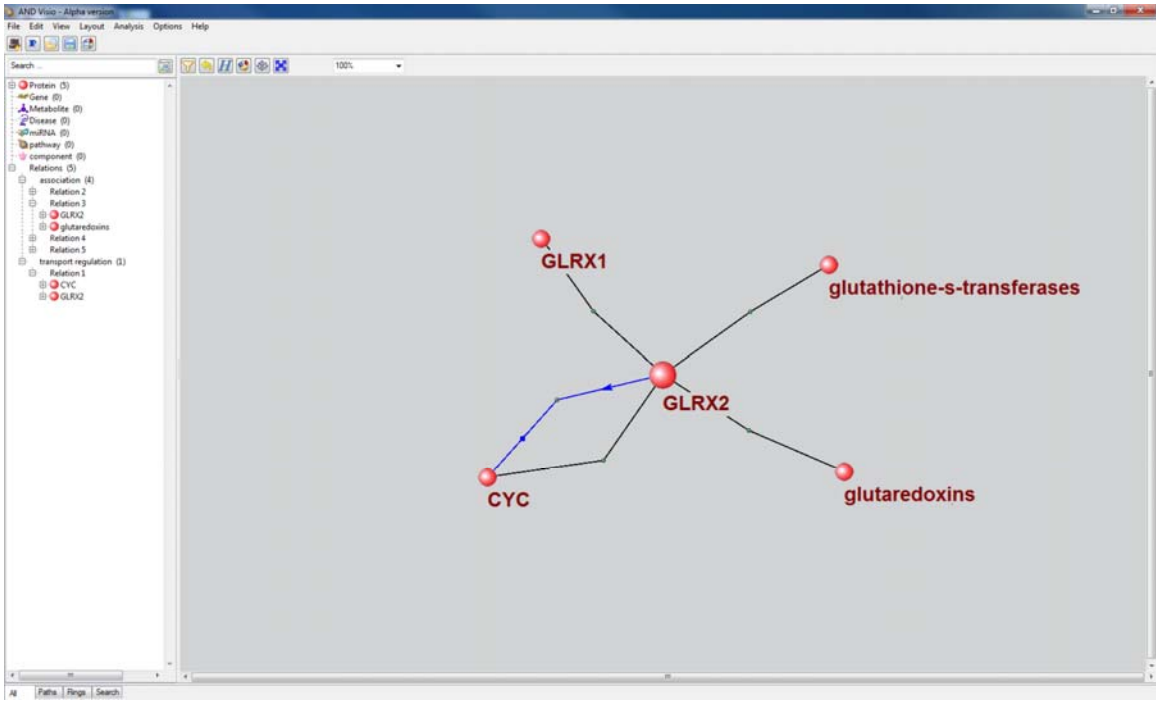


Figure 3.5 Associated proteins found by ANDVisio (**A**) for human mitochondrial Trx-2 and (**B**) for human mitochondrial Grx-2. The searched scope includes the literature and database content.

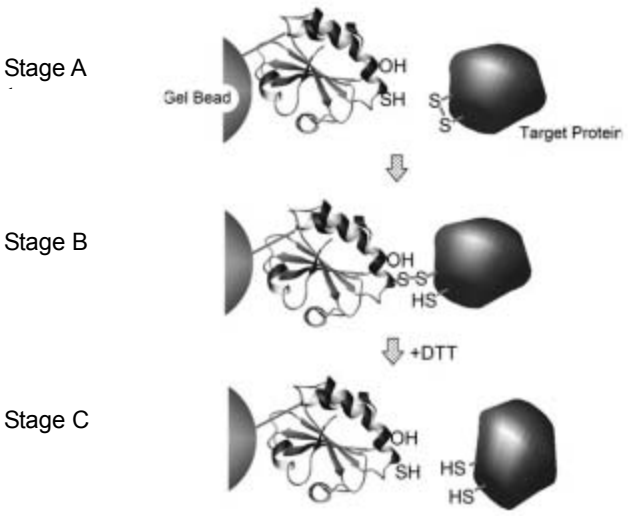


Figure 3.6 Thioredoxin affinity chromatography showing the steps of trapping and reductive elution. The mutated thioredoxin is immobilized on the resin in Stage A, and the intermolecular disulfide bond is formed between the target protein and immobilized thioredoxin in Stage B. The intermolecular disulfide bond is reduce by reductant (DTT), and the target protein is eluted in Stage C. (source:[HHF+05])

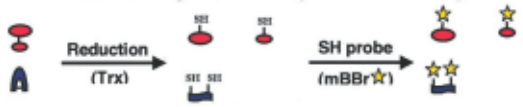
potential target proteins in the elution fractions are identified by mass spectrometry or gel-based methods (Fig. 3.6) [HHF+05].

3.1.3.1.2 Diagonal redox SDS or fluorescence-linked 2D polyacrylamide gel electrophoresis

Yano *et al*. [YWL+01] developed a method to identify Trx target proteins in peanut seeds based on two-dimensional gel electrophoresis. The sample of oxidized protein mixture is treated with specific redox system, either Trx or Grx, and the newly formed thiol residues after reduction are labeled with detectable marker. This marker could be the fluorescent dye (mBBr or cyanine) or radioactive molecule (^{14}C iodoacetamide). Trx/Grx-reduced targets strongly increase in fluorescence. An alternative method employs the often differential electrophoretic mobility of oxidized and reduced forms of the same protein [SD06]. After labeling, proteins are separated by first non-reducing and second reducing electrophoresis. The target proteins of specific redox system can be observed above or below on the diagonal by detecting the labeled marker (Fig. 3.7).

3.1.3.2 Computational method

A Reduction by thioredoxin and labeling of recovered thiol



B Protein separation by non-reducing and reducing electrophoresis

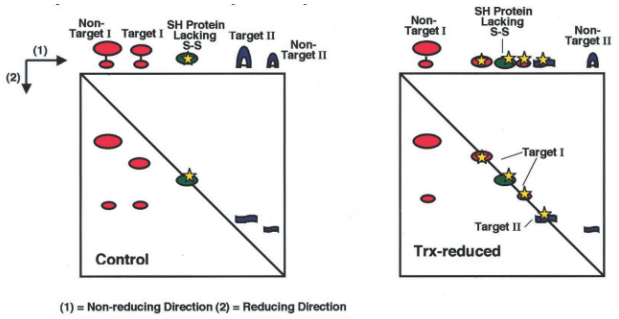


Figure 3.7 Diagonal redox SDS polyacrylamide gel electrophoresis (source: [YWL+01])

Due to the important role in modulating and regulating protein activity, computational prediction of redox-active cysteines is of significant interest. Most of the approaches for predicting redox-active cysteine thiols deal with the catalytic redox-active cysteines by analyzing the protein sequence, secondary structure, physiochemical properties, and database annotation.

Fomenko *et al.*[FXA+07] adopted the observation that selenocysteine is usually located in enzyme active sites and serves various redox functions and developed a procedure for high-throughput identification of catalytic redox-active Cys in proteins by searching for sporadic selenocysteine-Cys pairs in sequence databases. Marino and Gladyshev [MG09] have developed an integrative methodology in bioinformatics to detect thiol oxidoreductases and their catalytic redox-active cysteinyl residues. They tackled this problem by analyzing (i) the amino acid and secondary structure composition of the active site and its similarity to known active sites containing redox Cys, (ii) accessibility, reactivity, and active site location of tested cysteines. They applied this procedure to *Saccharomyces cerevisiae* proteins containing conserved Cys and identified the majority of known yeast thiol oxidoreductases.

A CxxC motif can be found in the active site of many thiol-disulfide oxidoreductases. Gopal *et al.*[GSZK09] used a pattern search of CxxC motif in the

bacterial *Listeria*-genome and allowed only its single appearance in the N-terminus of inspected protein. 29 candidate proteins were found. Followed by further reduction through 3D structure and phylogenetic analysis, almost half of the candidate proteins (14/29) had the structure or functional annotation related to known oxidoreductases or redox-regulated proteins.

Conour *et al*.[CGG04] first used redox related keywords like reduction, oxidation, redox, electron transfer, mental binding, heme, cysteine and disulfide to fish putative redox regulated protein motifs in the InterPro protein signature database. Then the motifs identified by keyword search were manually inspected by researchers.

The most-mentioned physicochemical parameters which relate to the redox capability of the cysteine are the vicinity of two cysteinyl residues, the acid dissociation constant (pK_a) value of the thiol, and the accessible surface area (ASA) of the cysteinyl residues on the inspected protein. Sanchez *et al.* [SRWM08] provided a discrimination rule with exact values of these three parameters–thiol-thiol distance, pK_a, accessible surface area–for predicting the oxidation susceptibility of the cysteine. The oxidation susceptibility of the cysteine indicates the possibility of cysteine oxidation, which is the prerequisite for the later re-reduction of ROC. Sanchez *et al*. used the data mining strategy to find three criteria which distinguish oxidation susceptible cysteine from non-oxidation susceptible ones. The training set was a balanced oxidation susceptible cysteine thiol database (BALOSCTdb) which collected 12 physicochemical parameters for 161 oxidation-susceptible cysteines (OSC) and 161 oxidation-non-susceptible ones (nOSC). The 12 physicochemical properties can be classified into the following categories: 1. distances between the target thiol and other atoms in its spatial neighborhood; 2. pK_a value of the target thiol and other amino acids in its spatial neighborhood; 3. ASA values of the target thiol and other amino acids in its spatial neighborhood; 4. electrostatic potential of target thiol; and 5. name of amino acids in its spatial neighborhood. pK_a is a measure of the strength of an acid in solution. The lower the pK_a value of a cysteine, the easier the cysteinyl thiol residue releases its H^+ cation, which is a step before disulfide bond formation. ASA (accessible surface area) is the surface area which is accessible to solvent for a biomolecule. Since the reagents of a chemical reaction are dissolved in solvent, it is critical for the reaction center to have minimum area exposed to the solvent. The calculated values of 12 physicochemical properties and the classification of the oxidation susceptibility for 161 OSCs and 161 nOSCs formed the training set for the decision tree learning by the C4.5 classifier. They concluded with three physiochemical

properties and their value ranges which help the identification of oxidation susceptible cysteines. The discrimination rule from their work is depicted in Fig. 3.8.

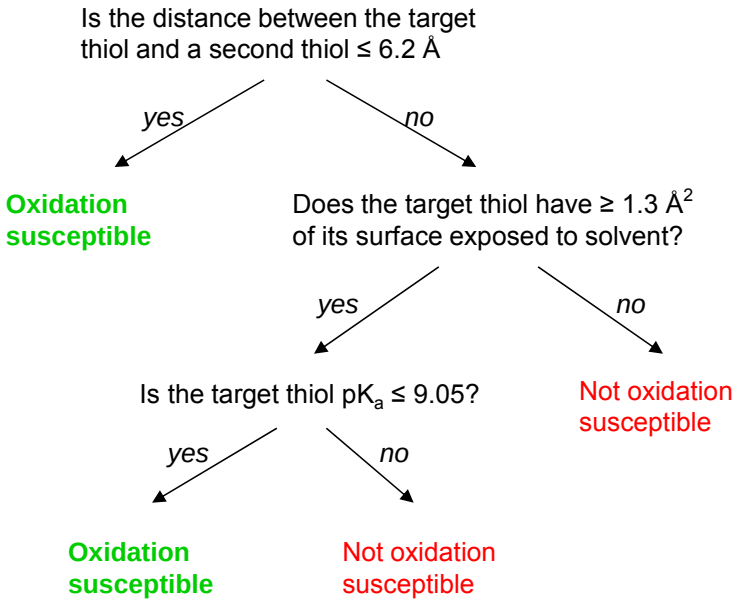


Figure 3.8 The decision tree to predict oxidation susceptible cysteine as developed by Sanchez *et al.* (modified from [SRWM08])

3.2 Document retrieval system for the biomedical research

Document retrieval, more commonly referred to as information retrieval, is the computerized process of producing a list of documents that are relevant to the inquirer's request by comparing the request to an automatically produced index of the textual content of documents in the system [Lidd05]. In the biomedical research field, the fast growing volume of scientific publications from thousands of journals makes it impossible for researchers to browse through all the publication when they search for relevant publication to their work. Thanks to the digitalization of the bibliographic information and the development of natural language processing, several document retrieval systems in the biomedical field have emerged.

Since the document retrieval system serves as the intermediate agent between user's query and the documents, the operation of the document retrieval system could be classified into three phases- document representation, query processing, and matching

[Lidd05] (Fig. 3.9). The document representation phase is to choose descriptive terms from a defined term set to represent each document, query processing to translate the free text query received from the user to the equivalent terms in the defined term set, matching to generate the document list by comparing the translated equivalent terms with the index.

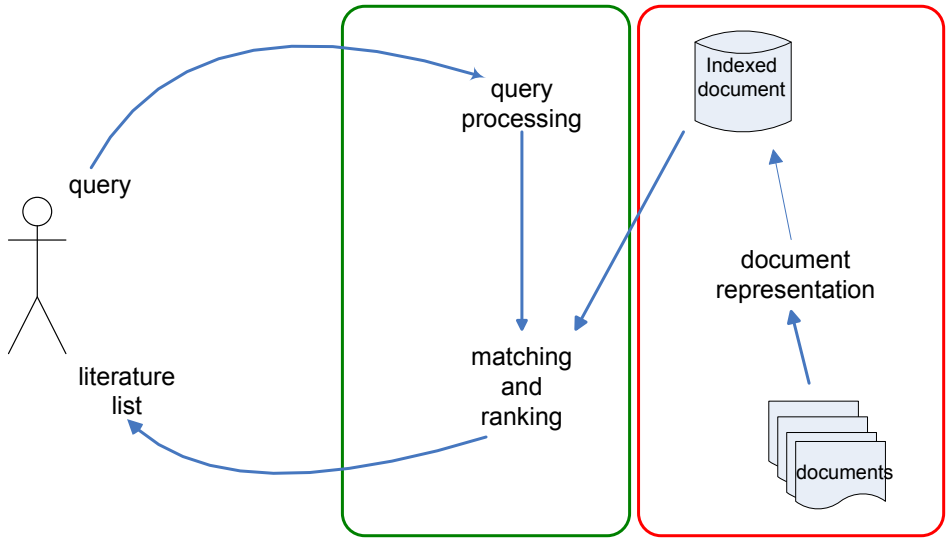


Figure 3.9 Components in document retrieval system.(modified from [Lidd05]). The green framed area represents the dynamic process and the red the static process. The indexed doucument could be saved in the database or flat file.

The heart of document representation phase is the document indexing process which represents each document by certain feature. The features are represented by the terms appearing or concepts/keywords discussed in the documents and are targeted during the computer search. The indexing process usually generates the final feature-target mapping in the form of table, in which the target refers to a document ID or a position in a document.

In the practice of existing biomedical document retrieval systems, two types of features are used as index for the document [LC09]. Some systems adopt the keyword-based indexing strategy by which only illustrative terms are selected as the indexes of an article, such as keywords of an article. This type of indexing requires the human intervention to comprehend the articles and choose the illustrative terms which best describe the article content. The full-text based indexing uses all phrases found in the article as the index and is adopted by full text document retrieval system and web search engines. The full-text based indexing returns not only the document containing the

indexed phrase but also can point out the position of its appearance in the document. During the full-text indexing process, some linguistic operations are carried out to avoid indexing unimportant phrases, such as stop words deletion, and deal with linguistic morphology (stemming). The index data structure of full-text based methodology is termed inverted file in the information retrieval society.

The query processing starts after receiving the free text query which users submit to the document retrieval system. Not all the words in the free text submitted by the user are informative and precise to the question. Some linguistic operations used in the document representation phase are also run in this phase, such as stop word deletion, stemming, and phrase recognition. Another important process of this phase is query expansion. Query expansion adds extra query terms to the original query term set derived from user's free text query. These extra terms may be synonymous terms or terms that are highly associated with the query term, based on co-occurrence statistics preferably computed on the same or a similar document collection as the one on which the search is being conducted [Lidd05]. Query expansion relieves the user from the need to generate all conceptual variants of their search terms [Lidd05].

During the matching phase, the processed user's query is compared with the index of each document, the similarity score is calculated between query and each candidate document, and a ranked list of documents is returned to the user.

The goal of a document retrieval system is to find all documents relevant to user's queries. Two effectiveness measures are intended to be maximized in the design of a document retrieval system: recall and precision. Recall measures the percentage ratio of the number of relevant records retrieved to the total number of relevant records in the database. Precision measures the percentage ratio of the number of relevant records retrieved to the total number of irrelevant and relevant records retrieved [LC09].

The following sections introduce the main document retrieval systems for biomedical research and are concentrated more on the systems which include the document indexing component (Table 3.2).

3.2.1 PubMed

PubMed is a document retrieval system accessing the most widely used bibliographic library in the biomedical research field–MedLine, which is developed by NLM. User

usually input query terms in the query box, and a list of bibliography is returned to the user. User then can browse each bibliographic record which includes title, authors, publication date, journal name, summary of the article and other bibliographic data. The bibliographic record might include a link to the website of journal publisher, and user may have the chance to see the full text article depending on the subscription status of the user's institute.

		PubMed	PubMed Central	BioText
Searchable field	*Bibliographic data(MeSH terms, journal name, author name, etc)*	Y	Y	Y
	Full text		Y	Y
	Table content and caption			Y
	Figure caption			Y
Indexing method		Keyword-bases	Keyword-bases	Full-text based
Automatic term mapping(translation, MeSH term expansion)		Y	Y	
Ranking method		Reverse chronical order	Reverse chronical order	Vector space model
API		eUtility	eUtility	web page

Table 3.2 Biomedical document retrieval system

Every new citation received from the journal publisher is assigned a PubMed Unique Identifier (PMID) and computationally indexed by PubMed. The indexing process is to create multiple machine-readable access points that refer to the different components of the journal citations for use when searching PubMed [Url8]. NLM adopts the keyword-based indexing methodology by using MeSH for the citation indexing. MeSH is a set of controlled vocabulary used for subject analysis of biomedical literature at NLM. MeSH terms are arranged in a hierarchical categorized manner called MeSH

Tree Structures. Each article entered in the MEDLINE is read by indexer who chooses the MeSH headings best describing the research subject of the article. These MeSH terms are added into the bibliographic record and later are used as one of the indexed fields of the article (Fig. 3.10). The other indexed fields are journal name, author name, and title/abstract [Url9] [Url10].

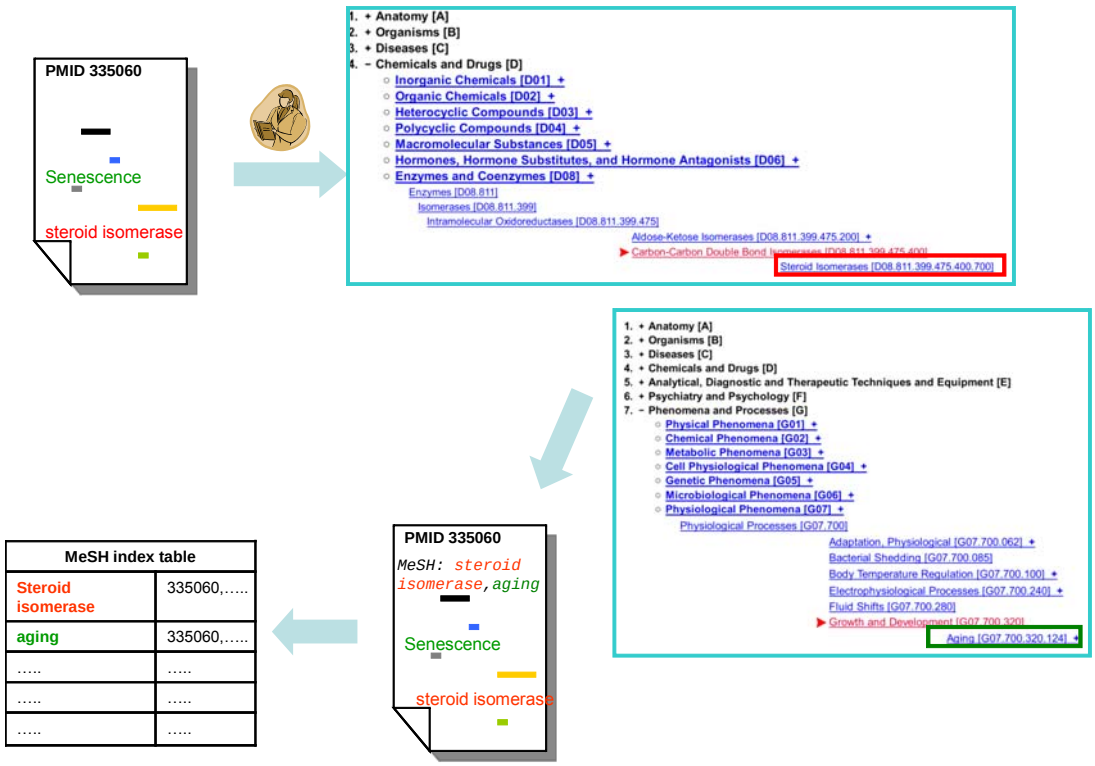


Figure 3.10 PubMed literature indexing process. After NLM receives the article citation, the indexer reads the article and chooses the proper MeSH terms to describe the article. The chosen MeSH terms are added to the bibliographic record and later are used as indices.

After users send their query to PubMed, the automatic term mapping (ATM) process is used on the terms entered without a qualifier. ATM looks for the translation tables and indexes in the following order: 1. MeSH translation table, 2. Journal translation table, 3. Author translation table. The translation table works as a synonym table to normalize various syntactic forms of the query terms. ATM first looks for the corresponding MeSH headings for the query terms in the MeSH translational table. Besides the corresponding MeSH headings, PubMed also does automatic MeSH exploding by adding more specific terms beneath those headings in the MeSH hierarchy. The literature list returned to the user contains the citation found in the index table for the found MeSH term set. If no citation is obtained until this stage, ATM keeps looking for

the translation tables and indexes of journal name and author name [Url8]. The retrieved documents are ranked in the reverse chronological order (Fig. 3.11).

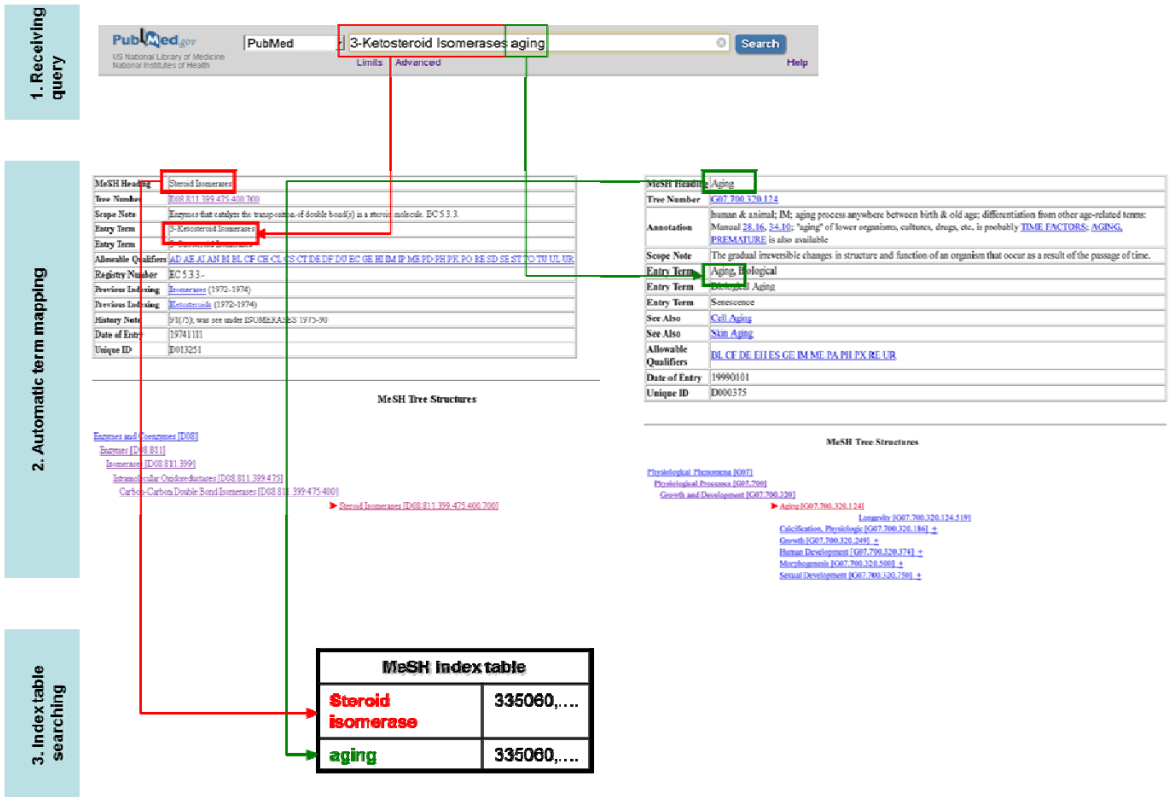


Figure 3.11 Automatic term mapping. The query terms entered without qualifiers are first looked up against the MeSH translation table and index followed by journal and author translation tables.

3.2.2 PubMed Central

PubMed Central (PMC) is a digital archive of life sciences journal literature. PMC provides free access to articles from journals that deposit their content in the archive. PMC differs from PubMed mainly in the following aspects [Url12]: 1. PubMed is a database of citations and abstracts for millions of articles from thousands of journals, while PMC archives over one million full-text journal articles. 2. Articles which were published prior to 1966 will be first digitally archived in PMC, and the citation of these articles will be included in PubMed several months after the PMC archiving. Each article in PMC is associated with a PMC unique identifier (PMCID), which is a different set of identifier from PMID. The literature searching process in PMC functions similarly as in PubMed including automatic term mapping and returning the MeSH indexed articles.

3.2.3 PubMed derivatives

PubMed hosts over 21-million citations as in 2012, and millions of queries are submitted each day by users around the world. Over one-third of PubMed queries result in 100 or more citations. Since its heavy usage in the biomedical research area, several PubMed derivatives are developed to facilitate a quickly and efficiently search and retrieval of relevant publications [Lu11].

Lu has reviewed other tools comparable to PubMed at the time of May 2011. Lu categorized the existing 28 PubMed derivatives into four groups according to their most notable features: ranking search results, clustering results into topics, enriching results with semantics and visualization, and improving search interface and retrieval experience. Lu summarized that improving ranking and the user interface seem to be the more popular directions. Most of the improved ranking algorithms rely on the extra information provided by the user. The extra information may be the user's feedback on the retrieved documents, a set of relevant documents provided by the user, and user's "click through" history. The presentation of search results in the 28 tools is primarily list based. Some tools provide the tabular or graph presentations when they are able to extract and display semantic relations.

3.2.4 BioText

Although its heavy usage in the biomedical research community, PubMed only searches over title, abstract, and document metadata, without making use of the full text [HDG+07]. However, the most important evidence supporting the argument of the article is displayed in the figure and table sections of the scientific publication. It is very common for the research to read the abstract in the first place and move to the figure and table caption. Divoli *et al*. conducted a survey over the desirable interface of a biomedical literature search engine [DWH10]. They concluded that 19 out of 20 of their survey participants expressed a desire to use a bioscience literature search engine that displays article' figures alongside the full text search results. The full text search distinguishes from the design of PubMed in that it tries to find a match between the user's query term and the full context of the article beyond the bibliographic data which is indexed by PubMed.

BioText [HDG+07] provides the full text article search which is beyond the searching scope in PubMed. Besides the full text content, BioText also searches for the

article title, table captions, table contents and figure captions. It indexes all Open Access articles available at PubMed Central. BioText uses Lucene [Url11] open source search engine to index, retrieve, and rank the biomedical literature.

3.3 Summary

The conventional strategy of biological pathway construction relies on the information stored in database and thorough review of primary literature [VSP+08]. However, biological databases often preferentially contain data of extensively investigated subjects. If the focus is placed on less investigated and more specialized subjects, the data deficiency issue emerges. On the other hand, manual review of primary literature is a tedious and time-consuming work. In addition to manual information extraction from literature, automated text mining is another choice but has the issue of low accuracy [VSP+08]. The construction of redox regulatory network faces the same data deficiency issue. Network construction is hindered if the expected data are missing from the database or text-mining tools or sometimes becomes too rich to handle, since reliable target selection may pose a problem.

The first part of this chapter summarizes that only few Trx/Grx target proteins could be found in the public protein-protein interaction databases by visiting either individual primary database or the integrated one. The same data deficiency problem is also observed through the text-mining tool. When the criteria of organelle, tissue, and species are added into the network construction, barely any target protein can be found. Besides the data scarcity problem from the public molecular interaction database, the uncertainty of the interaction type between proteins is another issue. The putative target proteins of Trx/Grx should interact with Trx/Grx through thiol-disulfide exchange mechanism. However, this type of interaction mechanism between molecules is not annotated in most databases. The lack of available data and uncertainty of the interaction mechanism from the database and text-mining tool thus becomes an obstacle in the network construction process. Therefore, the strategies of simple database integration and text-mining of the unstructured data in the literature currently is not applicable to the problem of identification of target protein of Trx/Grx in human mitochondrion.

One direction for target protein prediction is to use the bottom-up strategy by identifying the functional unit, such as motif and residue, essential for the reversibly oxidizable reaction (Fig. 1.1). Due to the diverse functionality that the known target proteins are involved (Appendix A), the trial of common motif search among the reported

target proteins will fail due to the non-conservative protein sequence. However, the conservative motif search is applicable to the identification of new oxidoreductase, as explained in the section 3.1.3.2. Some computational methods are also available for predicting redox-active cysteine, but most works are about predicting catalytic redox-active cysteines and less focus on regulatory cysteines. And unfortunately, the regulatory cysteine is the cysteine type which must be found in the target proteins of Trx/Grx. In order to identity all types of reversibly oxidizable cysteine, the decision rule devised by Sanchez *et al.* has provided a good starting point for firstly detecting "oxidizable" cysteine. The decision rule from Sanchez *et al.* is thus used for the computational pre-selection of thioredoxin/glutaredoxin target proteins by detecting oxidation susceptible cysteines. The implementation work is described in the next chapter.

The second part of this chapter talks about the document retrieval system in the biomedicine domain. The document retrieval activity is important for scientist to follow the latest development, idea generation, and explanation of experimental result. This part covers the main document retrieval systems in biomedicine as well as the explanation of some critical steps in those systems. The capability of the existing document retrieval system is to honestly return the document which contains the query terms syntactically or semantically. To enhance the recall of a document retrieval system, the query term expansion is sometimes carried out by incorporating the synonym of user's original query. Therefore, only document implying direct relationship between query terms is returned by the existing biomedical document retrieval systems. When the directly relevant literature doesn't exit, the retrieval of indirectly relevant literature may be beneficial.

Chapter 4

Pre-selection of target protein in the redox regulatory network *

In the workflow of redox regulatory network construction, the identification of target protein for thioredoxin/glutaredoxin is an essential process. The target proteins could be the catalytic enzymes in the biological pathways which take part in the metabolic turnover. The capability of the redox regulatory network will affect the state of its target proteins, and this influence will propagate to the corresponding pathway and lead to phenotype change or sickness. In spite of knowing the importance of identifying target proteins, the critical properties characterizing the thiol-disulfide exchange between thioredoxin/glutaredoxin and their target proteins are not clear. One certain thing about the target proteins is that they should contain reversibly oxidized cysteine residues, which implies that the expected cysteine should be "oxidizable" and "re-reducible". Besides the target proteins, the redox active cysteine also exists in the oxidoreductase enzyme and is located in the catalytic active site. Most existing work concerning redox active cysteine focused on the prediction of oxidoreductase enzyme class and left the proteins containing redox active cysteines residing in other regions out of consideration. And the proteins that were ignored are the potential candidates of target proteins.

Reversibly Oxidized Cysteine Detector (ROCD) is a tool for the pre-selection of thioredoxin/glutaredoxin target proteins in the construction of redox regulatory network. The target protein is pre-selected by detecting its bearing of oxidation susceptible

* Part of chapter 4 has been published in Journal of Integrative Bioinformatics [LDH10]

cysteine residues. The oxidation susceptibility is one of the prerequisites for the cysteine being reversibly oxidizable. ROCD adopted the Cysteine Oxidation Prediction Algorithm (COPA) developed by Sanchez *et al.* as shown in section 3.1.3.2 [SRWM08].

4.1 ROCD architecture

The architecture of ROCD is shown in Fig. 4.1. A command-line program and a web interface were created in this work. The program was implemented in Java and JSP, and the in-house HPRD and iProClass database are stored in MySQL. In the command-line mode, users need to provide: (i) List of SwissProt accession numbers, or to specify the targeted human tissue and organelle (ii) Criteria for the thiol-thiol distance, accessible surface area, and pK_a, (iii) File names to save the output files.

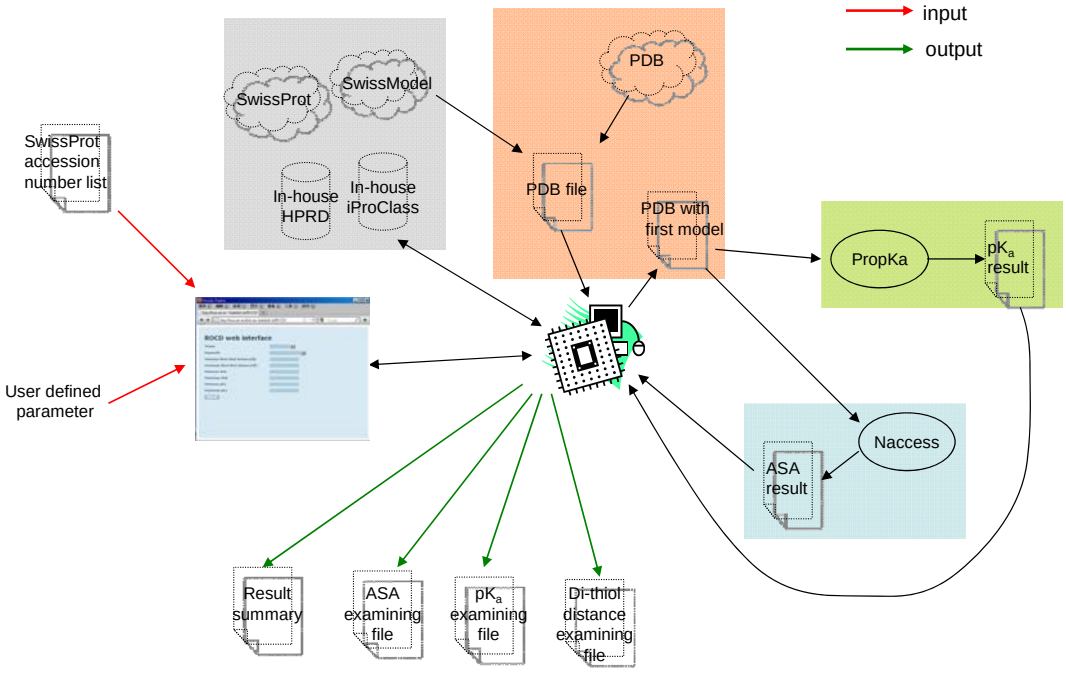


Figure 4.1 The architecture of ROCD. ROCD utilizes five public databases (HPRD, iProClass, PDB, SwissProt, SwissModel Repository) and two external tools (PropKa, Naccess) for the selection of oxidation susceptible cysteinyl residues. The grey shaded region indicate the database wrapper module, the orange shaded region the PDB processor module, the green shaded region the pK_a calculator module, and the blue shaded region the ASA calculator module.

After receiving the necessary parameters, ROCD queries the in-house HPRD for tissue and organelle specific protein data set and iProClass database to find the PDB ID for each SwissProt accession number. The SwissProt sequence entry and the PDB file are

downloaded from SwissProt, SwissModel and PDB through HTTP request. After tailoring the original PDB file, the tailored PDB file served as the input for PropKa and Naccess. Finally, it generates four output files – one file for the calculated distances of all thiol pairs, one for all the calculated pK_a for the cysteine residues, one for all the calculated ASA for the SG (cysteine sulfur) atom, and the result summary.

4.2 Implementation

4.2.1 Dependent external software

The values of three critical physicochemical properties have to be determined for the estimation of oxidation susceptible cysteines. The distance between cysteine thiols can be calculated from the atom coordinates in the PDB file. The other two properties are obtained by two external software–PropKa for the pK_a value and Naccess for the accessible surface area.

4.2.1.1 PropKa—pK_a calculator

The acid dissociation constant is a measure of the protonation/deprotonation tendency. The protonation of the amino acid residue determines several important properties including protein solubility, protein folding and catalytic activity [DTMF06]. Traditionally the pK_a is determined from the titration curves obtained in experiments. Several computational prediction programs for pK_a value are present. Davies *et al*. has evaluated four pK_a prediction programs: MCCE, MEAD (later renamed to PCE), PropKa and UHBD. They concluded that PropKa produced more accurate prediction in their overall test and computationally performed faster than the other three programs [DTMF06]. Besides the fore-mentioned programs, H++ and PKD are two other web-based programs. We chose PropKa as our pK_a calculator owing to our need of a standalone, accurate and efficient program for incorporating into our pipeline.

4.2.1.2 Naccess—ASA calculator

The concept of accessible solvent area was proposed by Lee and Richard and defined as the area composed of the trace of the center of a probe rolling over the protein. It can be considered as an expanded van der Waals surface, namely by increasing the van der Waals radius by the probe radius. Lee and Richards approximated the accessible surface area of each atom using the formula:

$$\text{accessible surface area} = \sum (R / \sqrt{R^2 - Z_i^2}) \cdot D \cdot L_i, \quad D = \Delta Z / 2 + \Delta' Z$$

where L_i is the length of the arc drawn on a given cross-section i, Z_i is the perpendicular distance from the center of the sphere to the cross-section i, ΔZ is the spacing between the cross-sections, and $\Delta' Z$ is $\Delta Z/2$ or R- Z_i, whichever is smaller. Summation is done over all arcs drawn for the given atom [LR71] (Fig. 4.2).

Naccess is an implementation of Lee and Richards's method. It calculates the ASA for each atom in the given PDB file and also provides the ASA for each residue by summing the atomic ASA over each protein or nucleic acid residue [HT93].

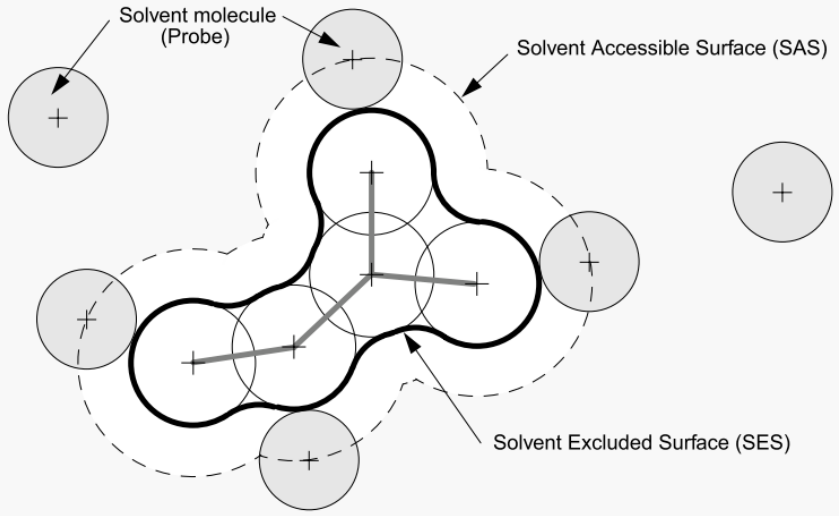


Figure 4.2 Definition of accessible solvent area. This figure shows a cross-section of a molecule. The arc drawn in dotted line is the ASA of one cross-section. (Source: [SOS96])

4.2.2 Construction of in-house databases

ROCD relies on five external databases- iProClass, SwissProt, PDB [RBB+11], HPRD, and SwissModel Repository (SMR)[KAK+09]. To shorten the time for the data retrieval, the databases which are queried frequently during the execution have a local copy of the data content on our server. Therefore, we set up the local MySQL database of iProClass and HPRD.

iProClass is a protein-centered database providing links and ID mapping to over 50 databases. We downloaded the iProClass tab-delimited flat file from its FTP site and parsed the selected columns, which correspond to the accession number/ID of other databases, into different data tables in our MySQL database.

HPRD is the other database for which we implemented a local copy. After downloading the HPRD flat file (downloaded on July 21st, 2010), we extracted the data which are essential for ROCD and parsed them into MySQL tables. The downloaded folder contained 13 text files, and each of the text file contained data for specific domains, such as protein-protein interaction, genetic disease, etc. We selected the text file containing intended data for ROCD and parsed each of selected files into MySQL table. In the end four tables for ID mapping, subcellular localization, protein-protein interaction, and tissue expression were created.

4.2.3 ROCD workflow

The algorithms implemented by ROCD are depicted in Fig. 4.3. The whole pipeline is composed of five modules–database wrapper, PDB processor, distance calculator, pK_a calculator, and ASA calculator.

ROCD retrieves data from its dependent databases by means of in-house MySQL query and external HTTP request, which are performed by the database wrapper. The in-house iProClass database is used to find the corresponding PDB ID and the residing chains for the queried SwissProt ID (SPID) or accession number (SPACC), the in-house HPRD to obtain human tissue/organelle specific protein set, SwissProt to retrieve the SMR link for the modeled protein structure through HTTP request, SMR to download the model in PDB format through HTTP request, PDB to download the protein structure through HTTP request. When a PDB structure is unavailable for the queried SPID/SPACC, the URL link to the SMR webpage is extracted from the HTML file of the SwissProt webpage, and the modeled structure is downloaded from SMR.

The PDB processor is responsible for modifying the download PDB file in order to comply with the requirement of input format for the subsequent calculation of thiol-thiol distance, pK_a, and ASA. The downloaded original PDB file might contain the atom coordinates from different models and peptide chains. One model represents one resolved structure. When the protein structure is determined by NMR (Nuclear Magnetic Resonance) technique, multiple models are possibly recorded. If the downloaded PDB structure contains multiple models, only the coordinates of the first model is retained so that the PDB file conforms the limitation on atom number in Naccess. In one structural model, it might contain several peptide chains. Some of the peptide chains belong to the ligand which the protein binds to and have to be removed. Only the coordinates of the peptide chains which are indicated from iProClass are maintained for the following

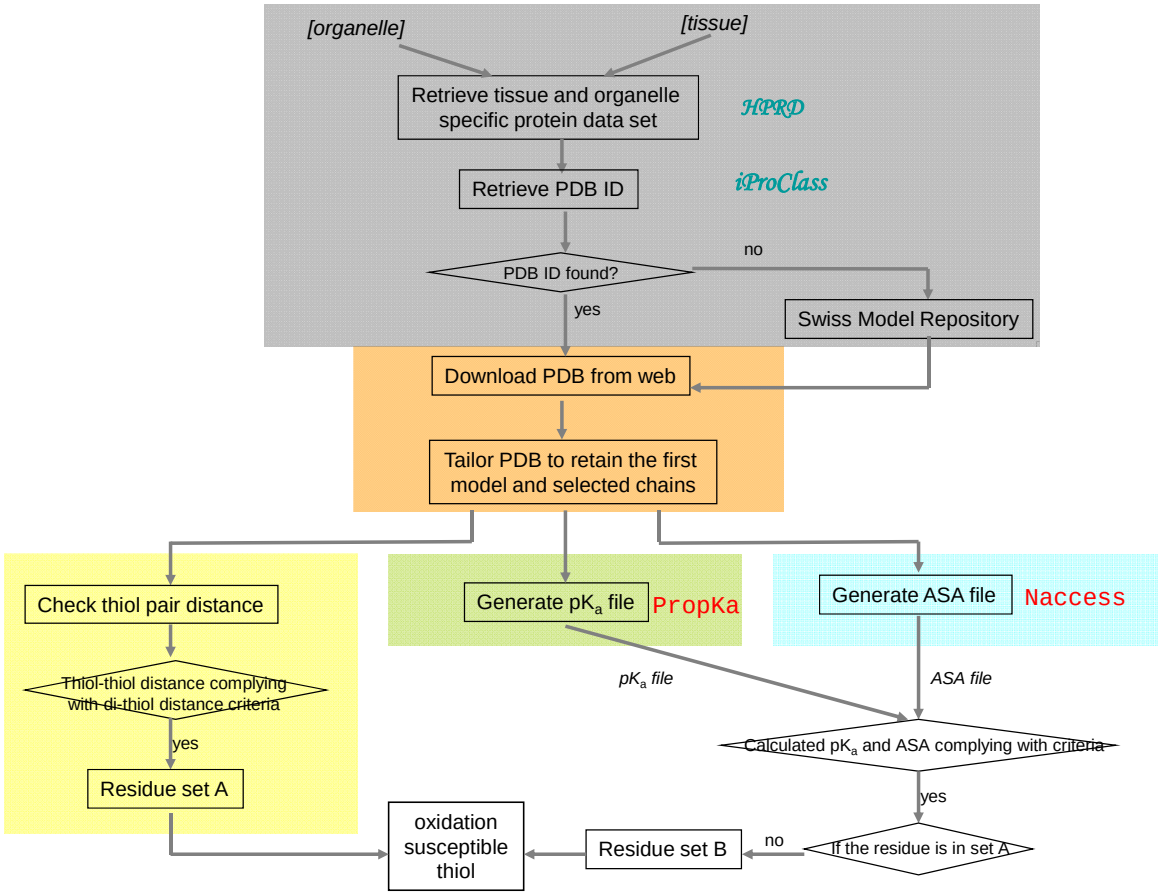


Figure 4.3 The workflow and the five modules to predict proteins with oxidation susceptible cysteinyl residues. The grey shaded region indicates the database wrapper module, the orange shaded region the PDB processor module, the yellow shaded region the distance calculator module, the green shaded region the pK_a calculator module, and the blue shaded region the ASA calculator module. The green colored italic font denotes the database, and the red font denotes another standalone program.

calculation. The output of the PDB processor is a tailored PDB file.

Distance calculator, pK_a calculator, and ASA calculator carry out the calculation of three physicochemical properties from the tailored PDB file. Distance calculator first extracts the coordinates of all thiols of a protein. Sometimes the cysteine residue of a wild type protein was mutated to serine before crystallography for certain reasons. Therefore, the coordinates of certain hydroxyls are also extracted according to the annotation in the "SEQADV" record of the PDB file. Afterward, the distance between any two extracted coordinates is calculated. pK_a calculator and ASA calculator make external calls to PropKa and Naccess programs.

If there is any calculated thiol-thiol distance falling in the user-defined range, the

residue numbers of the cysteine pair are stored in the residue set A. Following distance calculation, pK_a and ASA are calculated by PropKa and Naccess respectively with tailored PDB as the input. If the cysteine residue fulfills the following criteria: (i) being not listed in residue set A, (ii) pK_a and ASA falling in the user-defined range, (iii) locating on the chain as annotated in iProClass, it is stored in the residue set B. In the end, the protein entry and the qualified residues in residue sets A and B are written to the result summary. The result summary from ROCD keeps the following information for each tested protein– 1. SwissProt accession number, 2. SwissProt ID, 3. protein name, 4. PDB ID, 5. peptide coverage in PDB, 6. peptide sequence identity between the tested protein and template which the structure modeling is based on, 7. location of mature peptide, 8. chain symbols and residue numbers of cysteine pair which fulfills the thiol-thiol distance criteria, 9. chain symbol and residue numbers of cysteine which fulfills the pK_a and ASA criteria. During the execution of ROCD, the calculated thiol pair distance, pK_a, and ASA value for each cysteine residue are also written out to separate files for examination.

4.3 Validation of ROCD prediction

The annotation of cysteine residues in Balanced Susceptible Cysteine Thiol Database (BALOSCTdb) from Sanchez *et al.*'s study was used as the gold standard to validate ROCD. BALOSCTdb contains 161 cysteine thiols that are susceptible to oxidation and 161 cysteine thiols that are not.

To validate the ROCD implementation, each PDB ID in BALOSCTdb is intended to be tested by ROCD. However, ROCD only takes SPACC as the input. Therefore, the corresponding SPACCs for all PDB IDs in BALOSCTdb are first retrieved by the ID mapping service of iProClass. The retrieved SPACCs are sent to ROCD later. The parameters were chosen as specified in Sanchez *et al*.–6.2 for thiol-thiol distance, 1.3 for ASA, and 9.05 for pK_a. The result from ROCD concerning the PDB ID in BALOSCTdb is listed in Appendix B.

The compliance of our prediction with BALOSCTdb was checked (Appendix C), and the summary is shown in Table 4.1. Our prediction achieved 77.6% accuracy for the cysteine residues which are marked "oxidation susceptible" in BALOSCTdb and 74.5 % for "non-oxidation susceptible".

		Actual condition	
		Non-oxidation susceptible	**Oxidation susceptible**
Prediction result	**Non-oxidation susceptible**	120	36
	Oxidation susceptible	41	125

Table 4.1 Result from testing ROCD against BALOSCTdb

4.4 Examination of thioredoxin target proteins in plant mitochondrion

Balmer *et al.* [BVT+04] collected 46 thioredoxin target proteins in the plant mitochondrion. The target protein of thioredoxin should contain oxidation susceptible cysteines which are re-reduced by thioredoxin during interaction. We use ROCD to test the existence of oxidation susceptible cysteines for these 46 proteins using the following criteria: (i) thiol-thiol distance ≤ 6.2 Å, (ii) accessible solvent area ≥ 1.3 Å^2, (iii) $pK_a \leq 9.05$. Since the CxxC motif is found frequently in the enzyme active site of oxidoreductase and should consist of oxidation susceptible cysteines, we also scanned the existence of CxxC motif in these 46 proteins. The result is shown in Appendix D.

Among the 46 proteins, only four had a resolved protein structure. After searching SwissModel–a homology modelling database, we found the modeled structures for the other 26 proteins. For each of the proteins being modeled in SwissModel, the structure was created based on a structure "template", which is an experimentally-determined structure of a close homologue. The close homologue is the protein which shares certain sequence identity with the modeled protein. The structure modeling report from SwissModel indicates the region which is modeled and the sequence identity between the modeled region and the template. For the 26 proteins with modeled structures retrieved from SwissModel, the model could often be built only for a partial peptide sequence of the protein. In the structure model for SPACC Q9ZT91, only 7% of the entire mature peptide is assigned a predicted structure. And the sequence identities between the modeled region and the template range from 34 to 93% for these 26 proteins.

In the end we obtained the structures for 30 proteins, for either the entire or partial

peptide and through either experimental or homology modeling method. 20 out of these 30 proteins have passed the pre-selection of ROCD. In contrast to this ROCD result, there are only 3 proteins containing the CxxC motif in these 30 proteins.

According to this result from ROCD, we can separate 46 proteins into 3 groups: (A) 20 proteins which have an experimentally determined or modeled structure and are predicted to have OSC by ROCD; (B) 10 proteins which have an experimentally determined or modeled structure but no OSC is predicted (P16048 belongs to group A, since one of its three structures has OSC); (C) 16 proteins which have no experimental/modeled structure. When we inspected the ROCD result for proteins in group B, the protein entry "P17614" has an associated PDB ID "1HPC", but this associated structure is for the signal peptide not the mature peptide. Therefore, ROCD is unable to predict the existence of OSC in the mature protein of P17614. For another protein entry in group B, Q9ZT91 has a predicted structure only for a short fragment (29 amino acids) of the mature peptide (403 amino acids), representing only 7% of the mature peptide.

4.5 Summary

The Redox regulatory network is a central and evolutionarily conserved feature of the cell. The inventory of dithiol-disulfide transition proteins is important for *in silico* construction of the redox regulatory network [Diet08]. The goal of this part of work is to predict proteins that undergo reversible oxidization by examining the existence of oxidation susceptible cysteine. We implemented Sanchez *et al*'s algorithm to allow high-throughput and automatic *in silico* pre-selection of dithiol-disulfide transition proteins. There are several phases in the pipeline for achieving automation: database querying and data retrieving, external programs execution, and data processing.

The identification of oxidation susceptible thiols hints to potential thioredoxin/ glutaredoxin target proteins. In comparison to other prediction methods concerning redox active cysteine, Sanchez *et al.* provided a defined algorithm with numerical values to detect the necessary structure and physicochemical properties. In their leave-one-out cross-validation analysis, 80.1% accuracy was reported. By implementing the Sanchez algorithm and adopting their criteria, ROCD achieved an overall accuracy rate of 73.8% when using BALOSCTdb as the gold standard.

We have applied ROCD on 46 known thioredoxin target proteins in plant

mitochondrion. Due to the shortage of resolved structure, the testing on these proteins relied mostly on modeled structures. After omitting the structures with coverage lower than 10% of the entire mature peptide and the one for the signal peptide, 28 proteins with a valid structure were left, and 20 of them were detected to have OSC by ROCD when the Sanchez *et al.*'s criteria were applied. Besides the coverage issue, the sequence identity between the modeled protein and template is also worth inspecting. User should check these values and then choose a subset from the pre-selected proteins for further validation.

The identification of regulated target proteins is a critical step in the post-translational regulatory network construction. Since the network construction mostly relies on the data stored in the biological database or information extraction from the literature, the network construction work is hindered, once the required data is unavailable from these sources. This work demonstrates a strategy based on physicochemical and structural properties to fill the gap between specialized and limited knowledge deposited in literature and databases, and the advancement of network construction. ROCD is applicable for the thioredoxin/glutaredoxin mediated redox regulatory network. Other regulatory networks, such as the phosphorylation network, also require a target prediction tool like ROCD, if the prediction algorithm is available.

The result from ROCD just provides the predicted candidate for the proposed biological question. The prediction needs to be supported by direct or indirect evidence from the lab or literature. Besides experimental method, verification of the pre-selected candidates for the redox network construction could depend on manual curation by literature reading. However, searching for the directly relevant literature has limitations in particular if less investigated or novel research topic is involved. The indirectly relevant literature retrieved based on the established association between biological concepts is valuable under this literature deficiency scenario. In chapter 5, a literature search tool designated to provide indirectly relevant literature to support manual curation is introduced.

Chapter 5

Network-contexted Document Retrieval System

In the previous chapter, a pre-selection tool–ROCD–for the identification of thioredoxin/ glutaredoxin target protein (TTG) is introduced. ROCD implements the decision rule derived from data-mining on biochemical properties of protein. The computational prediction is to provide the possible candidate for proposed biological question and has to be validated by experimental or manual curation through literature reading. Searching for the directly relevant literature to support the prediction usually results in the literature deficiency when the research topic has been less investigated or is very specific or a novel research field, such as the topic of TTG identification in human which this thesis is tackling. This chapter is describing a network-contexted document retrieval system (ncDocReSy) to help biologists extend the literature search according to the established association between bioentities in the biological network.

5.1 Introduction

All living phenomena of the cell depend on the assembly or interaction of biological entities. Assemblies of interactions form biological network, and the variation of one single bioentity may influence the activity of the other members in the network. The construction of biological network allows the understanding of the biological phenomenon in a systematic way. Depending on the involved bioentities and molecular function of the interaction, biological networks can be classified into three main categories–protein-protein interaction networks, metabolic networks and gene expression networks. After several years of investigation in life science, the knowledge of biological

networks has been enriched and collected in several biological network databases. The availability of well-structured knowledge in biological network databases eases the computational access and advanced processing.

Besides the structured data deposited in molecular databases, the biomedical literatures is published on the scale of over 500,000 per year and hosts unstructured knowledge. Due to the wealth of the published literature, it is not feasible for researchers to manually read all literature and directly access the new knowledge in a comprehensive manner. Literature-mining tools have been developed to help researchers identify relevant papers, recognize the biological entities that are mentioned in the papers, and extract specific facts [JSB06]. The identification of relevant papers–a process known as information retrieval (IR)–is the selection of documents related to the researcher's interest, which is the activity carried out on a regular basis by biologists in order to interpret experimental result and keep up with the latest scientific development.

The vast amount of biomedical publication has provided rich material for the literature-based discovery. Literature-based discovery adopts the transitive inference [Rodr09], best known as Swanson's ABC model [SST06], to build connection between two implicitly linked biomedical concepts. ABC model states that "if A and B are related, and B and C are related, it follows that A and C might be indirectly related" [WKM05]. Swanson has made the connection between fish oil and Raynaud's disease by reading the literature and applying the ABC model [Swan86], which was validated experimentally later [DKS89]. Modern information extraction technique has been used to computationally construct the associations between A and B as well as between B and C by mining the literature and molecular databases and infer the distal relationship between A and C. Some indirect relationship discovery tools utilizing the ABC model have been developed, such as BITOLA [HPMH05] and FACTA+ [TMH+11].

In the scenario of specialized or less-investigated research field, direct supporting reference is usually lacking through conventional literature search engine, such as PubMed and its derivatives [Lu11]. The search of indirect supporting reference is beneficial under this scenario. Apart from the association discovered through text-mining of the literature, the established knowledge of biological network could be the alternative source providing biomedical concept association. The incorporation of biological networks into the literature search can be interpreted as bringing transitive inference into document retrieval. When the ABC model is used in literature search, the model states

that “if article A is related to molecule α, and molecule α is related to molecule β, it follows that article A and molecule β might be indirectly related”. This chapter presents the network-contexted document retrieval system (ncDocReSy), which is developed to present indirectly relevant publication by conducting the transitive inference following the biological network topology.

5.2 Criteria of ncDocReSy

ncDocReSy is designed to combine biological networks and literature search, where the relevant literature can be viewed in the context of biological network. Names of genes/proteins and chemicals/drugs have been recognized as one of the most frequent category of queries through query log analysis [DMNL09]. Therefore, ncDocReSy focuses on the bioentity networks which consist of genes, proteins, metabolites, and chemical compounds. Thus the inclusion of three major categories of biological networks of cellular process is planned for ncDocReSy. ncDocReSy adopts the more reliable data resources for the construction of biological networks. In the instance of protein-protein interaction network construction, the data resource of experimentally determined protein-protein interaction is more preferable than of computationally predicted one.

ncDocReSy is intended to maximize the retrieval of relevant articles which have mentioned the name of specific bioentities, either being an enzyme, metabolite, or chemical, with the correct semantics within the article context. To achieve this, three bibliographic databases are includes, and the named-entity recognition (NER) tool is used to ensure the full name of the expected entity is present and also in the correct semantic type in the article context.

ncDocReSy relies on several external resources for the network construction and literature search. The external resources must be internet-accessible, so that a local installation of the relied resources is waived. But for some resources which lack an API or whose API does not fit our need, a local installation of the resources is used. The document retrieval function of ncDocReSy borrows from existing literature search tools so that ncDocReSy does not depend on the implementation of its own document indexing component. ncDocReSy accesses the primary bibliographic databases PubMed, PMC and BioText directly and retrieves the generated literature list. Although many literature search tools, referred as PubMed derivatives by Lu [Lu11], have been implemented for quick and efficient search and retrieval of publications, these derivatives have

manipulated the raw literature list returned by PubMed according to certain criteria. Because ncDocReSy has implemented its own literature ranking algorithm based on the network context, the preliminary literature list from the primary bibliographic databases is sufficient.

Since ncDocReSy incorporates biological network into document retrieval, an integration platform has to be chosen for graphical representation of biological networks and display of literature search result. Cytoscape became the first choice because of the well-documented API and support from the vast user community.

According to Lu's classification of PubMed derivatives [Lu11], ncDocReSy is a document retrieval system more oriented towards the representation and ranking of literature list under the network context.

5.3 Architecture

ncDocReSy takes advantage of the existing bioinformatics resources in the public domain and avoids re-inventing the wheel. Besides our local database, most of the relied bioinformatics tools or databases are web-based and with APIs provided. The functions provided by ncDocReSy can be divided into three modules: network construction module, document retrieval module, and literature list refinement module. The different resources utilized in each module are described below (Fig. 5.1).

5.3.1 Network construction module

The network construction module is responsible for the construction of bioentity networks including metabolic networks, protein-protein, and protein-chemical interaction networks. The KEGG reference metabolic pathways are pre-constructed based on KEGG data in DAWIS-M.D., formatted in CSML, and saved in a MySQL table.

A CSML parser is created to parse the CSML-formatted pathway data. When the user requests a KEGG reference pathway, the pathway CSML file is obtained from the database and processed by the CSML parser. Besides the CSML file obtained from our local database, ncDocReSy allows users to provide their CSML file which also can be processed by the CSML parser. The protein-protein interaction data are retrieved from IntAct through the Representational State Transfer (REST) service at real time. For protein-chemical interaction information, a protein-chemical interaction MySQL table

has been constructed by parsing the downloaded file from STITCH database [KSF+12] and queried upon request.

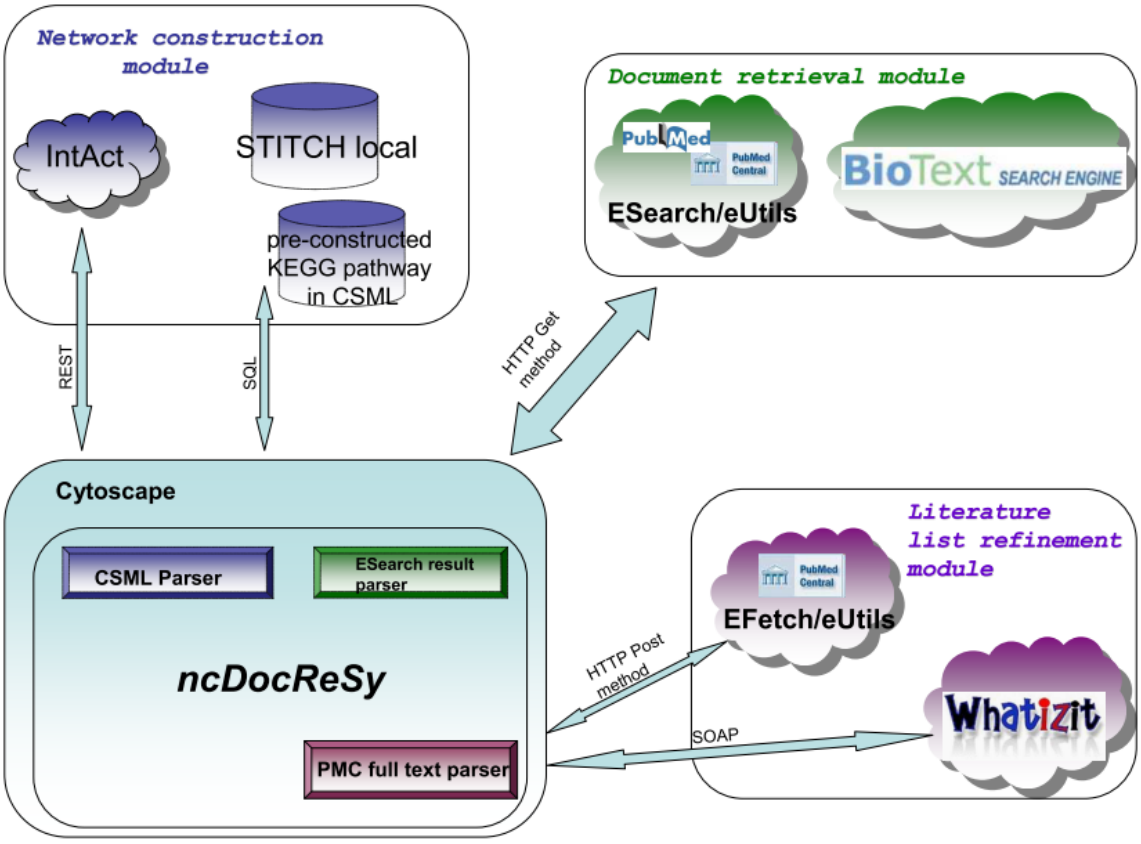


Figure 5.1 Three modules in ncDocReSy. The network construction module is responsible for the bioentity network construction, the document retrieval module for the retrieval of literature lists from the primary literature search engines, and the literature list refinement module for confirming the correct semantics of the matched terms.

5.3.2 Document retrieval module

The document retrieval module carries out the literature search function. ncDocReSy takes advantage of existing biomedical document retrieval systems. The KEGG and UniProt data in DAWIS-M.D. is used for the common name lookup for enzyme EC number/metabolite and protein, respectively. PubMed, PMC, and BioText are used as the external literature search engines. ncDocReSy uses the ESearch module of Entrez Programming Utilities (eUtils) for retrieving the literature list from PubMed and PMC. The query terms are included in a URL string, and the HTTP Get method is used for sending the query and receiving response from PubMed and PMC. The returned literature list from ESearch module is enclosed in XML and is further processed by an ESearch result parser which extracts the PMIDs and PMCIDs. The inquiry with BioText also

occurs through the inclusion of query terms in the URL and the use of HTTP Get method. The response from BioText is a web page in HTML which contains the PMIDs of the articles.

5.3.3 Literature list refinement module

The literature list refinement module is to check if the full name of the intended bioentity appears in the article context and is present in the correct semantic type, such as protein or metabolite. ncDocReSy takes advantage of the named entity recognition function of whatizit [RAG+08] for this purpose. Whatizit has implemented several pipelines which recognize the named entity in different semantic types. The web service of whatizit returns the result in which a semantic type tag and the URL pointing to the data entry in the biomolecular database are appended to the identified entity.

The text submitted to whatizit from ncDocReSy for named entity tagging is either the article abstract from PubMed or a segmented full text article from PMC. The full text article retrieved from PMC through eUtils contains extra tags which describe the content and metadata of journal articles [Url13]. Before sending the full text content to whatizit, different sections, such as introduction, methods, conclusion, and discussion, are identified by the PMC full text parser and submitted to whatizit separately. The full text parser is created by Java Architecture XML Binding (JAXB) [Url14] from the mixed XML schema recovered from two PMC articles (PMC2868029, PMC2584013). The semantic type tagged text from whatizit is parsed to produce the mapping between semantic type and a list of identified bioentity in that semantic type.

5.3.4 Network editor

ncDocReSy is implemented as a Cytoscape plugin, and all the function buttons are displayed in a new tab in the control panel of the Cytoscape workspace (Fig.5.2). The ncDocReSy tab has two main components: network construction component and document retrieval component. The network construction component asks for the user-provided ID of starting bioentity and its molecular type (protein or metabolite) for building the metabolic network. The network construction component also includes buttons for adding protein-protein and protein-chemical interactions. ncDocReSy also provides a CSML importer for constructing network from the user-provided CSML file. The document retrieval component allows users to type in free text query terms separated by space, to choose the document retrieval engines for literature search, to use the phrase search mode, and to restrict the publication year of the returned literature. The last

partition of this component asks for parameters for ranking the literature according to the network context. Besides, the literature list refinement button is also in this component.

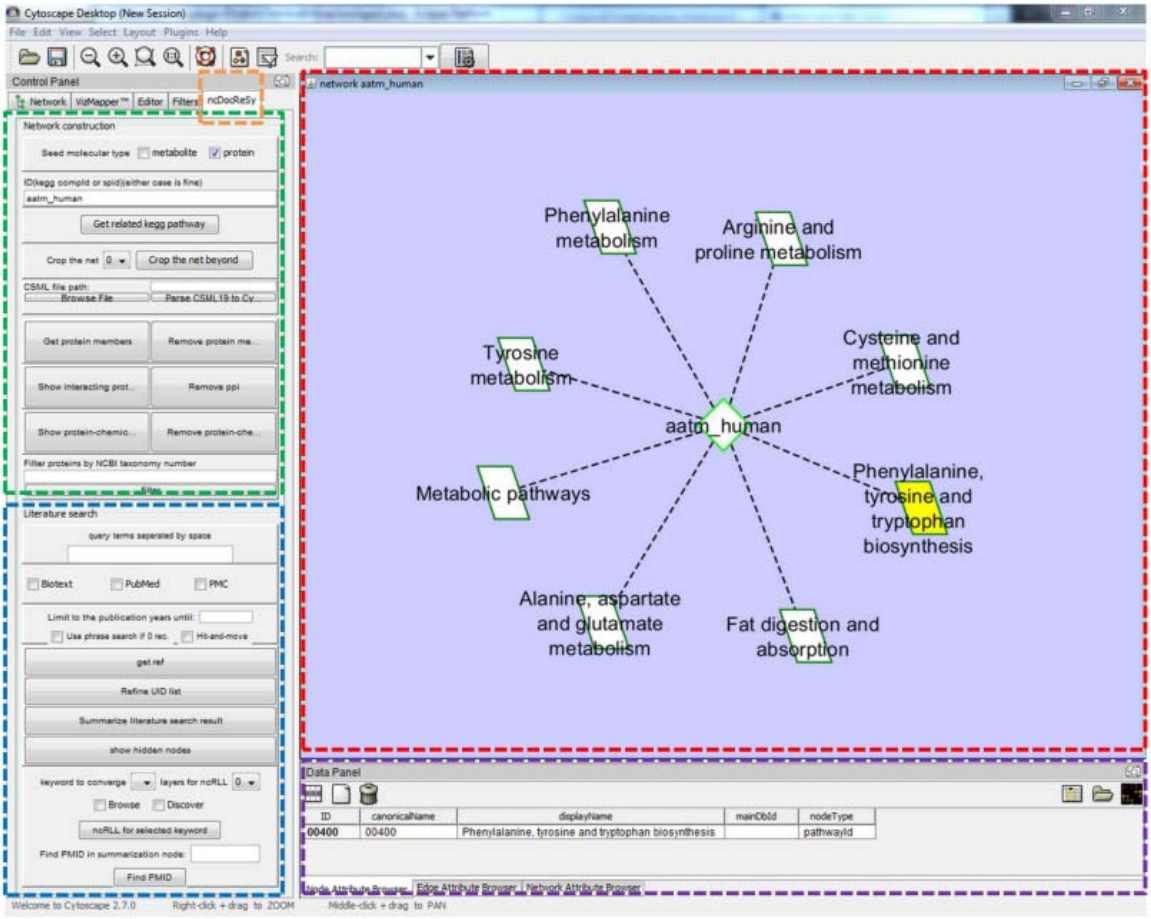


Figure 5.2 The graphical user interface of ncDocReSy. The ncDocResy tab in the control panel is marked in orange, network construction component in green, document retrieval component in blue, main network view window in red, and data panel in purple.

5.4 Implementation

This section explains the methods implemented in ncDocReSy to construct the biological network, to retrieve the preliminary literature list, to refine the preliminary literature list, to rank the literature based on the network context, and to organize the network layout.

5.4.1 Network construction and CSML importer

The KEGG reference metabolic network is pre-constructed in CSML format and saved in the database prior to user's request. CSML is a biopathway model exchange format and based on Hybrid Functional Petri net (HFPN) architecture [NSJ+10].

First, all available KEGG reference pathway IDs are obtained. For each of the pathway ID, the related reaction IDs are retrieved from DAWIS-M.D. DAWIS-M.D. contains a table where the reaction IDs of each reference pathway are easily found. For each reaction ID, the corresponding reaction equation and the associated enzyme EC number are retrieved. After the reaction equation is parsed and the involved metabolites are identified, a reaction object is created composed of 3-tuple (EC number, list of substrates, list of products), where the EC number is reaction-specific so that different catalytic reaction of the same EC number could be distinguished. Each reaction object corresponds to a process element in CSML. A CSML file is created for each metabolic reference pathway which is the assembly of the included reactions (Fig. 5.3).

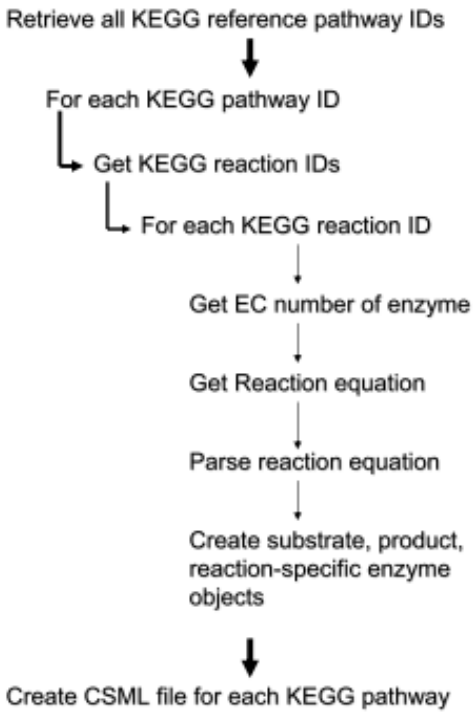


Figure 5.3 The pre-compilation of KEGG reference pathway in CSML format

Besides the metabolic network, ncDocReSy could also display the protein-protein and protein-chemical interactions. The protein-protein interaction is retrieved from the IntAct database through its REST service [ABK+11] in real time. The SwissProt accession number (SPACC) of selected protein and the deactivation of spoke mode are embedded in the request URL which is sent to IntAct. The interaction counterparts for the query SPACC are obtained after parsing the returned document. As for the protein-

chemical interaction, the PubChem IDs of the interacting chemicals are retrieved by querying the pre-constructed MySQL table of protein-chemical interaction.

5.4.2 Document retrieval

The document retrieval function is available for each bioentity in the constructed network. One of the query terms sent to external literature search engines should be the common name of the bioentity. The other query terms are any free text phrase which is provided by the user and whose relationship to the bioentity is of user's interest. Therefore, the synonyms of the selected bioentity have to be retrieved in the first place, according to the annotation from KEGG (for EC number and metabolite) and UniProt (for protein).

ncDocReSy provides the option to choose between three biomedical literature search engines–PubMed, PMC, and BioText. PubMed and PMC are accessed through eUtils which requires the query terms to be encapsulated in a fixed URL syntax. ncDocReSy generates two types of URLs for each bioentity's synonym and the free text query phrases (Fig. 5.4). The first URL includes a search field tag "mh" after the common name of the bioentity. Another URL is constructed for phrase search, in which the common name is enclosed in quotes. The search field tagged URL is first used as the query to PubMed/PMC. If no article is returned, the URL of phrase search is used then. The search result returned from eUtils is in XML format and is parsed to obtain the literature list.

BioText is another full text literature search engine which extends the search fields to the figure legends, table captions, and table contents. An URL where the common name and each of the free text query phrases are surrounded by quote is constructed and submitted to BioText through HTTP Get method. The literature list is extracted from the responding HTML page (Fig. 5.5).

The literature list generated in this step is termed preliminary literature list in which the PMIDs are sorted in numerically descending order. Each preliminary literature list is derived from 4-tuple (bioentity entry, common name, free text query phrases, literature search engine).

KEGG Compound ID: C01179
Synonyms: 3-(4-Hydroxyphenyl)pyruvate; 4-Hydroxyphenylpyruvate; p-Hydroxyphenylpyruvic acid
Free text query: oxidative stress

For PubMed and PMC

MeSH tagged URL:
http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?term=3-(4-Hydroxyphenyl)pyruvate%5Bmesh%5Doxidative%20stress&db=pubmed&retmax=1000000

Phrase search URL:
http://eutils.ncbi.nlm.nih.gov/entrez/eutils/esearch.fcgi?term=%223-(4-Hydroxyphenyl)pyruvate%22oxidative%20stress&db=pubmed&retmax=1000000

For BioText

Phrase search URL:
http://biosearch.berkeley.edu/index.php?q=%223-(4-Hydroxyphenyl)pyruvate%22%22oxidative%20stress%22&sumit=Search&view=abstract&sortedby=rel&r=1000&action=submit_search

Figure 5.4 Examples of two types of URLs for querying PubMed, PMC, and BioText. This example shows the different URLs constructed when the user searches literature on the metabolite node of KEGG Compound ID C01179 and the free text query phrase "oxidative stress".

5.4.3 Literature list refinement

When the preliminary literature list is retrieved through the phrase search of PubMed/PMC or through BioText, the semantic type of the query terms under the context of identified document is not defined. The identified document might just contain the words that match the query terms syntactically but not semantically correct. One example scenario is that the article mentioning "glutamate receptor", which is a protein, may be returned when only article containing "glutamate", which is a metabolite, is expected. ncDocReSy takes advantage of the semantic-type-specific named entity recognition function of whatizit to determine the semantic type of matched terms in the context of the found document (Fig. 5.5).

For each PMID that is returned by BioText or phrase search of PubMed/PMC, the abstract is retrieved first and submitted to whatizit for named entity tagging. Whatizit provides a web service access by which the query text and the choice of pipeline are submitted. The pipeline "whatizitChemicals" and "whatizitSwissprot" are used for recognizing the named metabolite and protein bioentity, respectively. The next step is to create a term frequency summarization for the identified named entity and free text query

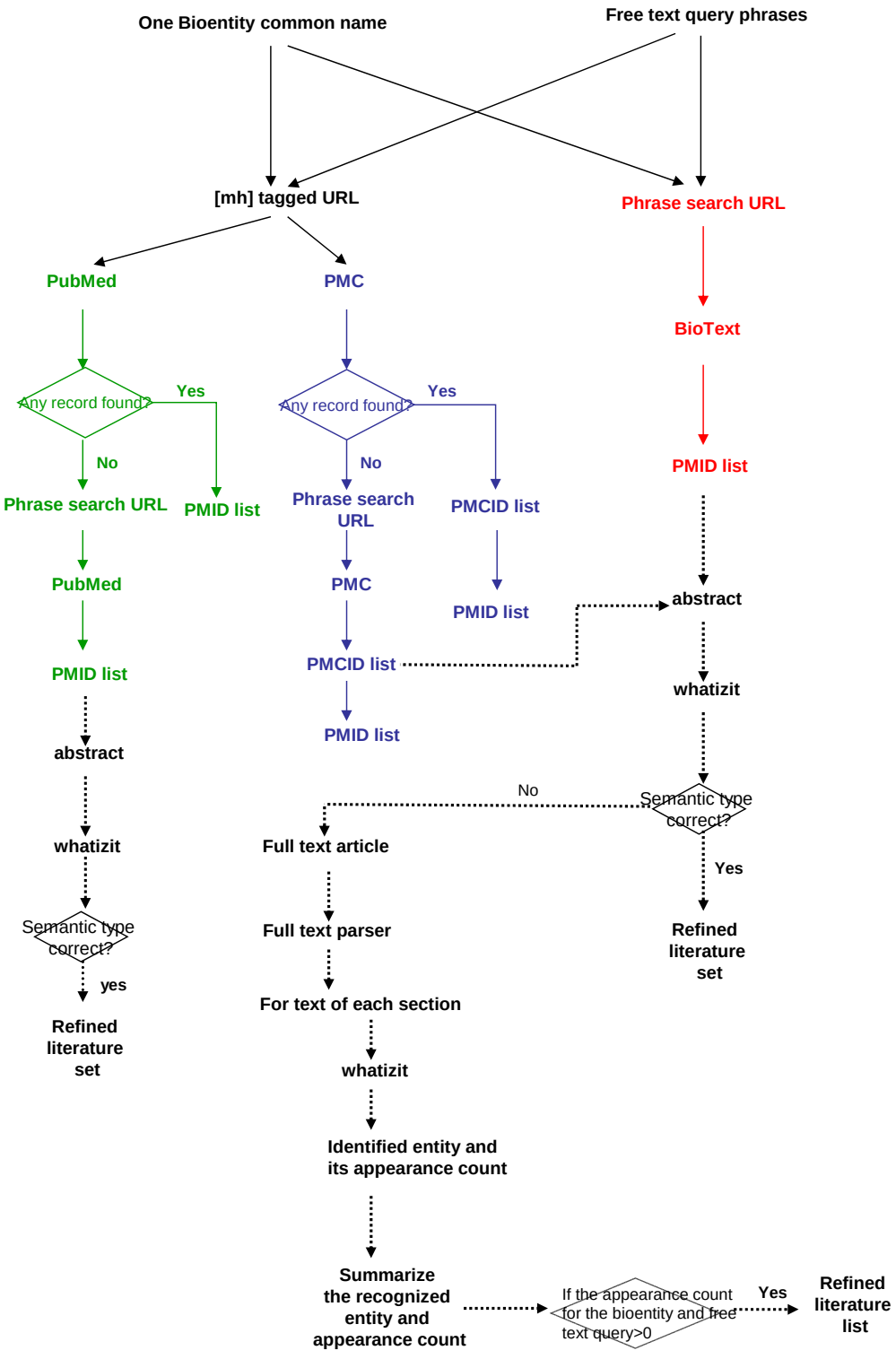


Figure 5.5 Literature search workflow. The literature querying process for PubMed, PubMed Central, and BioText are colored in green, blue, red, respectively. The dashed line represents the workflow of literature list refinement.

phrases. The semantic-typed abstract is parsed, and the frequency of each recognized named entity is recorded. The appearance count of each free text query phrase is searched through pattern match. Only PMID whose abstract contains bioentity name in the correct semantic type and all free text query phrases is saved in the refined document list.

For the document that is retrieved from phrase search of PMC and BioText and fails the abstract inspection, the semantic type checking in the full text article is followed. The full text article is retrieved from PMC. Since PMC uses PMCID as the document identifier, the corresponding PMCID for PMID has to be identified. After the full text article is obtained from PMC through eUtils, it is fed to a full text parser where the text in different sections of the article, such as introduction, methods, conclusion, and discussion, are acquired. The text of each section goes through the term frequency summarization as mentioned above. In the end the term frequency summarization from different sections are merged. Only the document which has the bioentity name in the correct semantic type and the appearance of all the free text query phrases is saved in the refined document list.

5.4.4 Literature summarization

After the document retrieval process, each pairing between different synonym of the bioentity and the free text query string has a preliminary literature list affiliated to. Due to the different synonyms indicating actually the same bioentity, the preliminary literature lists derived from different synonyms of the same bioentity and the same free text query string are merged to one single list. The literature list derived from different 4-tuple (bioentity entry, common name, free text query string, literature search engine) can be summed according to the same 2-tuple (bioentity entry, free text query string), which means the literature lists derived from different common names and literature search engines are merged if the rooted bioentity and the free text query string are the same. The merging process joins the preliminary literature list using PMID as the document identifier. After this process, each free text query string for each bioentity has a joint literature list in which the PMIDs are sorted in numerically descending order. The joint literature list is denoted by $D^{k}_{B\alpha}$, where k denotes certain free text query string that user provides, B_α the bioentity which the joint literature list is affiliated to.

5.4.5 Network-contexted article ranking

After the literature summarization, a network-context ranked literature list (ncRLL) is generated for every bioentity node by iterating all bioentity nodes in the constructed biological network. When the ncRLL is being generated for a certain bioentity node, this

node is called focused bioentity (B_f). The ncRLL of the focused bioentity considers also the joint literature list ($D^k_{B\alpha}$) from neighbor bioentities in the same network besides the one directly affiliated to the focused bioentity. In order to rank the literatures, a network-scaled score is calculated for each document ID based on the distance between the neighbor and focused bioentity as well as the upstream-downstream symmetry (Fig. 5.6).

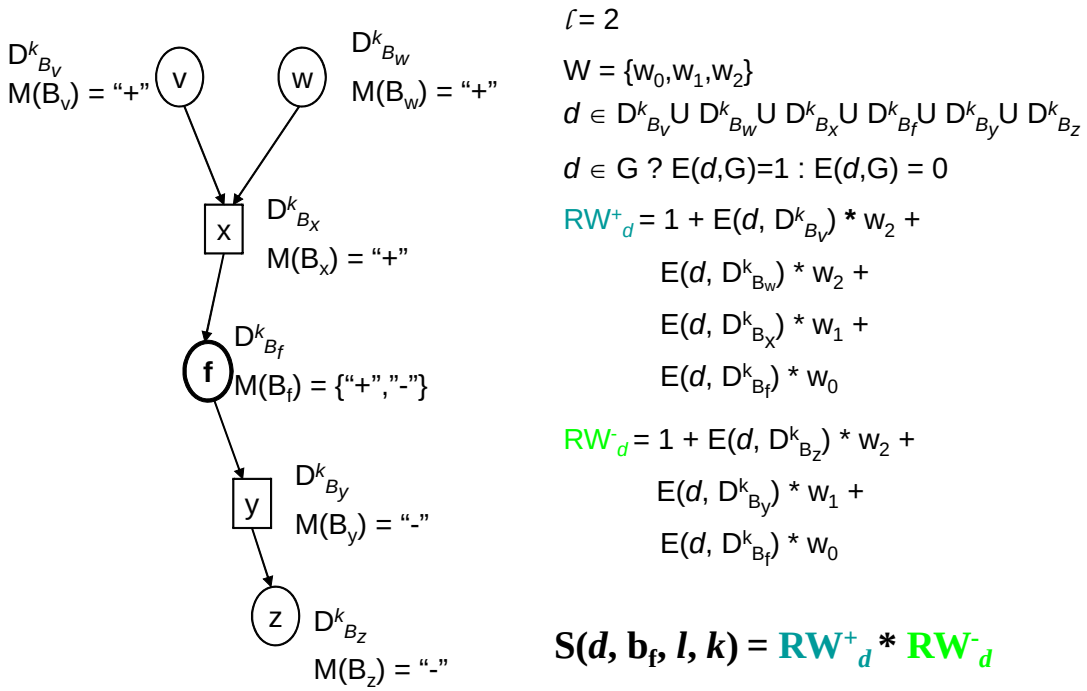


Figure 5.6 Exemplification of network-scaled score calculation for a document d retrieved with free text query string *k* and based on a two-layers network (*l*=2) centered at bioentity f. Abbreviation: W, surrogate ability weight list; RW, region-specific weight; S(*d*, b_f, *l*, *k*), network-scaled score; M, region mark mapping function.

The article ranking process transverses every bioentity node in the displayed network and produces a ncRLL for each bioentity node. The constructed network is split into two regions, plus-signed and minus-signed regions, relative to the focused bioentity node. The plug-signed region contains the upstream reactions and minus-signed region the downstream reactions relative to the focused bioentity node. Thus there are two region-specific weights for each document–one for the plus-signed region and one for the minus-signed region. If the focused bioentity is situated in a directed network, such as the metabolic network or signal transduction pathway, a region mark for the neighbor bioentity is determined in order to incorporate the upstream-downstream symmetry into the calculation of network-scaled score. The direct parent nodes which has an edge

pointing to the focused bioentity has the region mark "+", and the direct child nodes which the focused bioentity has an edge pointing to has the region mark "-", while the focused bioentity bears the region marks both "+" and "-". Any bioentity which connects to a plus-marked bioentity is marked as "+", and the same rule is applied to minus-marked bioentity. The region mark mapping function is represented by $M(B_\alpha)$, where $M(B_\alpha) \subseteq \{+,-\}$ but $M(B_\alpha) \neq \varnothing$, and B_α is any bioentity node in the network.

Before the literature ranking is proceeded, user has to select the number of layers, which is centered at the focused bioentity node and denoted by l, to be included in the creation of the ranked literature list. There is a surrogate ability weight list $W= \{w_0,...,w_l\}$, specifying the weight added to the region-specific weight of the document ID according to the distance between the neighbor node and the focused node. The region-specific weights are calculated for each PMID appearing in certain sub-region and is denoted by RW^m_d, where m represents the sub-region mark, and d represents certain PMID. Every RW^m_d has the base weight "1" and is incremented by w_i if the same PMID is in the joint literature list of certain bioentity B_α , where $d \in D^k_{B\alpha}$, $i = \mathrm{dis}(B_\alpha, B_f)$, $i \leq l$, $M(B_\alpha) = m$.

After the two region-specific weights are acquired, the network-scaled score of document d, based on a sub-network centered at b_f including the neighbor nodes within l layers with the free text query string k, denoted by $S(d, b_f, l, k)$ is calculated by multiplying RW^+_d and RW^-_d together. In the end, each PMID which is obtained from the document retrieval process will have a network-scaled score $S(d, b_f, l, k)$. The PMID is first ranked by $S(d, b_f, l, k)$ and then by PMID in descending numerical order when several PMIDs have obtained identical $S(d, b_f, l, k)$ scores.

5.4.6 Network layout

The constructed bioentity network, which is composed of bioentities, is shown in the main network view window of Cytoscape workspace. The topology of bioentity network mimics the one displayed in KEGG website that only main compounds undergoing chemical transformation in a reaction are shown. User then can select some bioentity nodes for document retrieval, and this will generate the document retrieval network shown together with the bioentity network. Different types of nodes in the bioentity and document retrieval network have their specific appearance, such as proteins in green diamond and metabolites in black circle. Each node has certain attributes associated with it, such as biomolecular database identifier, molecular type, PMID list, and URL link. These attributes can be found in the Data Panel in the Cytoscape workspace. The edges of

the bioentity network are drawn in solid lines and the document retrieval network in dashed lines (Fig. 5.7).

Since the literature search process in ncDocReSy is based on the selected bioentity node, the document retrieval network is rooted on the bioentity node. The child node of the bioentity node in the document retrieval network is the common name of the

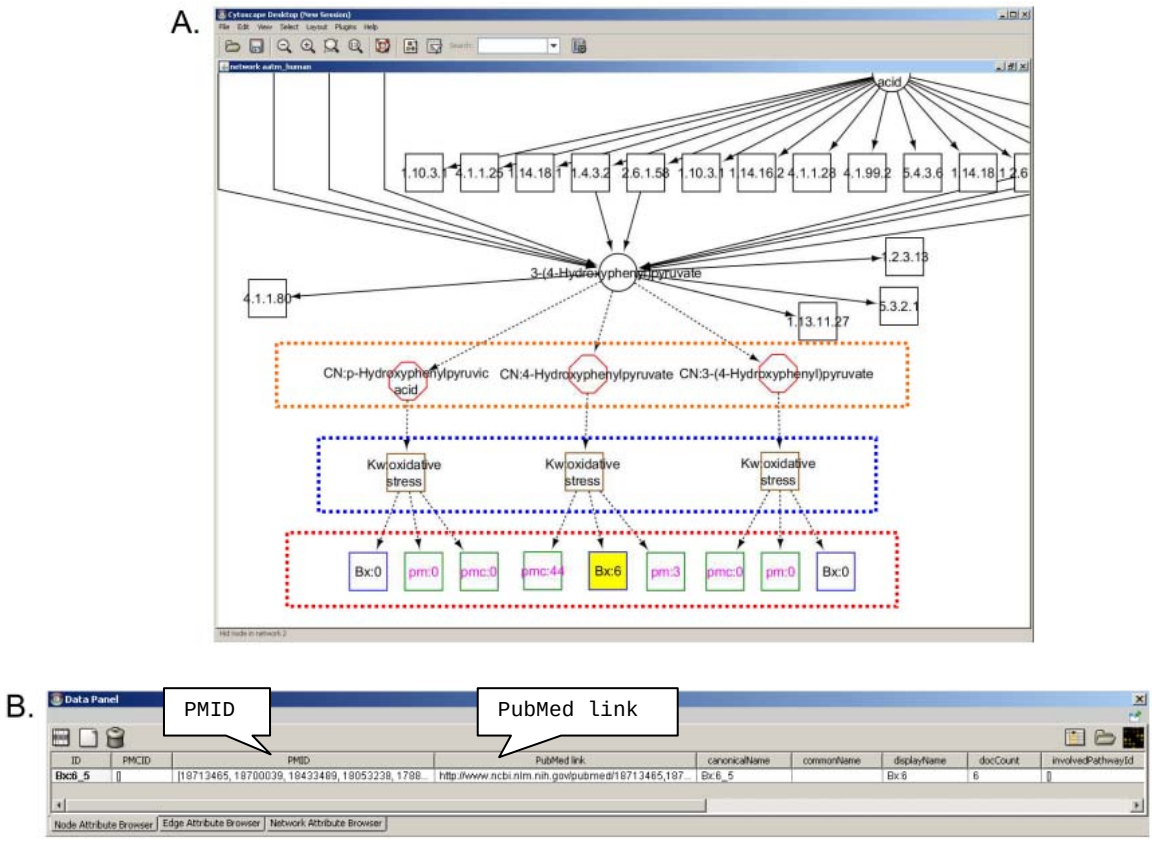


Figure 5.7 The exemplified bioentity network and document retrieval network and the associated node attribute. **A**. The main network view window has been isolated from the Cytoscape workspace. This view shows the partial network of human tyrosine metabolism pathway and the document retrieval network rooted at metabolite node 3-(4-Hydroxyphenyl)pyruvate. The common name nodes are indicated in orange dotted line, free text query nodes in blue, literature search result nodes in red. **B**. The isolated data panel shows the associated attributes of a preliminary literature list node (the square node filled by yellow in **A**), such as PMID and PubMed link.

bioentity. Each common name has a separate node representing it. The child node of the common name node is the node representing use-provided free text query string, and the child node of the free text query node represents the preliminary literature search result from the user-selected literature search engine. The node representing the preliminary literature search result has a “PubMed link” attribute that contains an URL link shown in

the data panel. Once this URL link is clicked, a PubMed webpage containing the preliminary literature list is opened in the web browser (Fig. 5.7).

After literature summarization, ncDocReSy creates a literature summarization node representing the merged literature list and connected directly to the bioentity node. The merged literature list is used later in the generation of ncRLL.

After the ncRLL is generated, two new node attributes — ncRLL and ncRLL_link—are added to each bioentity node and can be viewed in the data panel (Fig. 5.8). The ncRLL_link will show the ncRLL in the web browser once is clicked.

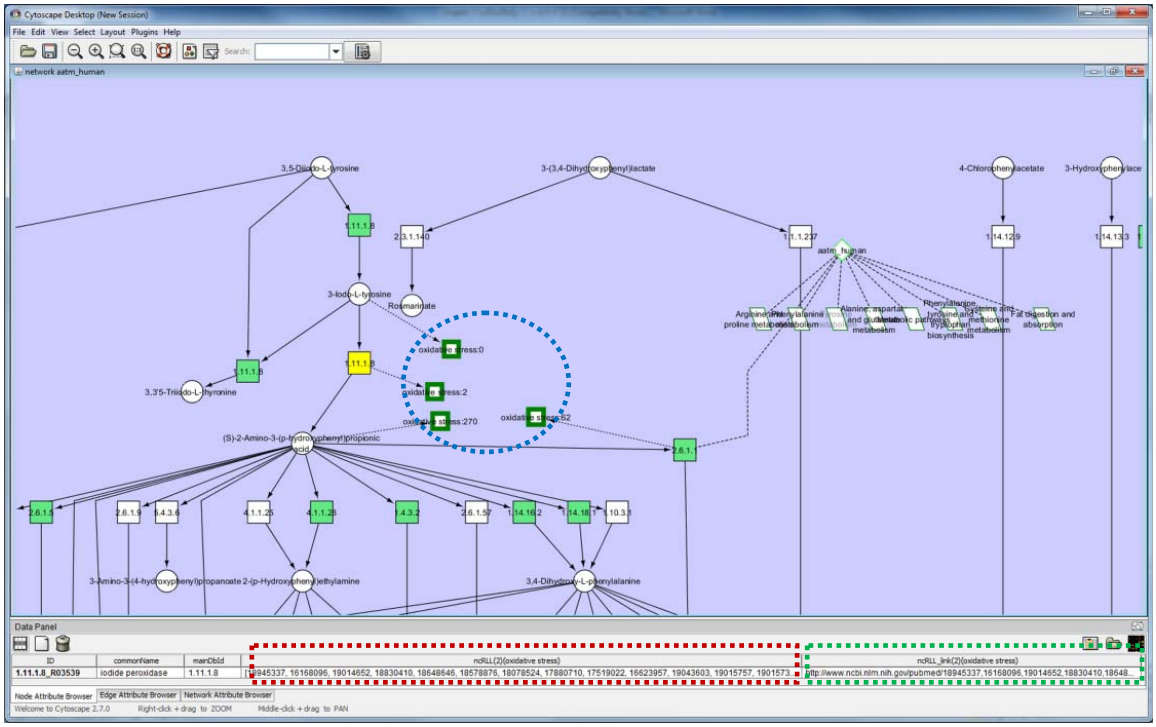


Fig. 5.8 The literature summarization nodes are outlined in blue. The node attributes of the ncRLL and of ncRLL_link are outlined in red and green, respectively.

5.5 Result

ncDocReSy is built on Cytoscape which is a well-known, well-documented, and internationally collaborated platform. Therefore, many functions provided by Cytoscape can be used together with ncDocReSy, such as the network layout algorithm and LinkOut function.

ncDocReSy is a bioentity-focused document retrieval system which can be used when a specific bioentity, such as a particular protein or metabolite, is of user's interest. Before the literature search is carried out, ncDocReSy requests a biological network to be constructed around the central bioentity of interest. After the user input the database identifier of the motivating bioentity, the name of related metabolic pathway will be shown to the user. Then the user can choose multiple pathways, and the content of selected pathway will be displayed inside one network. The enzyme node is represented by the EC number and is reaction-specific, which means the same EC number catalyzing different reaction will have a separate node. As for the metabolite, each metabolite is uniquely represented by one node.

When the user wishes to add the protein-protein or protein-chemical interaction into the network, the EC number is not enough to serve as the bait. The protein members of the EC class have to be obtained prior to the database search. There is a text input area in ncDocReSy where user can specify the NCBI taxonomy number when organism-specific protein members are expected. Once the specific proteins are selected, ncDocReSy can show the interacting proteins or chemicals. Since the number of interacting counterparts is sometimes numerous, ncDocReSy allows the interacting molecules later to be hidden.

In the current version of ncDocReSy, there are three types of bioentity nodes which are allowed to proceed with literature search—enzyme node, protein node, and metabolite node. Before the literature search is started, the user selects the bioentity nodes which they are interested in, gives some free text query string, and select the literature search engines they wish to use. ncDocReSy provides three alternatives of biomedical literature search engines– PubMed, PMC, and BioText. It's possible to select all three alternatives at the same time.

The document retrieval network is shown together with the bioentity network which is constructed before literature search. No matter which types of bioentity nodes user has selected, the different synonyms of the selected bioentities, the free text query string, and the returned literature list from each selected literature search engine are represented by separate nodes. One bioentity can be connected to several synonym nodes, one synonym nodes to several different free text query string node, and one free text query string node to maximum three preliminary literature list nodes. To avoid the exhausted iteration of all synonyms of the bioentity node, the hit-and-move mode allows

the literature search to skip the rest synonyms if any reference has been returned by one of the synonyms from PubMed or PMC. For every preliminary literature list node, it is labeled by the abbreviation of literature search engine ("pm" for PubMed, "pmc" for PMC, "Bx" for BioText) followed by the number of returned articles. If the literature list is retrieved by phrase search, the label of node is pink-colored. The preliminary literature list node has a "PubMed link" attribute that can be found in the data panel of Cytoscape. Once the user clicks on this attribute field, a PubMed web page is shown in the web browser with the literatures ranked in reverse chronological order. User can inspect the literature search network and remove the plausible nodes, such as the synonym node with an ambiguous common name or the preliminary literature list node derived from certain literature search engine.

For the preliminary literature list that is retrieved by phrase search and BioText, user can further filter the preliminary literature list by named entity recognition and semantic type checking. This filtering process can be applied on the preliminary literature list node, and a new node of refined literature list is attached to the preliminary literature list node.

The visually displayed network is used in the further process of literature summarization. Different preliminary literature lists that are retrieved by different synonym and different search engines but the same free text query and the same bioentity can be merged. ncDocReSy generates a literature summarization node for each free text query string that has ever been applied to each bioentity node and attaches this summarization node to the corresponding bioentity node. The literature summarization node has a joint literature list attribute which lists the merged PMIDs. After this process, all synonym nodes, free text query string nodes, and preliminary literature list nodes are hidden. User can press the "show hidden nodes" button to recover the hidden nodes.

Following the literature summarization, the network-context ranked literature list can be generated. User can select any free text query string that has been applied so far, and ncDocReSy will create the network-context ranked literature list concerning the chosen free text query string for each bioentity displayed in the current network. This literature list is sorted by the network-scaled score, and the PubMed link for this list is clickable from the data panel under the "ncRLL_link" attribute.

There are several types of node created by ncDocReSy and several attributes for each node. Depending on the node type, only some attributes hold a value. The data type of the attribute content could be a string or list. The Cytoscape LinkOut function can be used on the content of certain attributes, such as the string content of “mainDbId” attribute and the list content of “PMID” attribute, so that the corresponding records in the webpage-based public database can be displayed in the web browser.

5.6 Summary

A network-contexted document retrieval system—ncDocReSy—that combines biological network and literature search is presented. It extends the literature search beyond the motivating bioentity and allows user to browse relevant literature concerning its associated bioentities. These relationships between the motivating and the associated bioentities are confined to catalytic and physical interaction in the current version of ncDocReSy. In order to retrieve the maximal relevant literature, ncDocReSy incorporates PMC and BioText besides conventional PubMed and takes advantage of the advanced search feature from the literature search engines, such as search field qualifier and phrase search. Since one criterion from ncDocReSy is to retrieve the literature mentioning the intended bioentity semantically, the literature refinement function of ncDocReSy embraces the named entity recognition service from whatizit. Besides accessing these external web-based services for meeting the criteria, a heuristic network-contexted ranking algorithm is devised to incorporate network topology into literature ranking. ncDocReSy is implemented as a Cytoscape plug-in, so that the functions provided by Cytoscape can by used together with ncDocReSy.

Chapter 6

Application

Mitochondria represent cell organelles where citric acid cycle and respiratory electron transport chain catalyze energetic redox reactions at high rates. During the electron transport on the mitochondrial inner membrane, electrons may prematurely escape from the route to oxygen and produce superoxide which is a strong oxidant [VLM+07]. Thus the thioredoxin and glutaredoxin systems in the mitochondrion play an important role in protecting this organelle from oxidation and oxidative stress. Several studies have correlated the oxidation of mitochondrial proteins are associated with aging and neural degenerative disease [MHMF96][BG91][Ferr09]. Thus it is valuable to construct the redox regulatory network in human mitochondria and test its capability of maintaining redox homeostasis, and then infer the biological outcomes by the downstream regulated metabolic network in case of oxidative stress.

As mentioned in chapter 3, the target protein for thioredoxin/glutaredoxin system in plant mitochondria and chloroplasts has been reported in many publications but not in the human mitochondrion yet. Due to the specialty of the proposed biological question, the database methodology is not feasible, and mining of unstructured information in the literature also gives little information. Besides, the biological database doesn't specify the interaction type between two biological molecules, and the text-mining method has the problem of accuracy and uncertainty of semantic relation between text-mined biological concepts. Therefore a bottom-up strategy by implementing a decision rule discriminating the type of cysteine residue is adopted to provide pre-selected candidate proteins. After the pre-selection, looking for the relevant literature supporting further discrimination is

expected. And again, due to the less investigation and specialty of the research topic, it is hard to obtain the directly supportive literature through conventional literature search engines.

This chapter is devoted to the application of ROCD and ncDocReSy on the motivating biological question—pre-selection of thioredoxin/glutaredoxin target protein in human mitochondrion—to overcome the data and literature deficiency problem. ROCD implements the Cysteine Oxidation Prediction Algorithm (COPA) developed by Sanchez *et al.* [SRWM08] to pre-select the potential TTG by detecting the oxidation susceptible cysteine (Fig. 1.1). ncDocReSy is a bioentity-centered document retrieval system which incorporates the topology of biological network to extend the literature search beyond the initial query bioentity. First, ROCD is used on the protein inventory of human mitochondria to pre-select protein candidates. Then some of these candidates are sent to ncDocReSy for demonstration purpose, so that the relevant literature is retrieved to support the user in manual curation of pre-selected proteins from ROCD.

6.1 Pre-selection of target protein by ROCD

According to the discovery from Sanchez *et al.*, the following criteria were used on ROCD web interface to pre-select potential TTGs in human mitochondria: (1) tissue: liver; (2) organelle: mitochondrion; (3) distance of thiol pair: 0~6.2 Angstrom; (4) ASA: 1.3~999 square Angstrom; (5) pK_a: 0~9.05. The result is shown in Appendix E (run on May 21st, 2012).

After ROCD receives the required parameters, the protein set for the human liver mitochondrion is first retrieved according to the annotation in HPRD. This initial protein data set consists of 518 unique SPACC. Since the protein/peptide 3D structure is the essential input for the following calculation, ROCD tries to get the resolved PDB structure or a computationally predicted structure from SwissModel Repository. After this step, 196 SPACCs have experimentally determined PDB structures, and 223 SPACCs have computationally predicted structures. Thus in total, 419 SwissProt accession numbers could be linked to 3D structures described in PDB format. However, the corresponding structure does not always cover the complete peptide sequence of a protein. In many cases, only partial fragments of the mature peptide are experimentally determined or computationally predicted. The coverage of the peptide fragment with 3D coordinates for these 419 SwissProt accession numbers could be as low as 1% (PDB entry 3AGZ for SPACC P11142) of the length of mature peptide. Beside the sequence

coverage issue, there is also the sequence identity issue between the template and the targeted protein. If a computationally predicted structure is used for the calculation of physicochemical properties of the targeted protein, the chosen template might not have a 100% peptide sequence identity to the targeted protein. The sequence identity between the targeted and template proteins ranges from 20% (O60488, Q9BXM7) to 99% (P06576, Q9BUI6).

From the set of 419 SwissProt accession numbers annotated with either experimentally or computationally determined structures, only 309 unique SwissProt accession numbers fulfilled the pre-selection criteria adopted from Sanchez *et al*. The well-known redox transmitters (Trx2, Grx2, Trxr2) and redox sensors (Prx5) of RRN in the human mitochondrion were included in the set of 309 SwissProt accession numbers.

ROCD is accessible through its web interface, and the web page of the pre-selection result provides hyperlinks to other public molecular databases for certain data field. From the web page result, Cytoscape can be initiated through Java Web Start and preloaded with ncDocReSy, so that user can proceed with literature search after ROCD's pre-selection. The pre-selection result is downloadable in the XML format, and the XML schema of the result file could be found on the ROCD website. The XML-formatted result file and its schema information allow easy parsing of the result file through Java Architecture for XML Binding (JAXB). A tab-delimited format of the result file is also provided for direct loading into Microsoft Excel, and some data manipulation processes, such as sorting and filtering, can be applied from Microsoft Excel.

6.2 Literature search by ncDocReSy

After the pre-selection by ROCD, user might wish to look for relevant publication concerning the pre-selection result and carry out manual curation. This section is devoted to the application of ncDocReSy on the pre-selection result.

6.2.1 Generation of network context-ranked literature list

The first application will demonstrate the procedures leading to the generation of network context-ranked literature list using one of the pre-selected proteins, the aspartate aminotransferase (SPID: AATM_HUMAN), as an example. Before the metabolic network involving aspartate aminotransferase is displayed, ncDocReSy first searches for the metabolic pathways to which the aspartate aminotransferase is associated and

displays the name of each associated pathway as a node in the main network panel (Fig. 6.1). At this stage user can start to apply literature search function on the pathway name node. Figure 6.1 also shows the literature search result using the query terms “phenylalanine metabolism” and “human thioredoxin-2” and the literature search engine PubMed. The literature search returns three articles, and by clicking on the “PubMed link” attribute field, the PubMed records of these three articles are shown in the web browser.

Besides doing literature search on the pathway name nodes, the content of the chosen pathway can be displayed. Figure 6.2 gives the snapshot of phenylalanine metabolic pathway, where aspartate aminotransferase is associated with EC number 2.6.1.1 and carries out the KEGG reaction R00694. When the user requests the display of the pathway contents, ncDocReSy downloads the pre-compiled CSML file from our server and saves it in a local temporary file. Due to the CSML format of the downloaded file, the user can load this temporary file into CellIllustrator (Fig. 6.3).

If a pathway contains too many metabolites and enzymes, the whole network would be too complicated to be comprehended. ncDocReSy mimics the network topology of KEGG map and only displays the major metabolite which characterizes the reaction and hides the presumably less important ones. ncDocReSy also enables user to crop the network, so that only the nodes that are within a defined distance from a selected node are retained in the network view.

The literature search process starts from the selection of bioentity nodes which interest the user. Besides selection of interesting bioentity nodes, user has to give the free text query terms whose relationship to the selected bioentities are user’s interest. Figure 6.4 shows that the substrate and product nodes of enzyme EC 2.6.1.1 are selected, and “human thioredoxin-2” is used as the extra free text query terms. Only PubMed is selected as the literature search engine in this example.

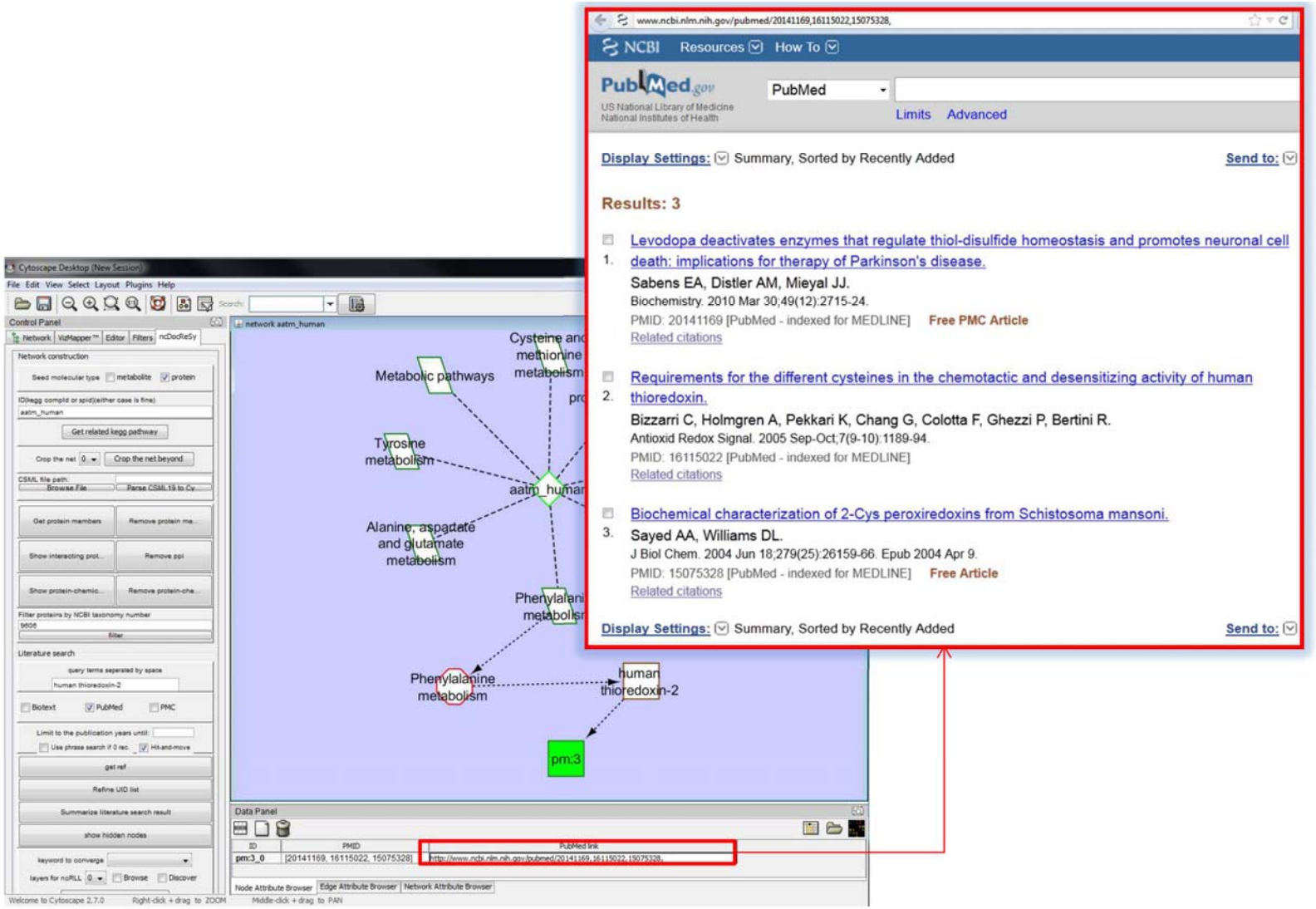


Figure 6.1 The associated metabolic pathway for UniProt entry AATM_HUMAN (human aspartate aminotransferase) and the literature search result on the pathway name nodes. The UniProt entry AATM_HUMAN is associated with 7 metabolic pathways as displayed in the main network view window (the "Metabolic pathways" is the generic name for any metabolic pathway). The string "Phenylalanine metabolism" and the free text query "human thioredoxin-2" are sent to PubMed, and three publications are retrieved. These three publications are displayed in the web browser by clicking the URL under the PubMed_link attribute in the data panel.

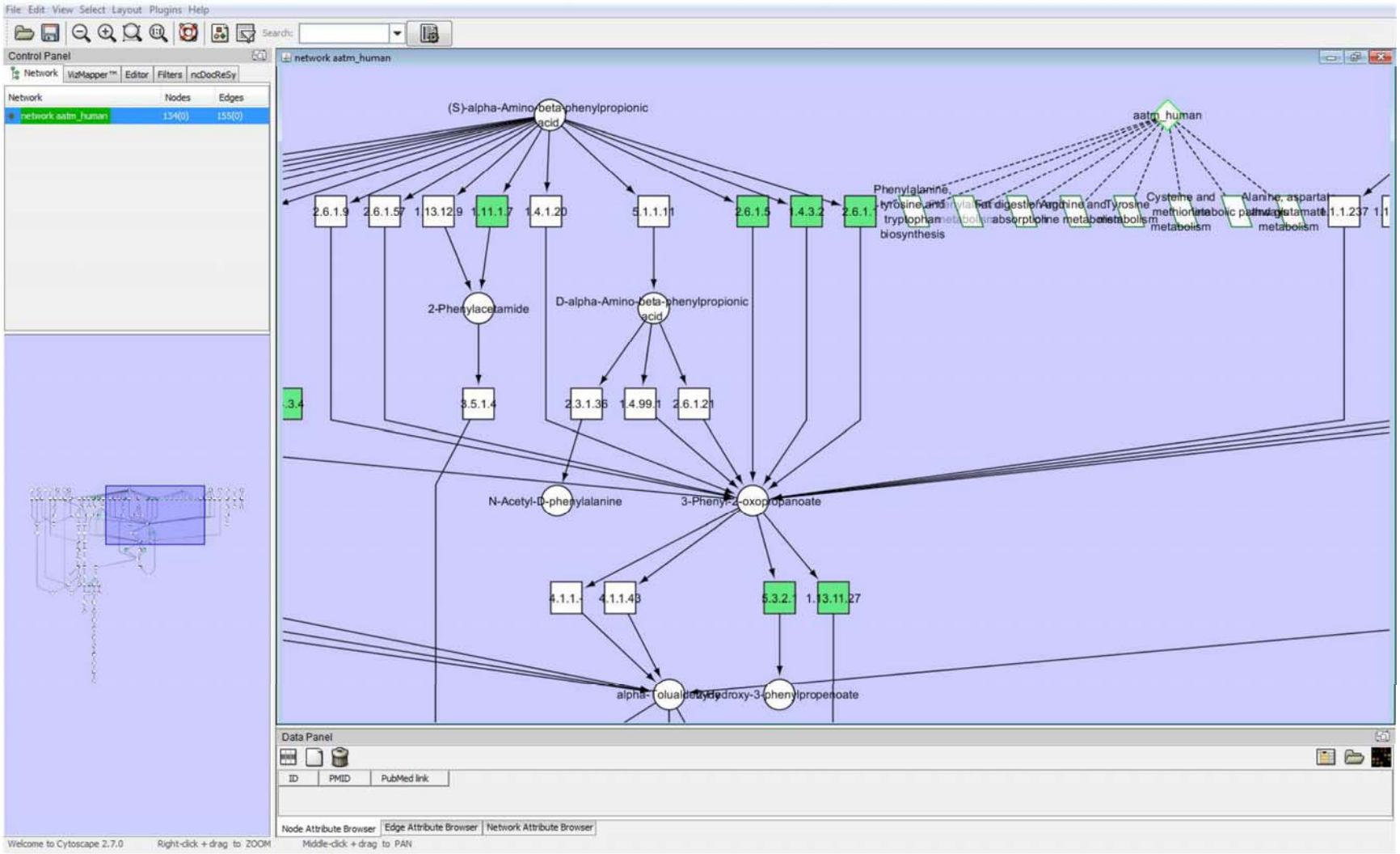


Figure 6.2 The network content of phenylalanine metabolic pathway. The whole phenylalanine pathway can be seen in the network overview pane (shown at the bottom left), and an enlarged area is shown in the main network view window.

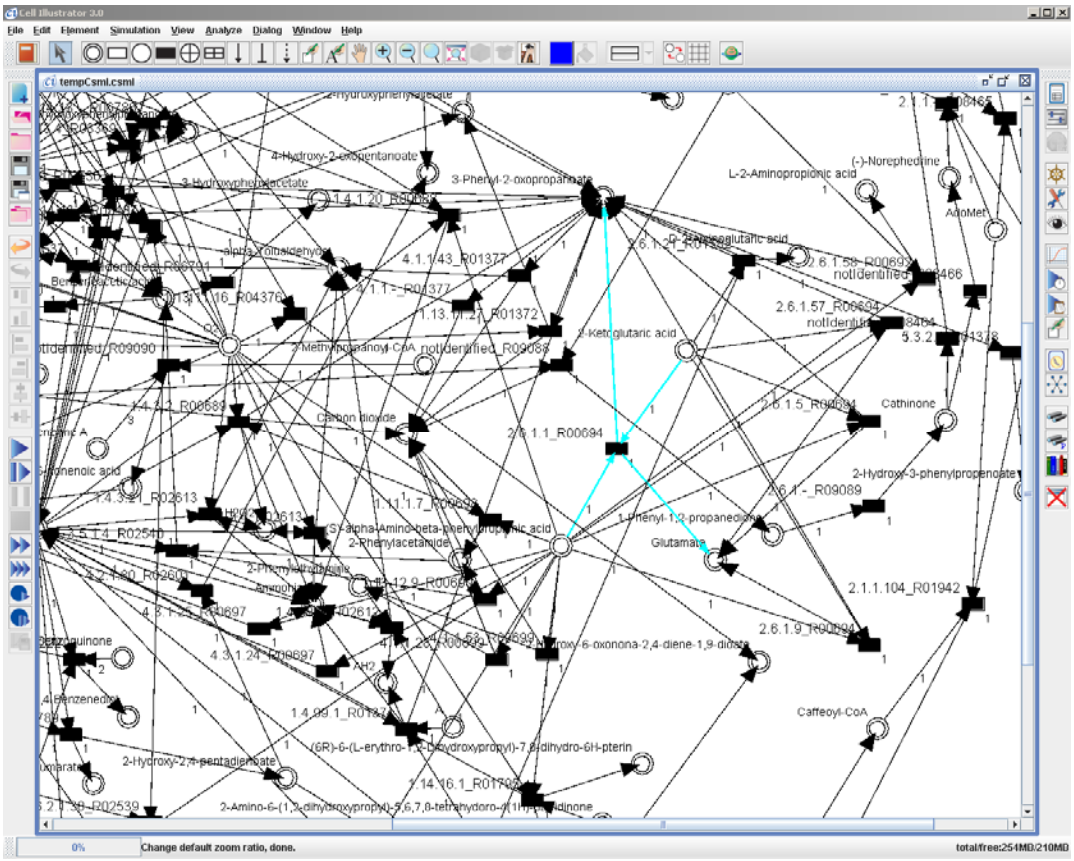


Figure 6.3 The downloaded CSML file of the phenylalanine metabolic pathway is loaded into CellIllustrator. The light blue-colored edges show the linkage between EC 2.6.1.1 and its reactants.

ncDocReSy tries to retrieve the different synonym for the selected bioentities before sending query to the literature search engine. Figure 6.5 shows that 5 synonyms are associated with 3-Phenyl-2-oxopropanoate (KEGG Compound ID: C00166) and more than 20 synonyms with EC 2.6.1.1. When we check the preliminary literature returned with the synonyms of EC 2.6.1.1, most of the synonyms generate the same literature set. The activation of hit-and-move mode can alleviate the exhausted iteration of all synonyms. Figure 6.4 actually shows the literature search on bioentity node EC 2.6.1.1 when the “hit-and-move” mode is on. A preliminary literature list with 3 articles is retrieved by the synonym “aspartate transaminase” and “human thioredoxin-2”.

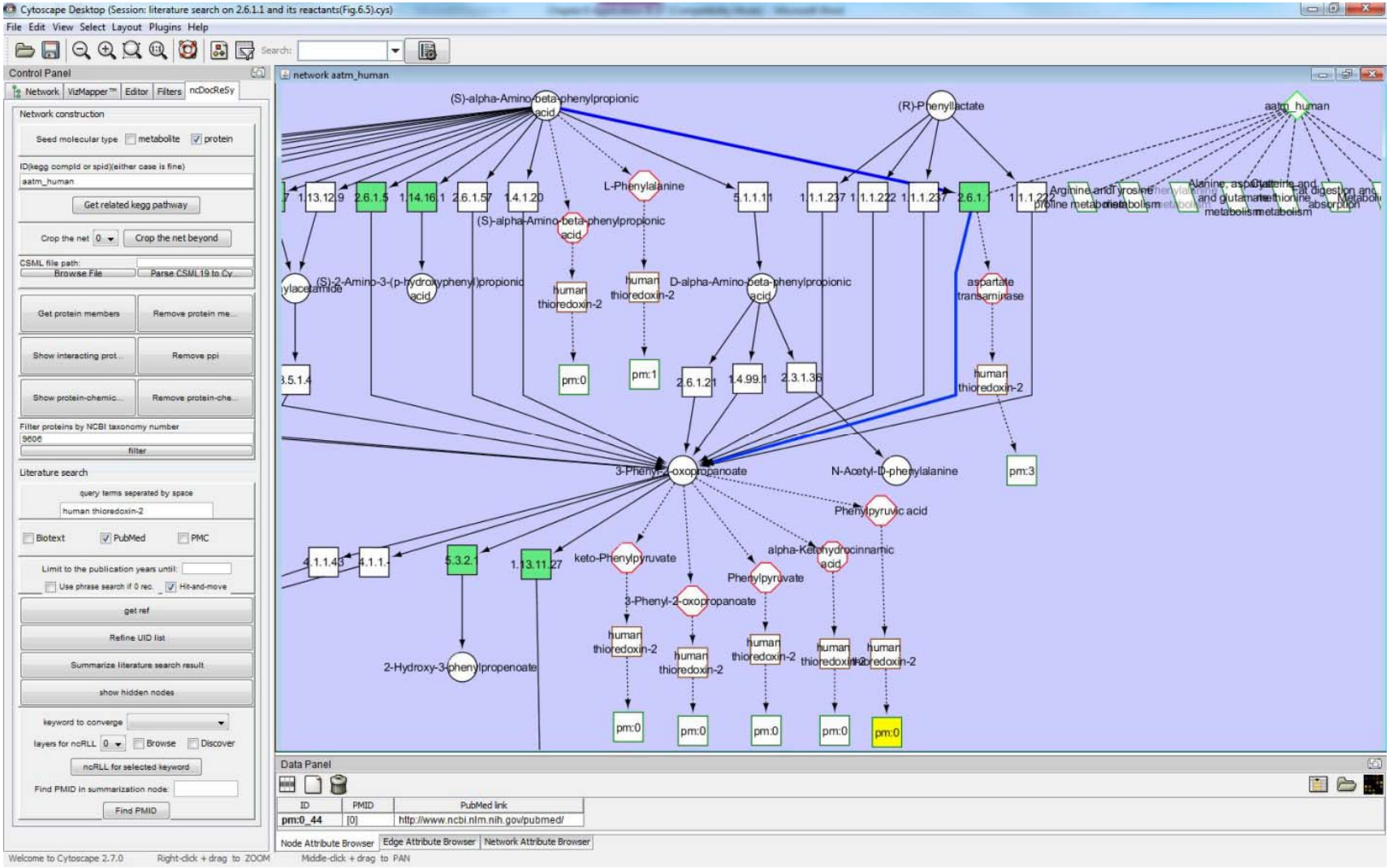


Figure 6.4 Literature search applied on the node of EC number 2.6.1.1 and also its substrate and product nodes. The blue-colored edges show the linkage between EC 2.6.1.1 and its reactants. The nodes EC 2.6.1.1 and its reactants are selected and are sent to PubMed with the free text query “human thioredoxin-2”. The literature search shows that there is no publication relating “3-Phenyl-2-oxopropanoate” to “human thioredoxin-2” but one publication relating “(S)-alpha-Amino-beta-phenylpropionic acid” to “human thioredoxin-2” and three publications relating “aspartate transaminase” to “human thioredoxin-2”.

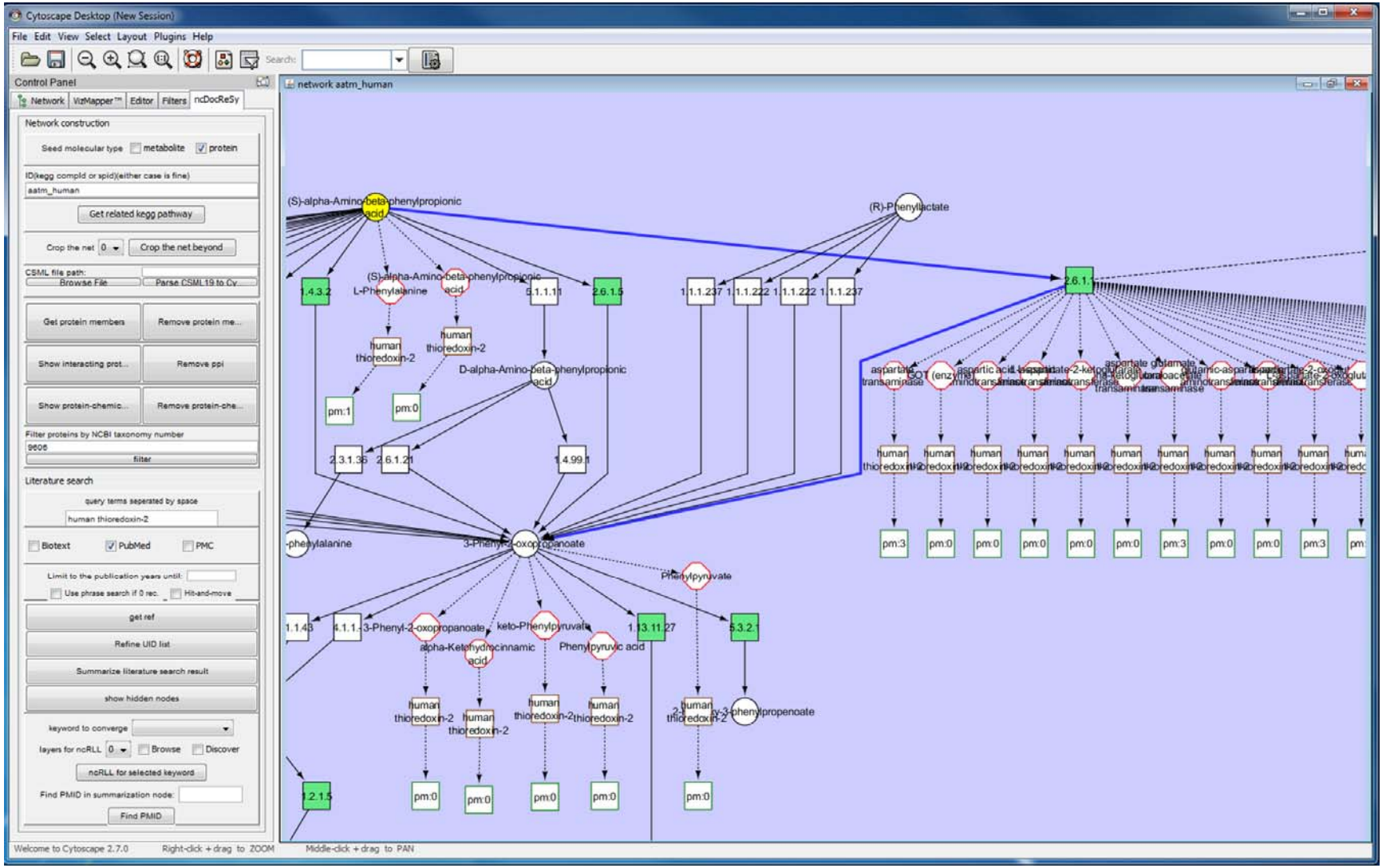


Figure 6.5 Synonyms associated with the EC 2.6.1.1 and its reactants. There are as many as 20 more synonyms for EC 2.6.1.1, two synonyms for (S)-alpha-Amino-beta-phenylpropionic acid (KEGG Compound ID: C00079) and five synonyms for 3-Phenyl-2-oxopropanoate (KEGG Compound ID: C00166)

After completing the literature search with the different synonyms of the selected bioentity nodes, the obtained preliminary literature lists can be merged based on the same bioentity and free text query string. The merged literature list is represented by a single summarization node as shown in Fig. 6.6. After the literature summarization, the network-context ranked literature list can be generated and displayed through the URL in the "ncRLL_link" attribute field of each bioentity node. Figure 6.7 shows the network-context ranked literature list considering the preliminary literature lists within one extension level.

6.2.2 Capability of network context-ranked literature list

The second application will show the capability of network context-ranked literature list using another pre-selected protein, namely the glycine amininotransferase (SPID: GATM_HUMAN), as an example.

After the relevant metabolic pathways for GATM_HUMAN are obtained, two relevant pathways—Glycine, serine and threonine metabolism as well as Arginine and proline metabolism—are extended. GATM_HUMAN is associated to EC 2.1.4.1 in the metabolic pathways. Since the enzyme node is reaction-specific, EC 2.1.4.1 are represented by two separate nodes for two reactions— KEGG reaction R01989 and R00565. The network is cropped by four layers centered at enzyme node 2.1.4.1_R00565, and ncRLL is obtained by the following steps: (1) The free text query term "human thioredoxin-2" is used. (2) PubMed is chosen. (3) The hit-and-move mode is used. (4) Document retrieval is applied on all the enzyme and metabolite nodes displayed in ncDocReSy. (5) The literature lists obtained by different synonyms are summarized. (6) ncRLL is obtained by considering the bioentity nodes within four layers in the neighborhood. After these steps, each bioentity node displayed in the ncDocReSy has a ncRLL affiliated to it concerning the free text query "human thioredoxin-2". As Fig. 6.8 shows, the input GATM_HUMAN belongs to EC 2.1.4.1. The ncRLL for EC 2.1.4.1 can be seen under the "ncRLL(4)(human thioredoxin-2)" attribute of EC 2.1.4.1 node in the data panel and can be displayed in the web browser by clicking the field under "ncRLL_link(4)(human thioredoxin-2)" attribute.

The document retrieval on the bioentity node "2.1.4.1" and with the free text query "human thioredoxin-2" retrieves no relevant literature list, but a ncRLL can be obtained through network-contexted literature search (Fig. 6.8). The first-ranked journal article in the ncRLL is PMID 20306272. When we check where this article comes from,

we find it is affiliated to metabolite node (S)-2-amino-5-guanidinovaleric acid, which is a substrate of EC 2.1.4.1, and also to other nodes, such as EC 1.14.13.39 and nitric oxide (Fig 6.9). These three nodes are all located in the upstream region of EC 2.1.4.1. The same region is also where the 2nd- to 10th- ranked journal articles are affiliated to. The 11th ranked article is PMID 12855383. This article is affiliated to the metabolite node (S)-2,5-diaminopentanoate, which is a product of EC 2.1.4.1, and also to other nodes like 1,4-butanediamine, glutamate and EC 1.4.1.2 in the downstream region (Fig. 6.10). Judging from these observations, the user can know that there are published articles that link human thioredoxin-2 to the substrate and product of the input protein GATM_HUMAN as well as other upstream and downstream bioentities in spite of no relevant article directly reporting the involvement of the input protein. ncDocReSy gives the user an overview of the published journal articles in the biological network context.

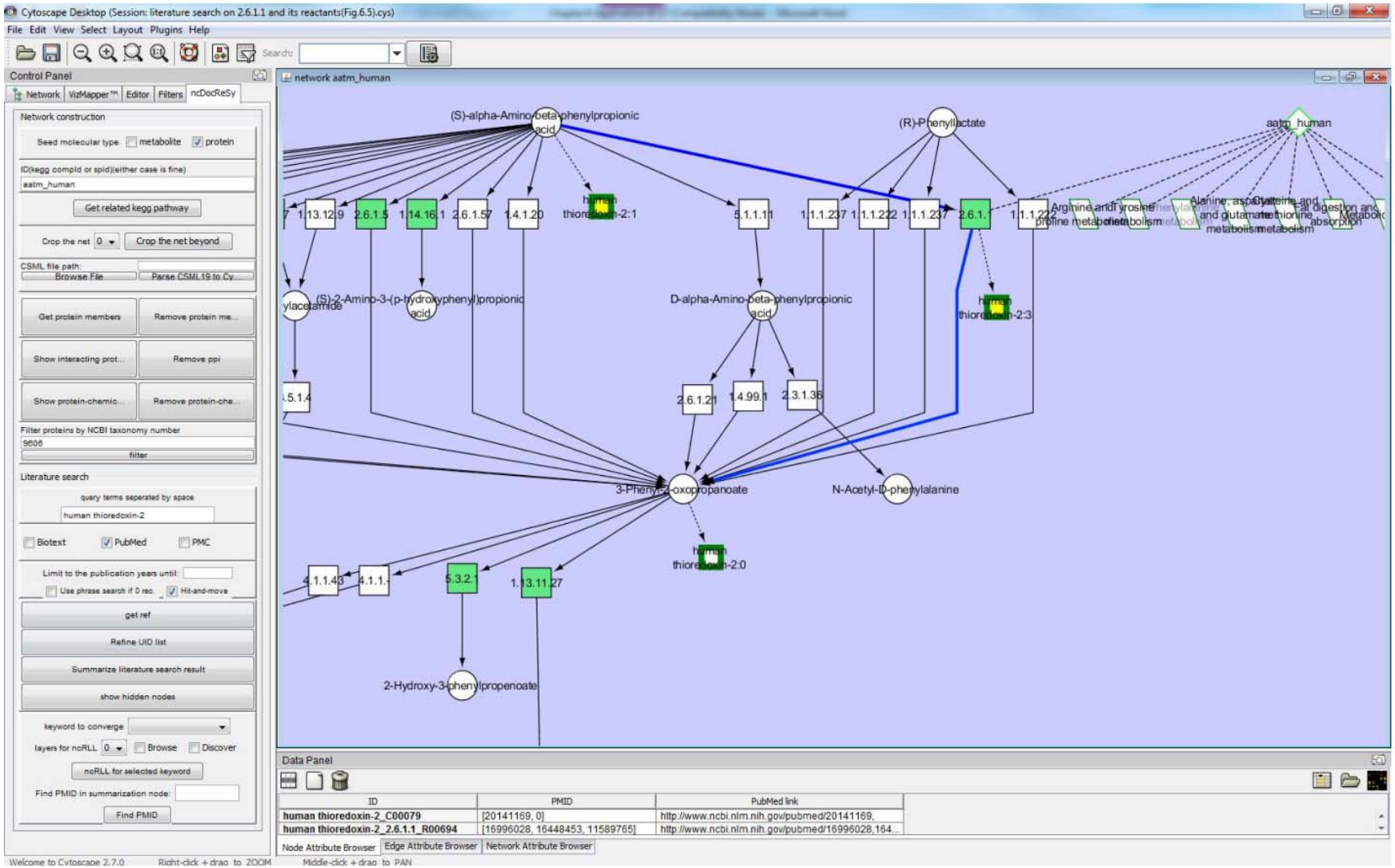


Figure 6.6 Literature summarization. The blue edges show the reaction of EC 2.6.1.1. The summarization node is shown in thickened square in green.

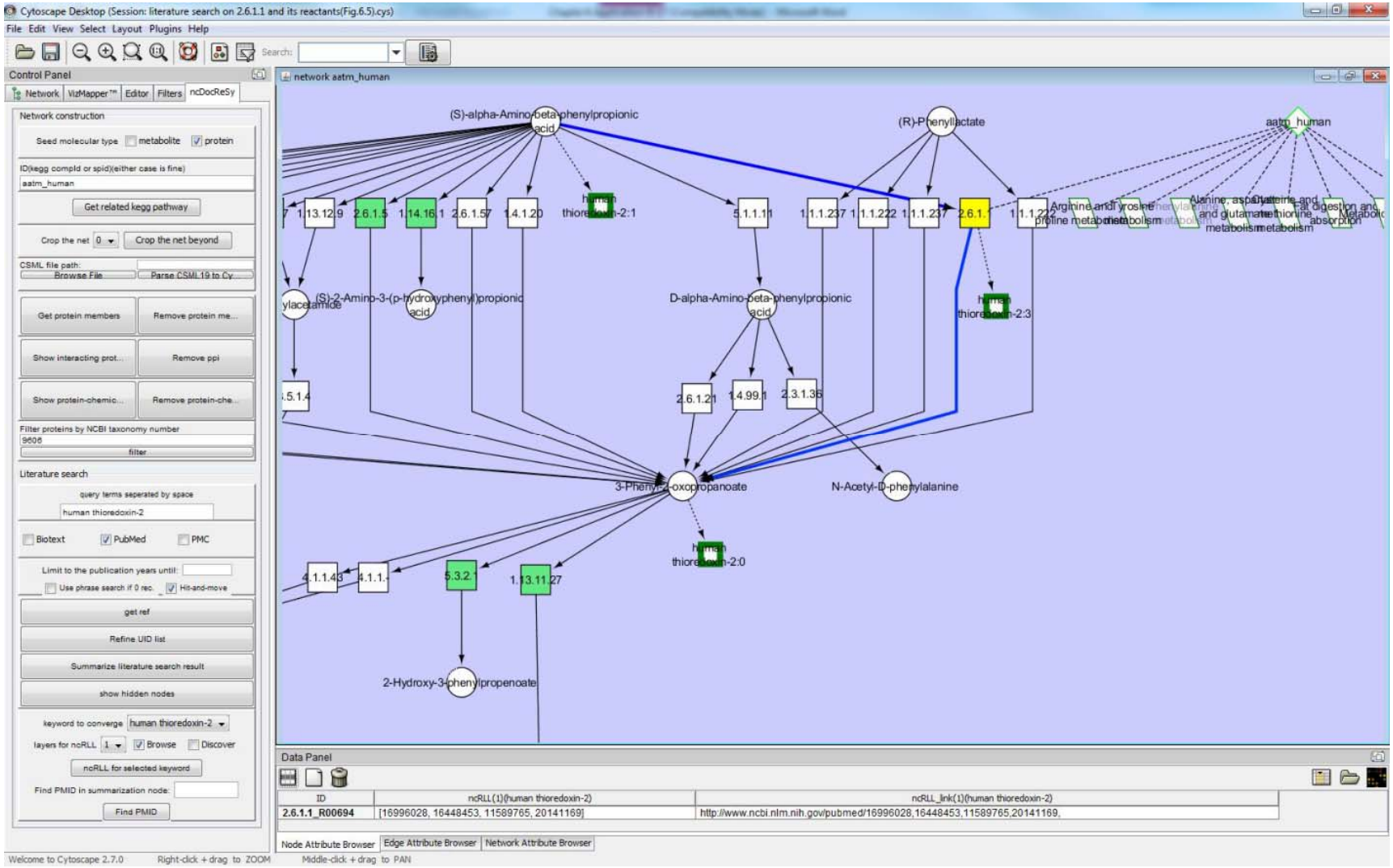


Figure 6.7 Network-context ranked literature list. The blue edges show the reaction of EC 2.6.1.1. The attributes of the yellow color-filled node (2.6.1.1) are shown in the data panel. The ncRLL can be viewed in the web browser by clicking the data field of “ncRLL_link”

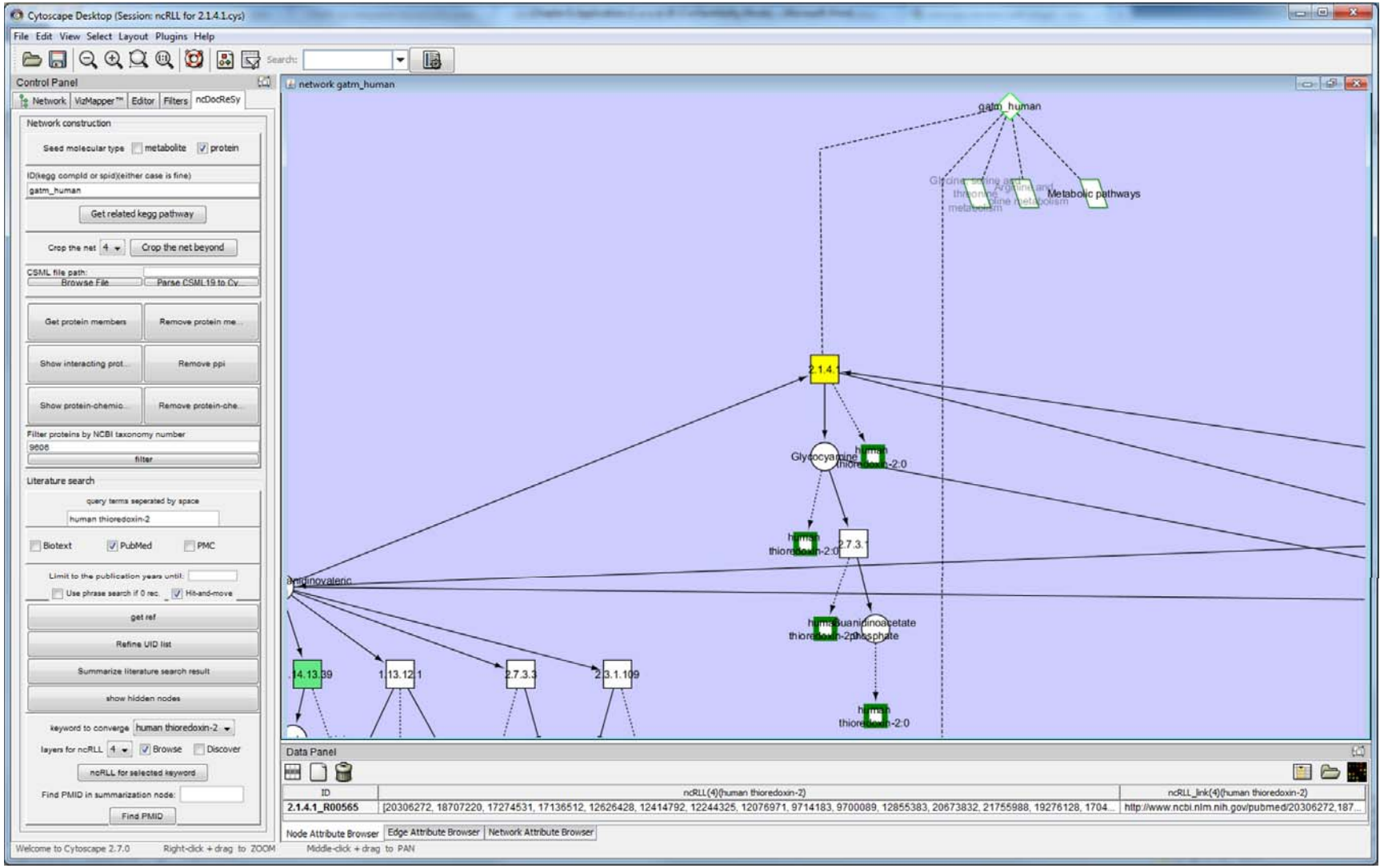


Figure 6.8 Network context-ranked literature list (ncRLL) for EC number 2.1.4.1. The parameters for obtaining ncRLL can be seen in the control panel of Cytoscape workspace. The EC 2.1.4.1 node is selected in this figure, and the ncRLL can be seen in the data panel.

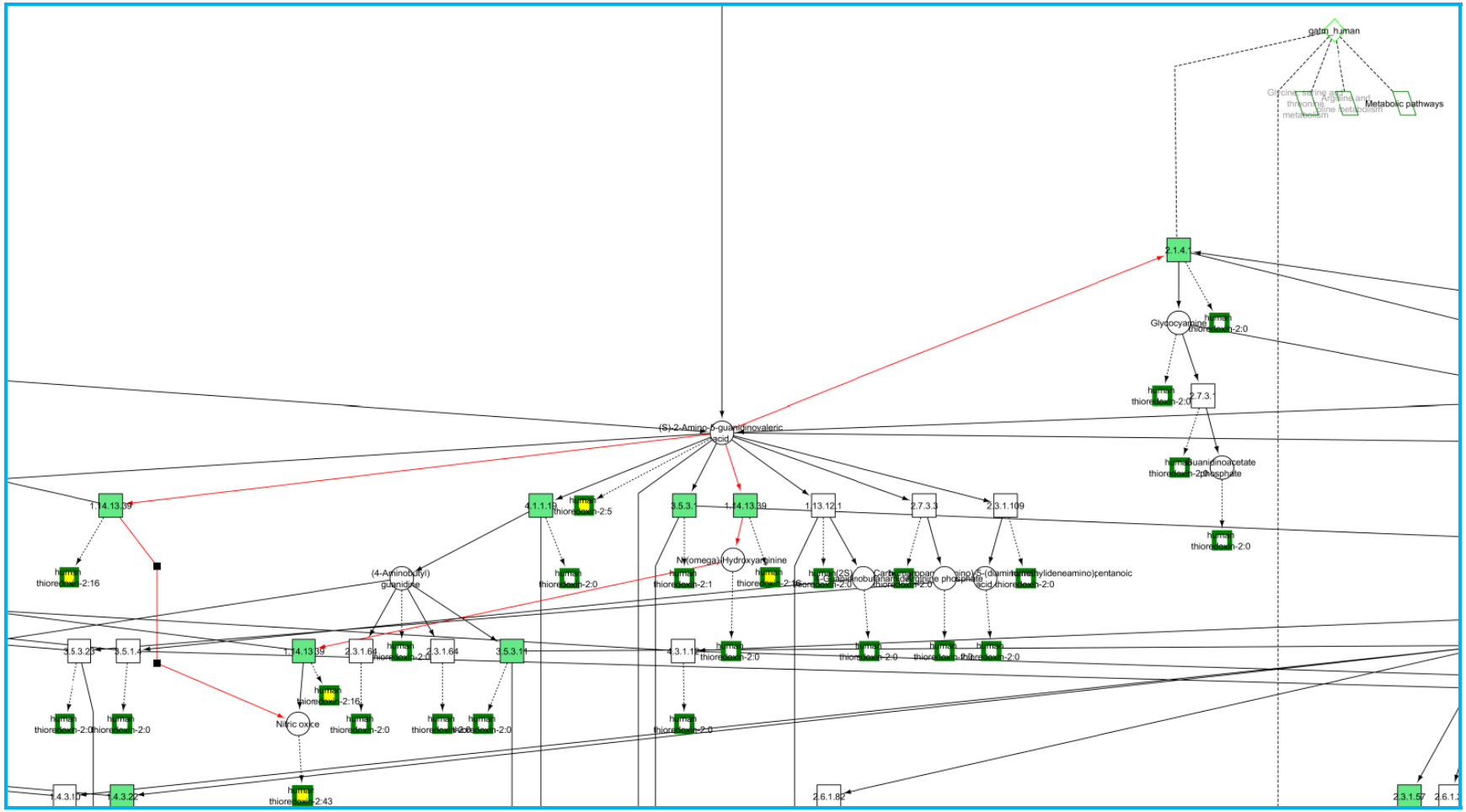


Figure 6.9 The affiliated bioentity nodes for the first-ranked article PMID 20306272 in the ncRLL of EC 2.1.4.1. Only the network view is extracted from Cytoscape and shown here. The yellow-filled literature summarization nodes are where PMID 20306272 is in, and the red edges show the linkage between where PMID 20306272 is derived from and the query protein GATM_HUMAN (EC 2.1.4.1). The yellow-filled literature summarization nodes are located in the upstream region of EC 2.1.4.1.

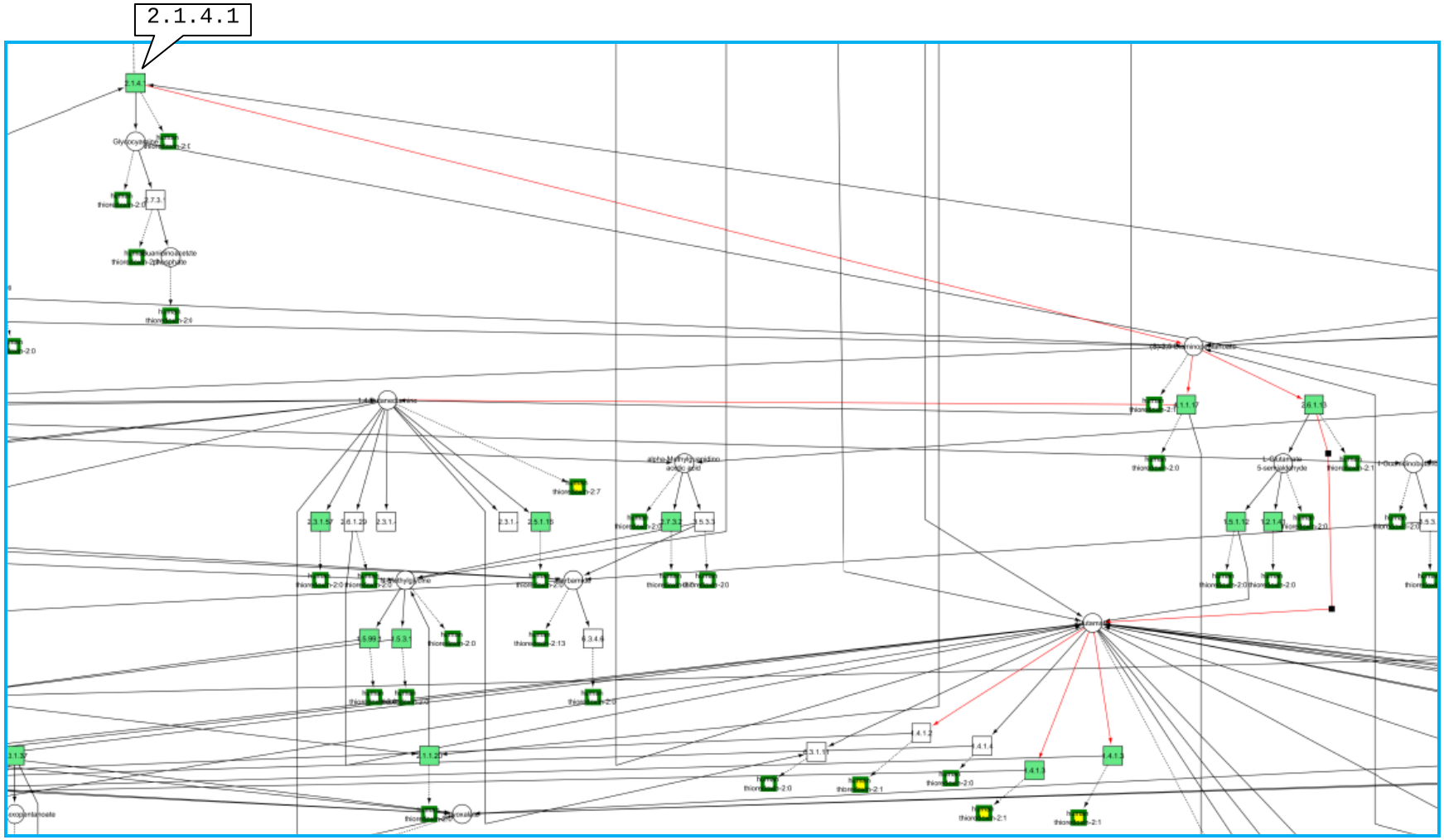


Figure 6.10 The affiliated bioentity nodes of the 11th-ranked article PMID 12855383 in the ncRLL of EC 2.1.4.1. The yellow-filled literature summarization nodes are where PMID 12855383 is in, and the red edges show the linkage between where PMID 12855383 is derived from and the query protein GATM_HUMAN (EC 2.1.4.1). The yellow-filled literature summarization nodes are located in the downstream region of EC 2.1.4.1.

6.2.3 Capability of full-text search and literature refinement

Another pre-selected protein is lanosterol synthase (SPID: ERG7_HUMAN). The third application chooses this protein for easy demonstration of the capability of the full-text literature search and the literature refinement function. Lanosterol synthase is associated with the EC number 5.4.99.7 as shown in Fig. 6.11. This node of EC number 5.4.99.7 and its direct substrate and product nodes are selected for preliminary and refined literature search (Fig. 6.11). One of the synonyms for EC number 5.4.99.7 is “lanosterol synthase”. The query terms “lanosterol synthase” and “thioredoxin” obtain no literature from PubMed but 3 articles from BioText and 13 articles from PMC. Since the preliminary literature list from the query term “lanosterol synthase” and “thioredoxin” is pink-colored, it means the phrase search mode has been applied to obtain the preliminary literature list. Therefore, the result solely indicates the appearance of “lanosterol synthase” and “thioredoxin” in the full text but does not identify the semantic type of them in the context of the article. To determine the correct semantic type of the matched terms, the refinement process is applied to the preliminary literature lists which are obtained by PMC and BioText. The full text of each article in the preliminary literature list is submitted to whatizit for semantic type tagging. Only the article which describes the queried bioentity in the expected semantic type, such as protein or chemical, is presented to the user. The refinement result shows that only one full-text article contains terms matching the query terms and being in the correct semantic type in the preliminary literature list derived from BioText. And two articles have passed the refinement process in the preliminary literature list derived from PMC.

The products of EC number 5.4.99.7 is “4,4',14 alpha-trimethyl-5 alpha-cholesta-8,24-dien-3 beta-ol”. When the literature search is done with one of its synonyms, lanosterol, and “thioredoxin”, there are preliminary literature returned from BioText and PMC. But after checking the semantic type, no article passes the refinement process. The reason of this outcome is that the metabolite name “lanosterol” occurs in the preliminary literature under incorrect context, such as “lanosterol synthase” or “lanosterol 14-alpha demethylase”, which refers to a protein name.

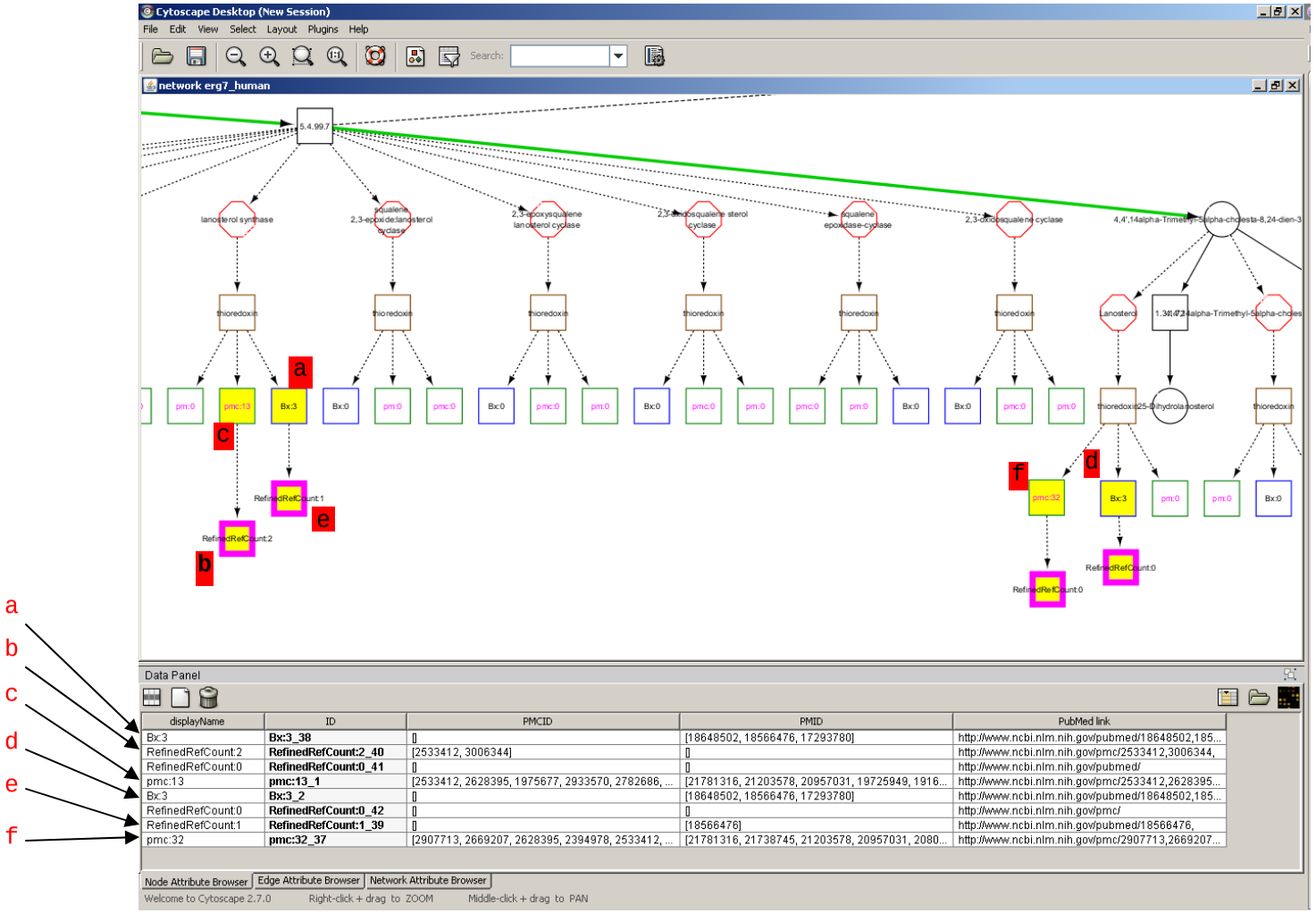


Figure 6.11 Capability of full-text literature search and the literature refinement function of ncDocReSy. The query with lanosterol synthase and thioredoxin retrieves no article from PubMed but 13 from PMC and 3 from BioText. After the literature list refinement, 2 PMC articles (label b) and 1 BioText article (label e) mention lanosterol synthase in the expected semantic type (as a protein). The similar outcome applies to lanosterol, but no article passes the semantic type check.

Chapter 7

Discussion

In chapter 6, two tools–ROCD and ncDocReSy–developed in this thesis work have been applied to the motivating biological question, namely the pre-selection of potential target proteins for thioredoxin/glutaredoxin. This chapter discusses the results obtained in chapter 6 and attempts a general discussion about each tool.

7.1 Discussion about Reversibly Oxidized Cysteine Detector (ROCD)

ROCD, the first tool, pre-selected 309 human mitochondrial proteins expressed in the liver as the potential candidates of thioredoxin- and glutaredoxin- mediated dithiol-disulfide transition for further validation. ROCD is based on protein-specific characteristics which allow for the calculation of three physicochemical properties. Even a structure for a structurally unresolved protein can be computationally modeled based on homologous structures deposited in the SwissModel repository, the model retrieved often only covers part of the protein. Particularly the N- and C-terminal domains often remain unpredictable. Therefore, it is impossible to detect the oxidation susceptibility of all cysteine residues provided some cysteines are absent from the model. The alternative for improving the structural model is to do manual model building with more user intervention. The present version of ROCD only retrieves the structural model generated from the automated modeling procedure of SwissModel repository for structurally unresolved protein. One future extension of ROCD is to allow users to provide their

manually built models, so that the availability and coverage of the protein structure are enhanced.

Due to their high reactivity with other thiols and compounds, cysteinyl residues exist as one of the least abundant amino acids and usually are conserved in functionally important sites. The function of cysteinyl thiols can be roughly classified as: (I) catalytic redox-active Cys, (II) regulatory Cys, (III) structural Cys, (IV) metal-coordinating Cys, (V) catalytic non-redox Cys, and (VI) posttranslationally modifiable Cys [FMG08]. However, in some cases a Cys can fall into more than one of the above categories. The catalytic redox active Cys residues are present in the active sites of thiol oxidoreductases and are directly involved in catalysis and are highly conserved in protein sequences. Glutathione reductase, Grx, thioredoxin reductase, and Trx in the redox regulatory network bear this type of cysteines. Several computational methods have been introduced in section 3.1.3.2 for prediction of thiol oxidoreductases and this type of Cys. The structural Cys residues are involved in the disulfide bond formation which is a major mechanism of protein structure stabilization. This class of Cys can be computationally predicted through the analysis of distance between any two sulfur atoms of Cys residues in the same protein. The metal-coordinating Cys residues are highly conserved and frequently present in the form of CxxC motif, which is also the typical motif of thiol oxidoreductases. The catalytic non-redox Cys residues participate in catalysis but don't change their redox state in the reaction and are highly conserved. This class of Cys could be predicted by sequence or structure similarity to functionally characterized proteins. The regulatory Cys residues reside in the non-catalytic region of a protein and may be reversibly oxidized. The redox state of regulatory Cys is changed by formation of intra- or intermolecular disulfide bonds, glutathionylation, and S-nitrosylation, and thus regulates protein activity. The TTG contains this class of Cys and is the focus of the thesis. The identification of regulatory Cys is mostly through experimental method. Several features, such as acid-base motifs and high frequency of charged amino acids in Cys-flanking regions, have been described but are not specific enough. As pointed out in Fomenko *et al*. [FMG08], computational identification of regulatory Cys is a challenge in redox biology. The ROCD tool developed in the thesis adopts the bottom-up strategy for computational identification of regulatory Cys by first pre-select the oxidation susceptible cysteine. Once the principle for the "re-reducible" property is discovered in the future, this principle can be added into ROCD workflow, so that the set of regulatory Cys can be finally determined.

With the advance of proteomics techniques, chromatography coupled with gel-based or mass spectrometry can experimentally identify TTGs or proteins that undergo dithiol-disulfide transitions [HHF+05][RVS+05][YWL+01][SD08]. Due to technical limitation, further experiments are needed to eliminate the false positive proteins. Specificity and sensitivity of the experimental techniques need to be improved in order to overcome the false positive problem and to identify target proteins with low abundance. ROCD has deployed the capability of bioinformatics to provide additional potential target protein candidates to complete the network and also to support the experimental approach. ROCD complements the high-throughput experimental approaches such as affinity chromatography and diagonal redox SDS gel electrophoresis, which can identify specific proteins that are under redox control but are incapable of identifying specific cysteine thiols that are redox-regulated.

ROCD is intended to serve as a TTG prediction module for the automatic construction of redox regulatory network. ROCD pre-selects the TTG by detecting the oxidation susceptible cysteine which is the initial step in the bottom-up TTG prediction (Fig. 1.1). To achieve the final goal *i.e.*, the prediction of TTG, the other additional steps are the prediction of cysteine re-reducibility and the interaction specificity between TTG and Trx/Grx. Some properties have been discussed and provide the plausible directions for resolving these issues, such as the redox potential and torsional energy of the disulfide bond for redox activity [WFH10], the hydrophobic groove on thioredoxin for recognition and docking of target protein [WCH+01] [MHF+06]. The computational implementation of these strategies will complete the bottom-up pipeline and achieve the automatic construction of redox regulatory network.

7.2 Discussion about network-contexted document retrieval system (ncDocReSy)

Following the acquirement of pre-selected protein by ROCD, ncDocReSy is used to provide relevant literature to assist biologists in manual curation, and some examples have been presented in section 6.2.

ncDocReSy combines the biological networks and literature retrieval services of three literature search engines. The criteria of ncDocReSy are to: (a) maximize the returned documents (b) which mention the intended bioentity with correct semantics in its context. The goal of criteria (a) is to increase the recall and (b) to increase the precision in the document retrieval.

7.2.1 The recall issue

The three search engines incorporated in ncDocReSy differ in the bibliographic fields to be searched in and their indexing and search methodology. The searchable field in PubMed is confined to bibliographic records, such as title, authors, journal name, and the narrowest comparing with the other two search engines but the broadest in term of article coverage. PMC and BioText allow the full text beyond the bibliographic records to be searched, but the size of article collection is confined by the number of open accessed journals. As mentioned in Url12, it is advisable to search PubMed and PMC separately for a comprehensive search. The utilization of these three search engines together will complement the shortcoming of each other and enhance the recall.

Besides the differences in searchable fields, the indexing methods adopted by these three search engine have different effects on recall and precision. Considering the two indexing methodologies used in the present biomedical literature retrieval systems, the full-text based indexing has a higher recall because any document whose content contains the terms matching the user's input will be returned. The keyword-based indexing system has higher precision because only documents whose keywords and thus the main topic match the user's query are returned.

The other strategy for enhancing recall in ncDocReSy is the phrase search function. Phrase search is to add double quote around the query terms or the adhesion of "tw"(standing for "text word") qualifier after the query terms. Phrase search is useful when the queried terms appear literally in the article but were not listed in the entry term of the MeSH heading. However, the phrase search will return some false positive results. The article which contains the string "3-chloro-4-hydroxyphenyl acetate" will be included in the returned list if the user's interest is only "4-hydroxyphenyl acetate", while the phrase search is used. Another example of false positive result would be the article containing "4-hydroxyphenylacetate 1-monooxygenase" or "4-hydroxyphenyl acetate decarboxylase", which is enzyme, is returned when "4-hydroxyphenylacetate", which is a metabolite, is intended. The literature refinement function of ncDocReSy enables to solve this problem.

7.2.2 The precision issue

To fulfill the criteria (b), ncDocReSy uses the advanced search features of PubMed by attaching MeSH qualifier, so that the preceding query term will be interpreted as protein

or metabolite under the “Chemicals and drugs” category of MeSH by PubMed/PMC. This advanced search feature was found to be only seldom used in an analysis of PubMed logs [BHTH07].

Another optional operation for assuring the correct semantic type is to do literature refinement on the preliminary literature list. The preliminary literature list node is marked differently by ncDocReSy when it is retrieved using phrase search mode. This specially marked preliminary search result and the result from BioText can be further filtered to remove some false positive records. ncDocReSy uses the named entity recognition and semantic type annotation functions from whatizit to identify the records containing the synonym of the intended bioentity with the correct semantic type (protein/metabolite). After this filtering process, the article which contains only “3-chloro-4-hydroxyphenyl acetate” will be removed when “4-hydroxyphenyl acetate” is intended, and so will the one with “4-hydroxyphenyl acetate decarboxylase” for “4-hydroxyphenyl acetate”, as exemplified before. As also shown in section 6.2.3, this literature list refinement function can filter out the article which mentions “lanosterol synthase” when only “lanosterol” is desired.

7.2.3 Network-contexted literature ranking

ncDocReSy devises a scoring schema for ranking the literature considering the distance-dependent association and upstream-downstream symmetry around the questioned bioentity. The distance-dependent association is represented by a factor which decreases with the shortest distance between the focused bioentity and its ancestor or descendent nodes. The idea of including this distance-dependent association is to use the associated ancestor/descendent bioentities as the surrogates of the focused bioentity, and farther the associated node from the focused bioentity, the weaker the surrogating ability, and the fewer scores the attached document receives. The consideration of upstream-downstream symmetry is from the theoretical reasoning: if the catalytic enzyme of a reaction malfunctions, the amount of substrates and products of the reaction will be accumulated and decreased, respectively. Therefore, if both the upstream and downstream substances of a focused bioentity in a directional network are mentioned in a document, this document is more worthy of an inspection. For each document, ncDocReSy multiplies the total distance-dependent association scores (RW^{+}_{d} and RW^{-}_{d} in Figure 5.6) from the upstream and downstream regions of the focused bioentity to reveal this upstream-downstream symmetry effect. After this multiplication, the network-scaled score for each

document is generated, and the document list is ranked by this score. The higher ranked document is expected to have closer neighbor nodes located both in the upstream and downstream regions to be co-mentioned together with the focused bioentity in the same document.

7.2.4 Perspective on ncDocReSy

ncDocReSy provides the user with the opportunity of looking for indirectly relevant literatures based on established knowledge of biological network. In the current version, ncDocReSy includes only the network of physical interaction, such as catalytic, protein-protein, and protein-chemical interaction network. The integrated biological network could be extended to include other types of biological networks, such as gene regulation and signal transduction. Besides the incorporation of heterogeneous type of biological network, integration of different resources for the same type of network could also be done in the future. The current version of ncDocReSy uses KEGG database as the only resource for the metabolic network, and the inclusion of Reactome and BioCyc data is planed.

During the calculation of region-specific weight, current ncDocReSy includes the surrogate ability weight considering the shortest distance between the ancestor/descendent node and the focused node. However, the surrogate ability depends not only on the distance but also the bioentity serving as the surrogate. If a bioentity is associated with numerous reactions, the surrogate ability from this bioentity is less representative for one specific reaction. The assignment of varied surrogate weight to different molecules in terms of its exploitation popularity would prioritize the more promising literature.

The implementation of literature list refinement downloads the full text and partitions the text into different sections, such as introduction, conclusion, discussion, before submitting to whatizit for the semantic typing. However, ncDocReSy doesn’t take advantage of the section where the query terms are found. This information could be utilized by ncDocReSy in the future.

Besides, the returned literature list could be further processed by text mining tools to extract new knowledge.

Chapter 8

Conclusion

Bioinformatics both in terms of application and tool development appears to be in the log phase of growth. Hundreds of biological databases and tools are developed aiming at different aspects of cell biology. With the current state of bioinformatics and the advance of information technology, the adaptation of integrative methodology to tackle biological problem is desirable and feasible. The integrative methodology is featured by the utilization of multi-omics data and multi-disciplines resources. Integrating information from different levels of functional hierarchy offers different perspectives for question resolution.

The starting point of this work was the increase in knowledge on the function and importance of the cellular thiol-disulfide redox regulatory network in controlling development and adaptation of organisms. Thus the motivating question was whether the structure of this regulatory network can be expanded by *in silico* approaches. However, biological databases and biomedical publication are populated with the result from popular research topics. When a less investigated or novel research topics are focused, little or no information can be obtained from the database and literature. The question of Trx/Grx target protein identification, which is a critical step in the redox regulatory network construction, faces the same difficulty that little information can be retrieved from the database and the literature. Concerning the data deficiency issue in the target protein identification, a bottom-up strategy is adapted in this thesis to pre-select some candidates. The pre-selected candidate fulfills partial prerequisites of being the promising

candidate, and the number of pre-selected candidate might be numerous. The pre-selection result thus requires further refinement by imposing other rules.

One strategy for pre-selection refinement is manual curation by reading relevant literature. However, the scarcity of relevant literature again becomes a problem. Concerning the literature scarcity problem, the network-contexted document retrieval system stated in chapter 5 extends the literature search by incorporating the biological network topology, so that the indirectly relevant literatures are fetched and ranked by network topology. The network-contexted document retrieval can be seen as bringing Swanson's ABC model into literature search (Fig. 8.1).

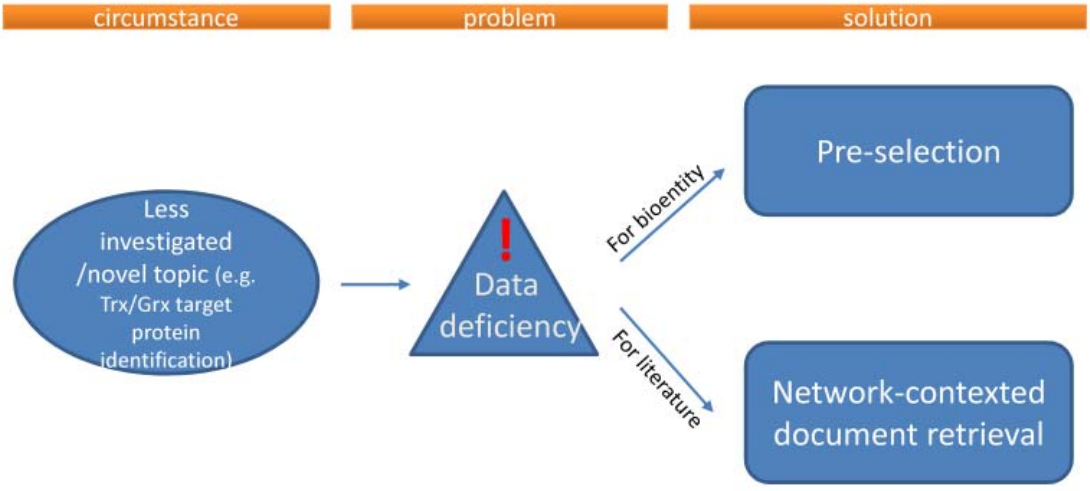


Figure 8.1 The resolution strategy for Trx/Grx target protein identification in the thesis

Chapter 2 exemplifies various bioinformatics resources for cell biology study and essential facilities for the integrative bioinformatics. The capability of an integrative methodology is fostered by the fast generation of biological data, biological databases, bioinformatics tools, and the advance of information technology. In chapter 3, the biological background of the motivating biological question is provided, and the currently relevant research is introduced. This chapter demonstrates that the method of database integration is not suitable for construction of specialized biological network. Besides the structured data from the biological databases, the restriction from mining the unstructured text in the biological literature is also shown. Due to this data deficiency issue, a bottom-up methodology by discriminating the type of amino acid residue, which belongs to the bottom level in the functional hierarchy (Fig. 2.1), is adopted (Fig. 1.1). The discovery from Sanchez *et al*. happens to provide us with this discrimination principle. Chapter 4

focused on the work of Reversibly Oxidized Cysteine Detector (ROCD) which implements the decision rule discovered by Sanchez *et al* [SRWM08] in order to automate the identification of putative elements of the redox regulatory network. The decision rule of Sanchez addresses the oxidation susceptible cysteine residue by considering three physicochemical properties, namely cysteine-cysteine distance, acid dissociation constant ($pK_{a)}$ and accessible surface area (ASA). ROCD combines the resources from biological databases, computational tools for biochemical properties, and the practice of contemporary web-based information techniques, such as the adoption of XML format in the result file and the web service to access dependent resources. ROCD requests the user to provide the protein list, the value ranges of three physicochemical properties before ROCD execution. The protein list can be obtained by either selecting from the lists of human tissue and organelle or user's manual input. Besides being displayed in the web browser, the ROCD result is downloadable in XML format for further processing. In chapter 5, a network-contexted document retrieval system (ncDocReSy) was introduced to provide complementary access to unstructured knowledge deposited in the biomedical literature. It works as the document retrieval system which is essential for literature-based curation. The purpose of ncDocReSy is to assist the scientist looking for indirectly relevant literature concerning their questioned bioentity for manual curation. ncDocReSy is a network-contexted document retrieval system which extends the document retrieval beyond the bioentity of use's interest and takes the other associated bioentities into account from the systematic perspective. The combination of biological network into biomedical document retrieval in ncDocReSy also allows the easy navigation of literatures concerning all bioentities in a biological network. Several essential elements for integrative bioinformatics have been used in ncDocReSy, such as the use of Cytoscape as the integration platform, the use of integrated database DAWIS M.D., web service as the techniques to access IntAct and DAWIS M.D. The specialty of biological resources used in ROCD and ncDocReSy spans the function level from atom/residue to pathway/network and deals with structured data in the biological database as well as the unstructured one present in the scientific publications (Fig 8.2). In the end, the biological application of ROCD on the human mitochondrial proteins has pre-selected 309 potential target proteins for thioredoxin/glutaredoxin-dependent thiol-disulfide transition. Any of these proteins can be inspected through review of relevant literature retrieved from ncDocReSy, and the promising candidates is finally decided after reading the returned literature.

In summary, this thesis shows that the database-driving or literature-driving biological network construction is currently not applicable to the very specific biological network due to the data deficiency in database and literature. However, a bottom-up pre-selection process based on certain known biochemical properties of the involved bioentity can identify the group of possible candidates for further detailed inspection. The incorporation of biological network into literature search helps to retrieve indirectly relevant literature while the directly relevant one is not available for the specialized research topic. The two tools covered in the thesis overcome the data deficiency problem in the specialized study.

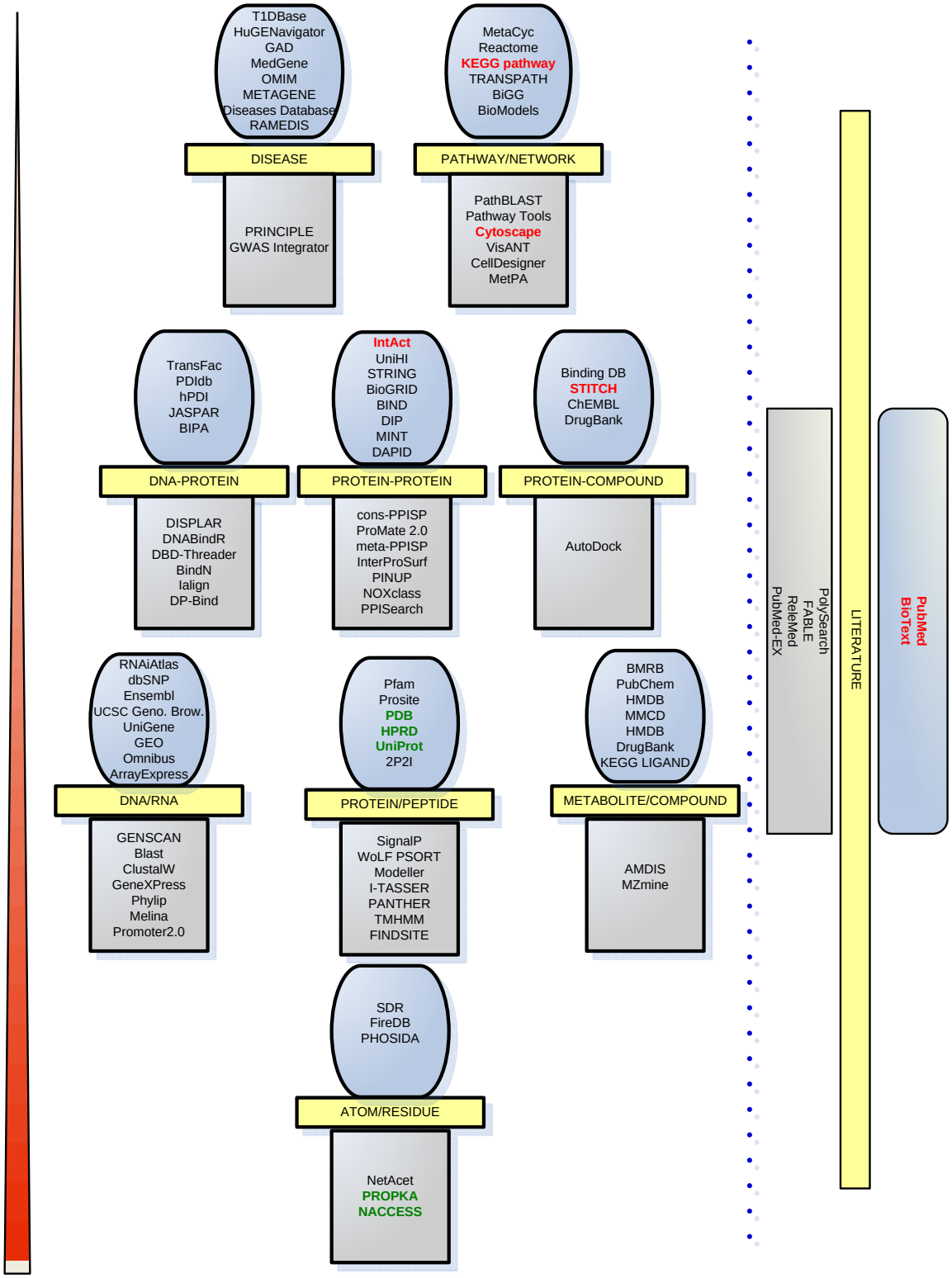


Figure 8.2 The bioinformatics resources utilized in the thesis. The resources used in ROCD are marked in green and ncDocReSy in red.

Acknowledgment

The accomplishment of this thesis is fueled by a lot factors. First I would like to thank the International Graduate School in Bioinformatics and Genome Research as well as the Graduate College Bioinformatic for offering me the opportunity of carrying out my PhD study in Germany. My supervisors– Prof. Dr. Ralf Hofestädt and Prof. Dr. Karl-Josef Dietz–assisted me to orient my research direction, provide insight in bioinformatics and biology, and the careful prove-reading of my thesis. The members from Prof. Hofestädt's and Prof. Dietz's groups– David Braun, Daniela Borck, Sridhar Hariharaputran, Klaus Hippe, Sebastian Janowski, Dr. Benjamin Kormeier, Klaus Kulitza, Andreas Lückner, Alban Shoshi, Björn Sommer, Dr. Meenakumari Muthuramalingam, Dr. Andrea Kandlbinder, Dr. Marie-Luise Oelze –offered me the scientific discussion and technical support. Mrs. Sabine Klusmann deserves the special thank for her help in the tedious administrative affair. Dr. Susanne Schneiker-Bekel, Mrs. Britta Quisbrok, and Mrs. Silke Kölsch from the Graduate School and Graduate College also helped me with the administrative affair.

Being a foreign student sometimes is not easy. Several friends–Chia-wang Lin, Shuo-hsiu Lee, Hui-ling Cheng, Dr. Ping-hua Ho, Chun-fu Lai, Dr. Li-ying Shih, Wei-ming Ho, Fan-ching Leung, Pei-lun Lee, Dr. Chia-lin Lee, Hui-fen Lo, Chia-chi Huang, Dr. Kai Essig, Dr. Huei-ling Yen, Dr. Ralf Joest, Chih-yu Chin, Shu-line Du, Ciao Li, Salvatore Annunziata– sometimes dragged me out of the intensive work and refreshed my mind. The special thanks go to Rong-chu Chen for the frequent guesting over her place and Chung-chih Chen for enriching the leisure time.

Most of all, the family support has no doubt being the main source of comfort and calm during this period. I am grateful for the unconditional love and unlimited support from my parents. I have to thank my brother and my sister-in-law for their attendance of my parents, so that I can more focus on the study.

The financial support during the thesis work was from the International Graduate School of Bioinformatics and Genome Research, Graduate College Bioinformatics, Bielefeld University, and the working group of Prof. Hofestädt.

Appendix A

Target proteins of thioredoxin in plants mitochondria. The GOs in green are from UniProt and annotated as "predicted". The GOs in red are predicted by GoPred [SAC10]

Pathway and protein name	*UniProt accession number*	*Gene Ontoloty: Molecular function*			
Photorespiration					
Glycine cleavage system H protein	P16048	methyltransferase activity			
Glycine cleavage system P protein	P26969	glycine dehydrogenase (decarboxylating) activity	pyridoxal phosphate binding		
Glycine cleavage system T protein	P49364	aminomethyltransferase activity	transaminase activity		
Serine hydroxymethyl transferase	P34899	glycine hydroxymethyl-transferase activity	pyridoxal phosphate binding		
Citric acid cycle-associated reactions					
Aconitase	Q8L784	4 iron, 4 sulfur cluster binding	ATP binding	aconitate hydratase activity	copper ion binding
Dihydrolipoamide acetyltransferase	Q8RWN9	copper ion binding	dihydrolipoyllysine-residue acetyltransferase activity	protein binding	
Dihydrolipoamide dehydrogenase	P31023	dihydrolipoyl dehydrogenase activity	flavin adenine dinucleotide binding		
Isocitrate dehydrogenase	Q7XK22	substrate-specific transporter activity			
Malate dehydrogenase	O48904	L-malate dehydrogenase activity	binding		
Malic enzyme	P37225	NAD or NADH binding	malate dehydrogenase (decarboxylating) activity	metal ion binding	
Pyruvate dehydrogenase E1, alpha su	P52902	pyruvate dehydrogenase (acetyl-transferring) activity			

Pyruvate dehydrogenase E1, beta su	P52904	pyruvate dehydrogenase (acetyl-transferring) activity				
Succinate dehydrogenase (flavoprotein su)	Q9ZPX5	electron carrier activity	flavin adenine dinucleotide binding	Succinate dehydrogenase (ubiquinone) activity		
Succinyl-CoA ligase, alpha su	P68209	ATP binding	ATP citrate synthase activity	copper ion binding	succinate-CoA ligase (ADP-forming) activity	succinate-CoA ligase (GDP-forming) activity
Succinyl-CoA ligase, beta su	Q8LAV0	ATP binding	ligase activity			
Lipid metabolism						
CoA-thioester hydrolase	Q9LKJ1	3-hydroxyisobutyryl-CoA hydrolase activity				
Electron transport						
Cytochrome c oxidase su 5b	Q9LW15	cobalt ion binding	cytochrome-c oxidase activity	zinc ion binding		
Cytochrome c oxidase su 6b	Q9SXV0	cytochrome-c oxidase activity				
NADH-ubiquinone oxidoreductase 75-kDa su	Q43644	2 iron, 2 sulfur cluster binding	4 iron, 4 sulfur cluster binding	NADH dehydrogenase (ubiquinone) activity	electron carrier activity	metal ion binding
Ubiquinol-cytochrome c reductase su II	P29677	Hydrolase	Metalloprotease	Oxidoreductase	Protease	
ATP synthesis/transform-ation						
Adenylate kinase	O82514	ATP binding	adenylate kinase activity	copper ion binding		
ATP synthase, alpha su	P05493	ATP binding	hydrogen ion transporting ATP synthase activity, rotational mechanism	proton-transporting ATPase activity, rotational mechanism		
ATP synthase, beta su	P17614	ATP binding	hydrogen ion transporting ATP synthase activity, rotational mechanism	hydrogen-exporting ATPase activity, phosphorylative mechanism	proton-transporting ATPase activity, rotational mechanism	
ATP synthase, delta su	Missing accession number					

Nucleoside diphosphate kinase	Q9SP13	ATP binding	nucleoside diphosphate kinase activity				
Membrane transport							
Porin (VDAC)	P42056	porin activity	voltage-gated anion channel activity				
Translation							
Elongation factor Tu	Q9ZT91	ATP binding	GTP binding	GTPase activity	cobalt ion binding	translation elongation factor activity	zinc ion binding
Protein assembly/folding							
Chaperonin HSP 60	Q05046	ATP binding	protein binding				
Dna-K molecular chaperone HSP 70	P37900	ATP binding	unfolded protein binding				
Nitrogen metabolism							
Alanine aminotransferase	Q8GRN4	entry deleted					
Aspartate aminotransferase	P46643	L-aspartate:2-oxoglutarate aminotransferase activity	copper ion binding	pyridoxal phosphate binding			
Branched-chain keto acid decarboxylase E1, beta su	O82450	3-methyl-2-oxobutanoate dehydrogenase (2-methylpropanoyl-transferring) activity					
Glutamate dehydrogenase	P93541	binding	glutamate dehydrogenase [NAD(P)+] activity				
Isovaleryl-CoA dehydrogenase	Q9FS87	flavin adenine dinucleotide binding	isovaleryl-CoA dehydrogenase activity				
Leucyl aminopeptidase	O65557	aminopeptidase activity	manganese ion binding	metalloexopeptidase activity			
Methylmalonate-semialdehyde dehydrogenase	Q9SI43	copper ion binding	methylmalonate-semialdehyde dehydrogenase (acylating) activity				

Sulfur metabolism					
Cysteine synthase	Q43153	cysteine synthase activity	pyridoxal phosphate binding	transferase activity	
Mercaptopyruvate sulfurtransferase	O64530	3-mercaptopyruvate sulfurtransferase activity	thiosulfate sulfurtransferase activity		
Hormone synthesis					
Allene oxide cyclase	Q9LEG5	allene-oxide cyclase activity			
Stress-related reactions					
Alcohol dehydrogenase CPRD12	P93697	binding	oxidoreductase activity		
Aldehyde dehydrogenase	P93344	aldehyde dehydrogenase (NAD) activity			
Catalase	P55312	catalase activity	heme binding		
Formate dehydrogenase	Q9S7E4	NAD or NADH binding	formate dehydrogenase activity	oxidoreductase activity, acting on the CH-OH group of donors, NAD or NADP as acceptor	
Glutaredoxin-like protein	Q8LBK6	2 iron, 2 sulfur cluster binding	electron carrier activity	metal ion binding	protein disulfide oxidoreductase activity
Peroxiredoxin	Q9XGP1	entry deleted			
Phospholipid hydroperoxide GSH reductase	O48646	glutathione peroxidase activity	phospholipid-hydroperoxide glutathione peroxidase activity		
Superoxide dismutase Mn	O81233	metal ion binding	superoxide dismutase activity		
Miscellaneous					
Hypothetical protein (29.8 kDa)	Q94GV5	entry deleted			
Hypothetical protein (36.2 kDa)	Q9SUK9	metal ion binding	phosphoprotein phosphatase activity		
Putative protein At5g10860.1	Q9LEV3	catalytic activity	cobalt ion binding		

Appendix B

The calculation result from ROCD for PDB IDs in BALOSCTdb. The PDB IDs in the second column are included in the BALOSCTdb data set. The corresponding SPACC for each PDB ID was obtained by ID mapping service.

SPACC	PDB ID	chain_coverage	resolvedStart_resolvedEnd	mature peptide location	residues matching cysteine-cysteine distance criteria	residues matching pka, asa criteria
P15034	1A16	A_1.00	A1_A440	2_441		A_249(249)_CYS|
P0AEG4	1A2L	A_1.00 B_1.00	A20_A208 B20_B208	20_208	B_30(49)_B_33(52)|A_30(49)_A_33(52)|	
P00883	1ADO	D_1.00 A_1.00 B_1.00 C_1.00	A1_A363 B1_B363 C1_C363 D1_D363	2_364	D_134(134)_D_177(177)|A_134(134)_A_177(177)|C_134(134)_C_177(177)|B_134(134)_B_177(177)|	A_201(201)_CYS|A_239(239)_CYS|A_338(338)_CYS|A_72(72)_CYS|B_201(201)_CYS|B_239(239)_CYS|B_338(338)_CYS|B_72(72)_CYS|C_201(201)_CYS|C_239(239)_CYS|C_338(338)_CYS|D_201(201)_CYS|D_239(239)_CYS|D_338(338)_CYS|
P02768	1AO6	A_0.99 B_0.99	A25_A609 B25_B609	19_609	A_168(192)_A_177(201)|B_245(269)_B_253(277)|A_461(485)_A_477(501)|A_437(461)_A_448(472)|B_265(289)_B_279(303)|B_461(485)_B_477(501)|B_316(340)_B_361(385)|A_75(99)_A_91(115)|A_360(384)_A_369(393)|A_476(500)_A_487(511)|B_90(114)_B_101(125)|A_316(340)_A_361(385)|A_124(148)_A_169(193)|A_90(114)_A_101(125)|A_53(77)_A_62(86)|A_392(416)_A_438(462)|B_514(538)_B_559(583)|A_514(538)_A_559(583)|B_53(77)_B_62(86)|B_168(192)_B_177(201)|A_265(289)_A_279(303)|B_200(224)_B_246(270)|B_360(384)_B_369(393)|A_245(269)_A_253(277)|B_392(416)_B_438(462)|A_200(224)_A_246(270)|B_437(461)_B_448(472)|B_278(302)_B_289(313)|A_278(302)_A_289(313)|B_75(99)_B_91(115)|B_558(582)_B_567(591)|A_558(582)_A_567(591)|B_476(500)_B_487(511)|B_124(148)_B_169(193)|B_278(302)_B_279(303)|A_278(302)_A_279(303)|	
P69924	1AV8	A_0.90 B_0.90	A1_A340 B1_B340	2_376	B_268(268)_B_272(272)|A_268(268)_A_272(272)|	A_305(305)_CYS|B_305(305)_CYS|
P00784	1BQI	A_0.65	A134_A345	19_345	A_153(286)_A_200(333)|A_22(155)_A_63(196)|A_56(189)_A_95(228)|	A_25(158)_CYS|
P18031	1BZH	A_0.68	A1_A298	1_435		A_215(215)_CYS|
P30304	1C25	A_0.30	A336_A495	1_524	A_384(384)_A_430(430)|	
P30074	1CGK	A_1.00	A1_A389	1_389	A_130(130)_A_190(190)|	A_164(164)_CYS|A_30(30)_CYS|
P0A9P4	1CL0	A_1.00	A1_A320	2_321	A_135(135)_A_138(138)|	A_303(303)_CYS|
P13650	1CRU	A_1.00 B_1.00	A25_A478 B25_B478	25_478	A_338(362)_A_345(369)|B_338(362)_B_345(369)|	
P05091	1CW3	D_0.99 E_0.99 F_0.99 G_0.99 A_0.99 B_0.99 C_0.99 H_0.99	A24_A517 B24_B517 C24_C517 D24_D517 E24_E517 F24_F517 G24_G517 H24_H517	18_517	B_301(318)_B_303(320)|H_301(318)_H_303(320)|A_301(318)_A_303(320)|C_301(318)_C_303(320)|F_301(318)_F_303(320)|D_301(318)_D_303(320)|E_301(318)_E_303(320)|G_301(318)_G_303(320)|G_302(319)_G_303(320)|B_302(319)_B_303(320)|D_302(319)_D_303(320)|F_302(319)_F_303(320)|A_302(319)_A_303(320)|C_302(319)_C_303(320)|E_302(319)_E_303(320)|B_301(318)_B_302(319)|A_301(318)_A_302(319)|E_301(318)_E_302(319)|H_302(319)_H_303(320)|C_301(318)_C_302(319)|F_301(318)_F_302(319)|G_301(318)_G_302(319)|D_301(318)_D_302(319)|H_301(318)_H_302(319)|	pka or asa file is not successfully generated
P60484	1D5R	A_0.11	A7_A285 A310_A353	1_403	A_71(71)_A_124(124)|	A_218(218)_CYS|A_250(250)_CYS|
P19080	1DBF	A_1.00 B_1.00 C_1.00	A1_A127 B1_B127 C1_C127	1_127		
P00276	1DE2	A_1.00	A1_A87	1_87	A_14(14)_A_17(17)|	
P04531	1DEL	A_1.00 B_1.00	A1_A241 B1_B241	1_241		A_55(55)_CYS|B_55(55)_CYS|

P07097	1DLU	D_0.99 A_0.99 B_0.99 C_0.99	A4_A391 B4_B391 C4_C391 D4_D391	2_392	D_89(89)_D_378(378)|B_89(89)_B_378(378)|A_89(89)_A_378(378)|C_89(89)_C_378(378)|	
P68688	1EGR	A_1.00	A1_A85	1_85		A_11(11)_CYS|
P10599	1ERT	A_0.99	A1_A104	2_105	A_32(31)_A_35(34)|	A_62(61)_CYS|A_73(72)_CYS|
P39593	1ESQ	A_1.00 B_1.00 C_1.00	A1_A272 B1_B272 C1_C272	1_272		
P25500	1F5A	A_0.69	A1_A513	2_739		A_160(160)_CYS|A_197(197)_CYS|A_293(293)_CYS|A_36(36)_CYS|
Q16667	1FPZ	D_1.00 E_1.00 F_1.00 A_1.00 B_1.00 C_1.00	A1_A212 B1_B212 C1_C212 D1_D212 E1_E212 F1_F212	1_212	E_119(119)_E_153(153)|A_119(119)_A_153(153)|B_119(119)_B_153(153)|D_119(119)_D_153(153)|C_119(119)_C_153(153)|F_119(119)_F_153(153)|A_79(79)_A_140(140)|F_79(79)_F_140(140)|D_79(79)_D_140(140)|B_79(79)_B_140(140)|C_79(79)_C_140(140)|E_79(79)_E_140(140)|	A_120(120)_CYS|A_129(129)_CYS|A_39(39)_CYS|B_120(120)_CYS|B_129(129)_CYS|B_39(39)_CYS|B_51(51)_CYS|B_70(70)_CYS|B_99(99)_CYS|C_129(129)_CYS|C_39(39)_CYS|D_39(39)_CYS|D_51(51)_CYS|E_120(120)_CYS|E_39(39)_CYS|E_51(51)_CYS|E_70(70)_CYS|E_99(99)_CYS|F_39(39)_CYS|F_70(70)_CYS|
P61586	1FTN	A_1.00	A1_A193	1_193	A_16(16)_A_83(83)|	
P0A5N4	1GU9	D_-1.00 E_-1.00 F_-1.00 G_-1.00 A_-1.00 B_-1.00 C_-1.00 L_-1.00 H_-1.00 I_-1.00 J_-1.00 K_-1.00	inconsis. annot. btw UniProt and PDB	1_177	K_130(130)_K_133(133)|I_130(130)_I_133(133)|G_130(130)_G_133(133)|L_130(130)_L_133(133)|H_130(130)_H_133(133)|D_130(130)_D_133(133)|A_130(130)_A_133(133)|J_130(130)_J_133(133)|E_130(130)_E_133(133)|B_130(130)_B_133(133)|C_130(130)_C_133(133)|F_130(130)_F_133(133)|	
P28593	1GXF	A_1.00 B_1.00	A1_A492 B1_B492	1_492	B_444(444)_A_444(444)|A_53(53)_A_58(58)|B_53(53)_B_58(58)|	
Q9Z2F5	1HL3	A_0.81	A1_A350	1_430	A_107(107)_A_339(339)|	A_123(123)_CYS|
P0A6R0	1HN9	A_1.00 B_1.00	A1_A317 B1_B317	1_317		A_112(112)_CYS|
P0A6Y5	1HW7	A_0.87	A1_A255	1_292	A_232(232)_A_234(234)|	
P0ACQ4	1I69	A_0.72 B_0.72	A87_A305 B87_B305	1_305		A_208(208)_CYS|
O55236	1I9S	A_0.35	A1_A210	1_597		A_126(126)_CYS|A_193(193)_CYS|A_97(97)_CYS|
P42212	1JC0	A_1.00 B_1.00 C_1.00	A1_A238 B1_B238 C1_C238	1_238	C_147(147)_C_204(204)|A_147(147)_A_204(204)|B_147(147)_B_204(204)|	
P50135	1JQD	A_1.00 B_1.00	A1_A292 B1_B292	1_292		A_82(82)_CYS|B_248(248)_CYS|
Q01745	1K3I	A_-1.00	inconsis. annot. btw UniProt and PDB	25_680	A_515(515)_A_518(518)|A_18(18)_A_27(27)|	
P0A006	1LJL	A_1.00	A1_A131	1_131	A_10(10)_A_82(82)|	
O26232	1LOR	A_1.00	A1_A228	1_228	A_84(84)_A_109(109)|	
P00439	1MMT	A_0.72	A103_A427	1_452	A_203(203)_A_334(334)|	A_237(237)_CYS|
P05109	1MR8	A_1.00 B_1.00	A1_A93 B1_B93	1_93		
Q9HZZ3	1N2F	A_-1.00 B_-1.00	A1_A142 B1_B142	-1_-1	A_60(60)_A_124(124)|B_60(60)_B_124(124)|	
P37062	1NHQ	A_1.00	A1_A447	1_447		
P00125	1NTM	D_0.65	D1_D241	85_325		D_40(40)_CYS|
P00126	1NTM	H_0.83	H1_H78	14_91	H_24(24)_H_68(68)|H_40(40)_H_54(54)|H_24(24)_H_30(30)|H_30(30)_H_68(68)|	
P00129	1NTM	F_0.99	F1_F110	2_111		
P00130	1NTM	J_0.97	J1_J62	2_64		
P00157	1NTM	C_1.00	C1_C379	1_379	C_40(40)_C_93(93)|	C_323(323)_CYS|
P07552	1NTM	K_1.00	K1_K56	1_56		
P13271	1NTM	G_0.99	G1_G81	2_82		G_44(44)_CYS|
P13272	1NTM	E_0.71	E79_E274	1_274	E_144(222)_E_160(238)|E_139(217)_E_158(236)|E_158(236)_E_160(238)|E_144(222)_E_158(236)|	
P23004	1NTM	B_1.00	B15_B453	15_453		
P31800	1NTM	A_1.00	A35_A480	35_480	A_346(380)_A_411(445)|	A_419(453)_CYS|
P12931	1O4B	A_0.20	A144_A251	2_536		A_44(187)_CYS|

P25799	1OOA	A_0.75 B_0.75	A39_A363 B39_B363	1_431	B_116(116)_B_121(121)|A_116(116)_A_121(121)|	A_259(259)_CYS|A_85(85)_CYS|B_259(259)_CYS|B_85(85)_CYS|
P00586	1ORB	A_1.00	A1_A296	2_297		
P0ABU5	1OY1	D_1.00 A_1.00 B_1.00 C_1.00	A1_A220 B1_B220 C1_C220 D1_D220	1_217	B_14(14)_A_14(14)|D_14(14)_C_14(14)|	A_114(114)_CYS|A_138(138)_CYS|B_114(114)_CYS|B_138(138)_CYS|C_138(138)_CYS|D_114(114)_CYS|D_138(138)_CYS|
P18052	1P15	A_0.31 B_0.31	A577_A829 B577_B829	20_829		A_660(695)_CYS|B_619(654)_CYS|B_660(695)_CYS|B_724(759)_CYS|
P0A790	1PT1	A_1.00 B_1.00	A1_A126 B1_B126	1_126		A_78(78)_CYS|B_78(78)_CYS|
P50097	1PVN	D_0.55 A_0.55 B_0.55 C_0.55	A2_A100 A227_A503 B2_B100 B227_B503 C2_C100 C227_C503 D2_D100 D227_D503	1_503	C_26(26)_C_459(459)|B_26(26)_B_459(459)|A_26(26)_A_459(459)|D_26(26)_D_459(459)|	A_319(319)_CYS|B_319(319)_CYS|C_319(319)_CYS|D_319(319)_CYS|
P23687	1QFS	A_1.00	A1_A710	1_710	A_573(573)_A_703(703)|	A_25(25)_CYS|A_532(532)_CYS|A_57(57)_CYS|A_601(601)_CYS|
P01009	1QLP	A_1.00	A26_A418	25_418		A_232(256)_CYS|
P00452	1R1R	A_1.00 B_1.00 C_1.00	A1_A761 B1_B761 C1_C761	1_761	C_225(225)_C_462(462)|A_225(225)_A_462(462)|B_225(225)_B_462(462)|	C_292(292)_CYS|
P69924	1R1R	D_0.05 E_0.05 F_0.05 P_0.05	D356_D375 E356_E375 F356_F375 P356_P375	2_376		
P68871	1RQA	D_0.99 B_0.99	B1_B146 D1_D146	2_147		B_112(112)_CYS|D_112(112)_CYS|D_93(93)_CYS|
P69905	1RQA	A_0.99 C_0.99	A1_A141 C1_C141	2_142		
P04585	1RTH	A_0.39	A587_A1146	2_1435		
Q04432	1RW7	A_1.00	A1_A237	1_237		A_138(138)_CYS|
P77150	1TD2	A_1.00 B_1.00	A1_A287 B1_B287	1_287		A_53(53)_CYS|B_53(53)_CYS|
P04406	1U8F	Q_1.00 P_1.00 R_1.00 O_1.00	O0_O334 P0_P334 Q0_Q334 R0_R334	2_335		P_152(151)_CYS|Q_152(151)_CYS|R_152(151)_CYS|
O58720	1UMJ	A_1.00 B_1.00	A1_A102 B1_B102	1_102		A_29(29)_CYS|B_29(29)_CYS|
P17559	1UTR	A_1.00 B_1.00	A1_A96 B1_B96	20_96		A_5(24)_CYS|A_71(90)_CYS|B_5(24)_CYS|B_71(90)_CYS|
Q9RRU8	1VH2	A_1.00	A1_A158	1_158		A_125(125)_CYS|A_82(82)_CYS|
Q9FD71	1X9E	A_-1.00 B_-1.00	A1_A383 B1_B383	-1_-1		
O33839	1XJN	D_-1.00 A_-1.00 B_-1.00 C_-1.00	A1_A644 B1_B644 C1_C644 D1_D644	-1_-1	B_134(134)_B_333(333)|A_134(134)_A_333(333)|D_134(134)_D_333(333)|C_134(134)_C_333(333)|	
Q9BYN0	1XW3	A_0.77	A32_A137	1_137		A_99(99)_CYS|
P24666	1XWW	A_1.00	A2_A158	2_158	A_145(146)_A_149(150)|	A_109(110)_CYS|A_12(13)_CYS|
P65688	1XXU	D_1.00 A_1.00 B_1.00 C_1.00	A1_A153 B1_B153 C1_C153 D1_D153	1_153		D_45(45)_CYS|
Q8NBK3	1Y1I	X_-1.00	inconsis. annot. btw UniProt and PDB	34_374	X_218(218)_X_365(365)|X_235(235)_X_346(346)|X_336(336)_X_341(341)|	
P66952	1Y25	A_0.99 B_0.99	A3_A165 B3_B165	1_165		A_93(93)_CYS|B_93(93)_CYS|
P37330	1Y8B	A_1.00	A1_A722	2_723		A_496(495)_CYS|A_617(616)_CYS|A_688(687)_CYS|
P25052	1YAF	D_1.00 A_1.00 B_1.00 C_1.00	A1_A236 B1_B236 C1_C236 D1_D236	1_236		A_149(149)_CYS|B_149(149)_CYS|C_149(149)_CYS|D_149(149)_CYS|
P0A9J4	1YJQ	A_1.00	A1_A303	1_303	A_8(8)_A_37(37)|	
P07451	1Z97	A_1.00	A0_A259	2_260		
Q9JLT4	1ZDL	A_1.00	A31_A524	35_524	A_86(86)_A_91(91)|	A_279(279)_CYS|A_483(483)_CYS|
P0A114	1ZI6	A_1.00	A1_A236	1_236		
P01112	1ZVQ	A_0.88	A1_A166	1_189		

P50163	2AE2	A_1.00 B_1.00	A2_A260 B2_B260	1_260	A_39(39)_A_65(65)|B_39(39)_B_65(65)|	A_10(10)_CYS|A_217(217)_CYS|A_258(258)_CYS|B_10(10)_CYS|B_217(217)_CYS|B_258(258)_CYS|
P38503	2AF3	D_1.00 C_1.00	C0_C332 D0_D332	2_333	C_277(276)_D_277(276)|	C_159(158)_CYS|D_159(158)_CYS|
P68826	2AI9	A_1.00 B_1.00	A1_A183 B1_B183	1_183		
P17618	2B3Z	D_1.00 A_1.00 B_1.00 C_1.00	A1_A361 B1_B361 C1_C361 D1_D361	1_361	B_74(74)_B_83(83)|D_74(74)_D_83(83)|C_74(74)_C_83(83)|A_74(74)_A_83(83)|	B_233(233)_CYS|C_233(233)_CYS|D_233(233)_CYS|
O89053	2B4E	A_0.87	A1_A402	2_461	A_40(40)_A_345(345)|	A_152(152)_CYS|A_332(332)_CYS|A_78(78)_CYS|A_90(90)_CYS|
P17967	2B5E	A_1.00	A23_A522	29_522	A_90(90)_A_97(97)|A_406(406)_A_409(409)|A_61(61)_A_64(64)|	
P21816	2B5H	A_1.00	A1_A200	1_200		
P02584	2BTF	P_0.99	P1_P139	2_140		
P60712	2BTF	A_1.00	A2_A375	1_375	A_217(217)_A_257(257)|	A_272(272)_CYS|
P21397	2BXS	A_1.00 B_1.00	A1_A527 B1_B527	1_527	B_321(321)_B_323(323)|A_321(321)_A_323(323)|	
Q9P2T1	2BZN	D_0.95 E_0.95 F_0.95 G_0.95 A_0.95 B_0.95 C_0.95 H_0.95	A10_A341 B10_B341 C10_C341 D10_D341 E10_E341 F10_F341 G10_G341 H10_H341	1_348	G_68(68)_G_95(95)|B_68(68)_B_95(95)|C_68(68)_C_95(95)|D_68(68)_D_95(95)|E_68(68)_E_95(95)|H_68(68)_H_95(95)|A_68(68)_A_95(95)|F_68(68)_F_95(95)|	A_224(224)_CYS|C_224(224)_CYS|D_224(224)_CYS|E_224(224)_CYS|F_224(224)_CYS|G_224(224)_CYS|
Q56839	2C3C	A_1.00 B_1.00	A1_A523 B1_B523	1_523	A_82(82)_A_87(87)|B_82(82)_B_87(87)|	A_156(156)_CYS|B_156(156)_CYS|
P54687	2COJ	A_1.00 B_1.00	A1_A386 B1_B386	1_386	B_335(335)_B_338(338)|A_335(335)_A_338(338)|	A_235(235)_CYS|B_221(221)_CYS|B_235(235)_CYS|
P00563	2CRK	A_1.00	A1_A381	1_381		A_146(146)_CYS|A_283(283)_CYS|
P21524	2CVX	A_1.00	A1_A888	1_888	A_218(218)_A_443(443)|	A_362(362)_CYS|
P46881	2CWT	A_1.00 B_1.00	A1_A638 B1_B638	1_638	B_317(317)_B_343(343)|A_317(317)_A_343(343)|	
Q9YA14	2CX4	D_-1.00 E_-1.00 F_-1.00 G_-1.00 A_-1.00 B_-1.00 C_-1.00 H_-1.00	inconsis. annot. btw UniProt and PDB	-1_-1	G_49(49)_G_54(54)|C_49(49)_C_54(54)|B_80(80)_A_80(80)|F_49(49)_F_54(54)|C_80(80)_D_80(80)|F_80(80)_E_80(80)|H_80(80)_G_80(80)|B_49(49)_B_54(54)|	A_54(54)_CYS|D_54(54)_CYS|E_54(54)_CYS|H_54(54)_CYS|
Q49610	2FHK	D_1.00 A_1.00 B_1.00 C_1.00	A1_A296 B1_B296 C1_C296 D1_D296	1_296	A_50(50)_B_96(96)|C_50(50)_D_96(96)|D_50(50)_C_96(96)|B_50(50)_A_96(96)|	
P16088	2FIV	A_0.10 B_0.10	A39_A154 B39_B154	1_1124		A_84(122)_CYS|B_84(122)_CYS|
P36655	2FWF	A_0.23	A438_A565	20_565	A_461(480)_A_464(483)|	
P0A744	2GT3	A_1.00	A1_A211	2_212		A_198(198)_CYS|A_206(206)_CYS|A_86(86)_CYS|
Q8JLF5	2HZE	A_1.00 B_1.00	A1_A108 B1_B108	1_108	A_23(23)_A_26(26)|	B_26(26)_CYS|
P29557	2IDV	A_0.82	A39_A215	1_215	A_113(113)_A_151(151)|	
P32322	2IZZ	D_0.94 E_0.94 A_0.94 B_0.94 C_0.94	A1_A300 B1_B300 C1_C300 D1_D300 E1_E300	2_319	E_95(95)_E_120(120)|C_95(95)_C_120(120)|B_95(95)_B_120(120)|A_95(95)_A_120(120)|D_95(95)_D_120(120)|	E_262(262)_CYS|
P04191	2OA0	A_0.99	A1_A993	1_1001	A_636(636)_A_675(675)|	A_674(674)_CYS|A_70(70)_CYS|A_876(876)_CYS|A_938(938)_CYS|
P09211	3PGT	A_1.00 B_1.00	A1_A210 B1_B210	2_210	A_101(102)_B_101(102)|	

Appendix C

Statistics for the validation of ROCD calculation against BALOSCTdb. The data in the first four columns is extracted from BALOSCTdb. The last column shows the compliance of ROCD prediction with BALOSCTdb.

PDB ID	CHAIN	SEQUENCE POSITION	CLASS	*comply with ROCD prediction (1: yes, 0: no)*	PDB ID	CHAIN	SEQUENCE POSITION	CLASS	*comply with ROCD prediction (1: yes, 0: no)*
1A16	A	202	0	***1***	1A16	A	249	1	***1***
1A16	A	240	0	***1***	1A2L	A	30	1	***1***
1ADO	A	134	0	***0***	1A2L	A	33	1	***1***
1ADO	A	149	0	***1***	1ADO	A	72	1	***1***
1ADO	A	177	0	***0***	1ADO	A	201	1	***1***
1ADO	A	289	0	***1***	1ADO	A	338	1	***1***
1AV8	A	196	0	***1***	1AO6	A	34	1	***0***
1AV8	A	214	0	***1***	1AV8	A	268	1	***1***
1BZH	A	92	0	***1***	1AV8	A	272	1	***1***
1BZH	A	121	0	***1***	1BQI	A	25	1	***1***
1BZH	A	231	0	***1***	1BZH	A	215	1	***1***
1C25	A	441	0	***1***	1C25	A	384	1	***1***
1C25	A	476	0	***1***	1C25	A	430	1	***1***
1C25	A	480	0	***1***	1CGK	A	164	1	***1***
1CGK	A	130	0	***0***	1CL0	A	135	1	***1***
1CGK	A	190	0	***0***	1CL0	A	138	1	***1***
1CL0	A	105	0	***1***	1CRU	A	338	1	***1***
1CW3	A	19	0	***1***	1CRU	A	345	1	***1***
1CW3	A	162	0	***1***	1CW3	A	301	1	***1***
1CW3	A	455	0	***1***	1CW3	A	303	1	***1***
1D5R	A	83	0	***1***	1D5R	A	71	1	***1***
1D5R	A	136	0	***1***	1D5R	A	124	1	***1***
1DBF	A	88	0	***1***	1DBF	A	75	1	***0***
1ERT	A	69	0	***1***	1DE2	A	14	1	***1***
1ERT	A	73	0	***0***	1DE2	A	17	1	***1***
1ESQ	A	9	0	***1***	1DEL	A	55	1	***1***
1ESQ	A	208	0	***1***	1DLU	A	89	1	***1***
1F5A	A	118	0	***1***	1EGR	A	11	1	***1***
1F5A	A	160	0	***0***	1EGR	A	14	1	***0***
1F5A	A	293	0	***0***	1ERT	A	32	1	***1***
1FPZ	A	70	0	***1***	1ERT	A	35	1	***1***
1FPZ	A	119	0	***0***	1ESQ	A	198	1	***0***
1FPZ	A	148	0	***1***	1F5A	A	36	1	***1***
1FPZ	A	153	0	***0***	1F5A	A	197	1	***1***
1FTN	A	83	0	***0***	1FPZ	A	79	1	***1***
1FTN	A	107	0	***1***	1FPZ	A	140	1	***1***
1FTN	A	159	0	***1***	1FTN	A	16	1	***1***
1GXF	A	176	0	***1***	1FTN	A	20	1	***0***
1GXF	A	375	0	***1***	1GU9	A	130	1	***1***
1GXF	A	469	0	***1***	1GU9	A	133	1	***1***
1HL3	A	226	0	***1***	1GXF	A	53	1	***1***
1HN9	A	103	0	***1***	1GXF	A	58	1	***1***
1HN9	A	146	0	***1***	1HL3	A	27	1	***0***
1HN9	A	282	0	***1***	1HN9	A	112	1	***1***
1HW7	A	141	0	***1***	1HW7	A	232	1	***1***
1I9S	A	97	0	***0***	1HW7	A	234	1	***1***

1I9S	A	110	0	***1***	1I69	A	199	1	***0***
1I9S	A	138	0	***1***	1I69	A	208	1	***1***
1I9S	A	193	0	***0***	1I9S	A	126	1	***1***
1JC0	A	70	0	***1***	1JC0	A	147	1	***1***
1JQD	A	31	0	***1***	1JC0	A	204	1	***1***
1JQD	A	229	0	***1***	1JQD	A	82	1	***1***
1LJL	A	15	0	***1***	1JQD	A	217	1	***0***
1LOR	A	109	0	***0***	1JQD	A	248	1	***0***
1MMT	A	357	0	***1***	1K3I	A	18	1	***1***
1NTM	H	30	0	***0***	1K3I	A	27	1	***1***
1OKG	A	33	0	***spacc missing***	1LJL	A	10	1	***1***
1OKG	A	134	0	***spacc missing***	1LJL	A	82	1	***1***
1OKG	A	260	0	***spacc missing***	1LOR	A	65	1	***0***
1OKG	A	278	0	***spacc missing***	1MMT	A	217	1	***0***
1OOA	A	59	0	***1***	1MMT	A	334	1	***1***
1OOA	A	259	0	***0***	1MR8	A	42	1	***0***
1ORB	A	63	0	***1***	1N2F	A	60	1	***1***
1OY1	A	176	0	***1***	1N2F	A	124	1	***1***
1P15	A	660	0	***0***	1NHQ	A	42	1	***0***
1P15	A	735	0	***1***	1NTM	H	40	1	***1***
1PS4	A	46	0	***spacc missing***	1NTM	H	54	1	***1***
1PS4	A	53	0	***spacc missing***	1O4B	A	44	1	***1***
1PT1	A	26	0	***1***	1OKG	A	247	1	***spacc missing***
1PVN	A	356	0	***1***	1OOA	A	116	1	***1***
1PVN	A	461	0	***1***	1OOA	A	121	1	***1***
1QFS	A	78	0	***1***	1ORB	A	247	1	***0***
1QFS	A	225	0	***1***	1OY1	A	138	1	***1***
1QFS	A	526	0	***1***	1P15	A	724	1	***0***
1QFS	A	532	0	***0***	1PS4	A	106	1	***spacc missing***
1QFS	A	703	0	***0***	1PT1	A	78	1	***1***
1R1R	A	179	0	***1***	1PVN	A	26	1	***1***
1R1R	A	420	0	***1***	1PVN	A	459	1	***1***
1R1R	A	439	0	***1***	1QFS	A	25	1	***1***
1R1R	A	565	0	***1***	1QFS	A	57	1	***1***
1RTH	A	38	0	***1***	1QFS	A	175	1	***0***
1TD2	A	76	0	***1***	1QFS	A	255	1	***0***
1TD2	A	111	0	***1***	1QFS	A	601	1	***1***
1TD2	A	122	0	***1***	1QLP	A	232	1	***1***
1U8F	O	156	0	***1***	1R1R	A	225	1	***1***
1U8F	O	247	0	***1***	1R1R	A	462	1	***1***
1U8F	O	284	0	***1***	1RQA	B	93	1	***0***
1VH2	A	125	0	***0***	1RTH	A	280	1	***0***
1XJN	A	322	0	***1***	1RW7	A	138	1	***1***
1XWW	A	62	0	***1***	1TD2	A	53	1	***1***
1XWW	A	109	0	***0***	1U8F	O	152	1	***0***
1XWW	A	149	0	***0***	1UMJ	A	29	1	***1***
1Y25	A	93	0	***0***	1UTR	A	5	1	***1***
1Y8B	A	271	0	***1***	1UTR	A	71	1	***1***
1Y8B	A	496	0	***0***	1VH2	A	82	1	***1***
1YAF	A	7	0	***1***	1X9E	A	111	1	***0***
1YJQ	A	181	0	***1***	1XJN	A	134	1	***1***
1Z97	A	66	0	***1***	1XJN	A	333	1	***1***
1ZDL	A	54	0	***1***	1XW3	A	99	1	***1***
1ZDL	A	233	0	***1***	1XWW	A	12	1	***1***

1ZDL	A	407	0	***1***	1XWW	A	17	1	***0***
1ZDL	A	446	0	***1***	1XXU	A	45	1	***0***
1ZDL	A	500	0	***1***	1Y1I	X	336	1	***1***
1ZI6	A	60	0	***1***	1Y1I	X	341	1	***1***
1ZVQ	A	51	0	***1***	1Y25	A	80	1	***0***
2AE2	A	10	0	***0***	1Y8B	A	617	1	***1***
2AE2	A	217	0	***0***	1Y8B	A	688	1	***1***
2AE2	A	236	0	***1***	1YAF	A	149	1	***1***
2AE2	A	258	0	***0***	1YJQ	A	8	1	***1***
2AF3	C	312	0	***1***	1YJQ	A	37	1	***1***
2AF3	C	325	0	***1***	1Z97	A	183	1	***0***
2B3Z	A	233	0	***1***	1Z97	A	188	1	***0***
2B3Z	A	304	0	***1***	1ZDL	A	86	1	***1***
2B4E	A	24	0	***1***	1ZDL	A	91	1	***1***
2B4E	A	51	0	***1***	1ZI6	A	123	1	***0***
2B4E	A	104	0	***1***	1ZVQ	A	118	1	***0***
2B4E	A	192	0	***1***	2AE2	A	39	1	***1***
2B4E	A	195	0	***1***	2AE2	A	65	1	***1***
2B4E	A	285	0	***1***	2AF3	C	159	1	***1***
2B5H	A	93	0	***1***	2AI9	A	110	1	***0***
2B5H	A	130	0	***1***	2B3Z	A	74	1	***1***
2BTF	A	217	0	***0***	2B3Z	A	83	1	***1***
2BTF	A	257	0	***0***	2B4E	A	78	1	***1***
2BTF	A	285	0	***1***	2B4E	A	90	1	***1***
2BXS	A	201	0	***1***	2B4E	A	152	1	***1***
2BXS	A	266	0	***1***	2B4E	A	332	1	***1***
2BXS	A	374	0	***1***	2B5E	A	406	1	***1***
2BXS	A	398	0	***1***	2B5E	A	409	1	***1***
2BXS	A	406	0	***1***	2B5H	A	164	1	***0***
2BZN	A	127	0	***1***	2BTF	A	272	1	***1***
2BZN	A	186	0	***1***	2BTF	A	374	1	***0***
2BZN	A	205	0	***1***	2BXS	A	321	1	***1***
2BZN	A	222	0	***1***	2BXS	A	323	1	***1***
2BZN	A	224	0	***0***	2BZN	A	68	1	***1***
2BZN	A	316	0	***1***	2BZN	A	95	1	***1***
2C3C	A	96	0	***1***	2C3C	A	82	1	***1***
2C3C	A	156	0	***0***	2C3C	A	87	1	***1***
2C3C	A	231	0	***1***	2COJ	A	335	1	***1***
2COJ	A	221	0	***1***	2COJ	A	338	1	***1***
2CRK	A	146	0	***0***	2CRK	A	283	1	***1***
2CVX	A	429	0	***1***	2CVX	A	218	1	***1***
2CVX	A	620	0	***1***	2CVX	A	443	1	***1***
2CWT	A	315	0	***1***	2CWT	A	317	1	***1***
2CX4	A	80	0	***0***	2CWT	A	343	1	***1***
2FHK	A	228	0	***1***	2CX4	A	49	1	***0***
2FHK	A	242	0	***1***	2CX4	A	54	1	***1***
2FIV	A	90	0	***1***	2FHK	A	58	1	***0***
2GT3	A	86	0	***0***	2FIV	A	84	1	***1***
2IDV	A	123	0	***1***	2FWF	A	461	1	***1***
2IZZ	A	159	0	***1***	2FWF	A	464	1	***1***
2IZZ	A	262	0	***1***	2GT3	A	51	1	***0***
2OA0	A	12	0	***1***	2GT3	A	198	1	***1***
2OA0	A	70	0	***0***	2GT3	A	206	1	***1***
2OA0	A	268	0	***1***	2HZE	A	23	1	***1***

2OA0	A	417	0	***1***	2HZE	A	26	1	***1***
2OA0	A	498	0	***1***	2IDV	A	113	1	***1***
2OA0	A	774	0	***1***	2IDV	A	151	1	***1***
2OA0	A	876	0	***0***	2IZZ	A	95	1	***1***
2OA0	A	910	0	***1***	2IZZ	A	120	1	***1***
3PGT	A	14	0	***1***	2OA0	A	674	1	***1***
3PGT	A	47	0	***1***	3PGT	A	101	1	***1***
SUM				***120***	**SUM**				***125***

Appendix D

Result from applying ROCD on 46 thioredoxin target proteins in plant mitochondria. The PDB ID in upper case letters are the exact structure entry for the protein, and the PDB ID in lower case is obtained by homology modeling. The fifth column from the left records the sequence identity and the position of the fragment being modeled, and the sixth column indicates the position of mature peptide. The two digits separated by underscore in the seventh column mean the cysteine residue pair, and the eighth column records the qualified residue number and residue name. The last column indicates the position of CxxC motif. If "-1" is shown in the data value, it indicates the missing information. The data syntax for each column is:

peptideCoverage column-- "chainSymbol"_"peptideCoverage"

modelSimilarity column-- "chainSymbol""peptideResidueNumber"_"chainSymbol""peptideResidueNumber" or "spacc"_"pdbModelTemplate"_"sequenceSimilarity"_"modeledResidueStart"_"modeledResidueEnd"_ SwissModel

maturePeptide column-- "peptideResidueNumberStart"_"peptideResidueNumberEnd"

NoteG1 column-- [list of chains]_"pdbResidueNumber"("uniprotResidueNumber")_[list of chains]_"pdbResidueNumber"("uniprotResidueNumber") or "peptideResidue"_"peptideResidue"

NoteG2 column-- [list of chains]_"pdbResidueNumber"("uniprotResidueNumber")_"aminoAcid" or "peptideResidue"_"aminoAcid"

CxxC motif column-- "matchedMotif"_"motifLocation"

spacc	proteinName	pdbId	peptide coverage	model_similarity	maturePeptide	NoteG1	NoteG2	CxxC motif
P52904	Pyruvate dehydrogenase E1 component subunit beta, mitochondrial	2ozl	0.95	P52904_2ozl_60_25_347_SwissModel	20_359			
O82450	Branched-chain alpha-keto acid decarboxylase E1 beta subunit	1dtw	-1	O82450_1dtw_67_30_352_SwissModel	-1_-1		249_CYS|86_CYS|	
Q9ZPX5	Succinate dehydrogenase [ubiquinone] flavoprotein subunit 2, mitochondrial	2fbw	0.99	Q9ZPX5_2fbw_67_39_632_SwissModel	30_632	175_177|	292_CYS|	
O81233	Superoxide dismutase	NoRecordInSMR						
P16048	Glycine cleavage system H protein, mitochondrial	1HTP	A_1.00	A35_A165	35_165			
P16048	Glycine cleavage system H protein, mitochondrial	1HPC	A_1.00 B_1.00	A35_A165 B35_B165	35_165		[A]_124(158)_CYS|	
P16048	Glycine cleavage system H protein, mitochondrial	1DXM	A_1.00 B_1.00	A35_A165 B35_B165	35_165			
P37225	NAD-dependent malic enzyme 59 kDa isoform, mitochondrial	NoRecordInSMR						
P26969	Glycine dehydrogenase [decarboxylating], mitochondrial	NoRecordInSMR						
Q9ZT91	Elongation factor Tu, mitochondrial	2hcj	0.07	Q9ZT91_2hcj_93_65_93_SwissModel	52_454			

Q7XK22	OSJNBa0044K18.23 protein	NoRecordInSMR						
Q9SIB9	Aconitate hydratase 2, mitochondrial	2b3x	0.96	Q9SIB9_2b3x_62_116_989_SwissModel	79_990		146_CYS|914_CYS|	CTTC_599
P68209	Succinyl-CoA ligase [ADP-forming] subunit alpha-1, mitochondrial	1euc	0.96	P68209_1euc_71_52_345_SwissModel	43_347			
P31023	Dihydrolipoyl dehydrogenase, mitochondrial	1DXL	D_1.00 A_1.00 B_1.00 C_1.00	A32_A501 B32_B501 C32_C501 D32_D501	32_501	[A, B, C, D]_45(76)_[A, B, C, D]_50(81)|		
P17614	ATP synthase subunit beta, mitochondrial	1PYV	A_-0.00	A1_A54	55_560			
Q8LBK6	Monothiol glutaredoxin-S15, mitochondrial	3ipz	0.79	Q8LBK6_3ipz_43_63_166_SwissModel	38_169		91_CYS|	
P93344	Aldehyde dehydrogenase (NAD+)	NoRecordInSMR						
P93541	Glutamate dehydrogenase	NoRecordInSMR						
P93697	CPRD12 protein	2bgk	-1	P93697_2bgk_51_12_264_SwissModel	-1_-1	68_79|	154_CYS|	
Q95P13	Secreted frizzled-related protein precusor	NoRecordInSMR						
Q9LEV3	CBS domain-containing protein CBSX3, mitochondrial	2rc3	0.8	Q9LEV3_2rc3_34_55_188_SwissModel	40_206			
Q8RWN9	Dihydrolipoyllysine-residue acetyltransferase component 2 of pyruvate dehydrogenase complex, mitochondrial	3b8k	0.44	Q8RWN9_3b8k_53_301_539_SwissModel	1_539		342_CYS|385_CYS|510_CYS|	
O48904	Malate dehydrogenase	1smk	0.97	O48904_1smk_64_32_340_SwissModel	24_343		281_CYS|	
P05493	ATP synthase subunit alpha, mitochondrial	2jdi	0.94	P05493_2jdi_73_27_501_SwissModel	1_507		390_CYS|	
P52902	Pyruvate dehydrogenase E1 component subunit alpha, mitochondrial	NoRecordInSMR						
Q9LKJ1	3-hydroxyisobutyryl-CoA hydrolase 1	3bpt	0.95	Q9LKJ1_3bpt_39_6_365_SwissModel	1_378		308_CYS|	
Q9LW15	Cytochrome c oxidase subunit 5b-1, mitochondrial	1v54	0.78	Q9LW15_1v54_37_64_158_SwissModel	56_176	122_146|146_149|122_149|		CPVC_146

Q05046	Chaperonin CPN60-2, mitochondrial	1aon	0.97	Q05046_1aon_56_33_557_SwissModel	33_575	244_305|	377_CYS|	
Q43644	NADH dehydrogenase [ubiquinone] iron-sulfur protein 1, mitochondrial	NoRecordInSMR						CRMC_111 CPIC_164 CIQC_212
O48646	Probable phospholipid hydroperoxide glutathione peroxidase 6, mitochondrial	2p5q	0.9	O48646_2p5q_72_71_230_SwissModel	55_232		105_CYS|105_CYS|153_CYS|153_CYS|	
O82514	Adenylate kinase 1	1ak2	0.88	O82514_1ak2_51_31_246_SwissModel	1_246	60_110|	58_CYS|	
P37900	Heat shock 70 kDa protein, mitochondrial	2v7y	0.84	P37900_2v7y_62_54_574_SwissModel	53_675	364_367|		CKSC_364
P42056	Mitochondrial outer membrane protein porin of 36 kDa	NoRecordInSMR						
Q9SXV0	cDNA clone:J013033A17, full insert sequence	NoRecordInSMR						
Q9FS87	Isovaleryl-CoA dehydrogenase 2, mitochondrial	1ivh	0.97	Q9FS87_1ivh_65_23_395_SwissModel	18_401	177_230|326_336|		
Q944P7	Leucine aminopeptidase 3, chloroplastic	3t8w	0.98	Q944P7_3t8w_36_77_580_SwissModel	71_583		209_CYS|368_CYS|451_CYS|468_CYS|	
Q8LAV0	Succinyl-CoA ligase beta subunit	NoRecordInSMR						
Q43153	Cysteine synthase	1z7w	0.93	Q43153_1z7w_61_47_362_SwissModel	31_368			
P34899	Serine hydroxymethyltransferase, mitochondrial	1eji	0.94	P34899_1eji_60_50_509_SwissModel	32_518			
Q9S7E4	Formate dehydrogenase, mitochondrial	3JTM	A_0.98	A34_A384	28_384			
Q9S7E4	Formate dehydrogenase, mitochondrial	3N7U	D_0.98 E_0.98 F_0.98 G_0.98 A_0.98 B_0.98 C_0.98 L_0.98 H_0.98 I_0.98 J_0.98 K_0.98	A34_A384 B34_B384 C34_C384 D34_D384 E34_E384 F34_F384 G34_G384 H34_H384 I34_I384 J34_J384 K34_K384 L34_L384	28_384		pka or asa file is not successfully generated	
Q9S7E4	Formate dehydrogenase, mitochondrial	3NAQ	A_1.00 B_1.00	A28_A384 B28_B384	28_384			

P49364	Aminomethyltransferase, mitochondrial	1wsr	0.98	P49364_1wsr_52_36_406_SwissModel	31_408			
O64530	Thiosulfate/3-mercaptopyruvate sulfurtransferase 1, mitochondrial	3olh	0.92	O64530_3olh_38_78_372_SwissModel	58_379		305_CYS|333_CYS|	
P29677	Mitochondrial-processing peptidase subunit alpha	NoRecordInSMR						
Q9LEG5	Allene oxide cylase	2brj	0.94	Q9LEG5_2brj_66_72_244_SwissModel	61_244		233_CYS|	
Q9SUK9	Probable protein phosphatase 2C 55	NoRecordInSMR						
Q95I43	MHC class Ia antigen	NoRecordInSMR						
P55312	Catalase isozyme 2	NoRecordInSMR						
P46643	Aspartate aminotransferase, mitochondrial	7aat	0.99	P46643_7aat_56_30_428_SwissModel	29_430			

Appendix E

The pre-selected proteins from ROCD in human liver mitochondrion following the criteria from Sanchez *et al* . This table shows the partial result from the raw ROCD result due to page limitation. First the protein entry without the experimentally or computationally resolved structure was removed from the raw result. Then only one record is selected when multiple PDB structures are associated with one SPACC. The NoteG1 column is for the cystein pair complying with the dithiol distance criteria. The NoteG2 column is for the cystein residue complying with the pK_a and ASA criteria. The data syntax for each column is:

peptideCoverage column-- "chainSymbol"_"peptideCoverage"

modelSimilarity column-- "chainSymbol""peptideResidueNumber"_"chainSymbol""peptideResidueNumber" or "spacc"_"pdbModelTemplate"_"sequenceSimilarity"_"modeledResidueStart"_"modeledResidueEnd"_ SwissModel

maturePeptide column-- "peptideResidueNumberStart"_"peptideResidueNumberEnd"

NoteG1 column-- [list of chains]_"pdbResidueNumber"("uniprotResidueNumber")_[list of chains]_"pdbResidueNumber"("uniprotResidueNumber") or "peptideResidue"_"peptideResidue"

NoteG2 column-- [list of chains]_"pdbResidueNumber"("uniprotResidueNumber")_"aminoAcid" or "peptideResidue"_"aminoAcid"

spacc	pdbId	peptideCoverage	modelSimilarity	maturePeptide	NoteG1	NoteG2
P08559	3EXG	E_1.00 G_1.00 A_1.00 C_1.00 M_1.00 O_1.00 I_1.00 K_1.00 3_1.00 U_1.00 1_1.00 W_1.00 Q_1.00 5_1.00 S_1.00 Y_1.00	A30_A390 C30_C390 E30_E390 G30_G390 I30_I390 K30_K390 M30_M390 O30_O390 Q30_Q390 S30_S390 U30_U390 W30_W390 Y30_Y390 130_1390 330_3390 530_5390	30_390	[3, 5, C, E, G, K, S, U, W]_65(94)_[3, 5, C, E, G, K, S, U, W]_72(101)|[E]_161(190)_[E]_189(218)|[Y]_65(94)_[Y]_72(101)|[5, M]_161(190)_[5, M]_189(218)|[I, O]_65(94)_[I, O]_72(101)|[1, I, U, W]_161(190)_[1, I, U, W]_189(218)|[1, A]_65(94)_[1, A]_72(101)|[G, K, O]_161(190)_[G, K, O]_189(218)|[M, Q]_65(94)_[M, Q]_72(101)|[A, C, Q, S, Y]_161(190)_[A, C, Q, S, Y]_189(218)|[S, Y]_152(181)_[S, Y]_161(190)|[3]_161(190)_[3]_189(218)|[W]_12(41)_[W]_232(261)|[3, 5, A, C, E, G, I, K, M, Q, U, W]_152(181)_[3, 5, A, C, E, G, I, K, M, Q, U, W]_161(190)|[G]_12(41)_[G]_232(261)|	pka or asa file is not successfully generated
P60709	3LUE	D_1.00 E_1.00 F_1.00 G_1.00 A_1.00 B_1.00 C_1.00 H_1.00 I_1.00 J_1.00	A2_A375 B2_B375 C2_C375 D2_D375 E2_E375 F2_F375 G2_G375 H2_H375 I2_I375 J2_J375	1_375	[A, B, C, D, E, F, G, H, I, J]_217(217)_[A, B, C, D, E, F, G, H, I, J]_257(257)|	pka or asa file is not successfully generated

P05091	3N80	D_1.00 E_1.00 F_1.00 G_1.00 A_1.00 B_1.00 C_1.00 H_1.00	A18_A517 B18_B517 C18_C517 D18_D517 E18_E517 F18_F517 G18_G517 H18_H517	18_517	[A, B, C, D, E, F, G, H]_301(318)_[A, B, C, D, E, F, G, H]_303(320)|	pka or asa file is not successfully generated
Q99757	1W89	D_1.00 E_1.00 F_1.00 A_1.00 B_1.00 C_1.00	A60_A166 B60_B166 C60_C166 D60_D166 E60_E166 F60_F166	60_166	[A, B, C, D, E, F]_31(90)_[A, B, C, D, E, F]_34(93)|	
Q9UKU7	1RX0	D_1.00 A_1.00 B_1.00 C_1.00	A24_A415 B24_B415 C24_C415 D24_D415	23_415	[A, B, C, D]_124(146)_[A, B, C, D]_128(150)|	
Q7Z4W1	1WNT	D_1.00 A_1.00 B_1.00 C_1.00	A1_A244 B1_B244 C1_C244 D1_D244	1_244	[A, B, C, D]_138(138)_[A, B, C, D]_150(150)|	[B, C, D]_58(58)_CYS|[A, B, C, D]_51(51)_CYS|
P26440	1IVH	D_1.00 A_1.00 B_1.00 C_1.00	A30_A423 B30_B423 C30_C423 D30_D423	30_423	[A, B, C, D]_318(347)_[A, B, C, D]_323(352)|[A, B, C, D]_307(336)_[A, B, C, D]_338(367)|[B]_323(352)_[B]_328(357)|[D]_318(347)_[D]_328(357)|[C]_323(352)_[C]_328(357)|[C]_318(347)_[C]_328(357)|[D]_323(352)_[D]_328(357)|[A, B]_318(347)_[A, B]_328(357)|[A]_323(352)_[A]_328(357)|	
P07954	30000	D_1.00 A_1.00 B_1.00 C_1.00	A44_A510 B44_B510 C44_C510 D44_D510	45_510	[A, B, C, D]_377(377)_[A, B, C, D]_434(434)|	
P30044	2VL9	D_0.68 A_0.68 B_0.68 C_0.68	A2_A162 B2_B162 C2_C162 D2_D162	53_214	[A, B, C, D]_47(48)_[A, B, C, D]_151(152)|	
P49327	2JFK	D_0.16 A_0.16 B_0.16 C_0.16	A422_A831 B422_B831 C422_C831 D422_D831	1_2511	[A, B, C, D]_496(496)_[A, B, C, D]_587(587)|[A, B, C]_630(630)_[A, B, C]_634(634)|	
O94925	3UO9	D_0.81 A_0.81 B_0.81 C_0.81	A71_A598 B71_B598 C71_C598 D71_D598	17_669	[A, B, C, D]_500(500)_[A, B, C, D]_518(518)|	
P11498	3BG3	D_0.60 A_0.60 B_0.60 C_0.60	A482_A1178 B482_B1178 C482_C1178 D482_D1178	21_1178	[A, B, C, D]_752(752)_[A, B, C, D]_786(786)|	[A, B, C, D]_622(622)_CYS|
O15294	3PE3	D_0.69 A_0.69 B_0.69 C_0.69	A323_A1041 B323_B1041 C323_C1041 D323_D1041	2_1046	[A, B, C, D]_835(845)_[A, B, C, D]_911(921)|	[A, B, C, D]_313(323)_CYS|[A, B, C, D]_610(620)_CYS|[A, B, C, D]_748(758)_CYS|
P11310	1T9G	D_1.00 A_1.00 B_1.00 C_1.00	A26_A421 B26_B421 C26_C421 D26_D421	26_421	[A, B, C, D]_91(116)_[A, B, C, D]_219(244)|	[A]_70(95)_CYS|

P10109	3P1M	D_1.00 E_1.00 F_1.00 G_1.00 A_1.00 B_1.00 C_1.00 H_1.00	A61_A184 B61_B184 C61_C184 D61_D184 E61_E184 F61_F184 G61_G184 H61_H184	61_184	[A, B, C, E, F, G]_115(115)_[A, B, C, E, F, G]_152(152)||[B, C, G]_106(106)_[B, C, G]_112(112)||[D]_115(115)_[D]_152(152)||[F]_106(106)_[F]_112(112)||[H]_115(115)_[H]_152(152)||[A, D, E, H]_106(106)_[A, D, E, H]_112(112)||[A, B, C, D, E, F, G, H]_112(112)_[A, B, C, D, E, F, G, H]_115(115)||[A, B, C, D, E, F, G, H]_106(106)_[A, B, C, D, E, F, G, H]_152(152)||[B, D, E, F, H]_106(106)_[B, D, E, F, H]_115(115)||[C]_112(112)_[C]_152(152)|	[A, B, C]_155(155)_CYS|
P33316	3EHW	A_-1.00 B_-1.00 C_-1.00 Y_-1.00 X_-1.00 Z_-1.00	inconsis. annot. btw UniProt and PDB	70_252	[A, B, C, X, Y, Z]_78(78)_[A, B, C, X, Y, Z]_134(134)|	
O75880	1WP0	A_0.54 B_0.54 C_0.54	A138_A301 B138_B301 C138_C301	1_301	[A, B, C]_169(169)_[A, B, C]_173(173)|	
P33316	2HQU	A_0.87 B_0.87 C_0.87	A94_A252 B94_B252 C94_C252	70_252	[A, B, C]_78(171)_[A, B, C]_134(227)|	
P05091	3SZ9	D_1.00 E_1.00 F_1.00 G_1.00 A_1.00 B_1.00 C_1.00 H_1.00	A18_A517 B18_B517 C18_C517 D18_D517 E18_E517 F18_F517 G18_G517 H18_H517	18_517	[A, B, D, E, F, G]_301(318)_[A, B, D, E, F, G]_303(320)||[A, D]_301(318)_[A, D]_302(319)||[C]_302(319)_[C]_303(320)|	pka or asa file is not successfully generated
P78310	2W9L	T_0.34 V_0.34 G_0.34 P_0.34 A_0.34 B_0.34 O_0.34 Y_0.34 X_0.34 J_0.34 Z_0.34 K_0.34	A16_A139 B16_B139 G16_G139 J16_J139 K16_K139 O16_O139 P16_P139 T16_T139 V16_V139 X16_X139 Y16_Y139 Z16_Z139	20_365	[A, B, G, J, K, O, P, T, V, X, Y, Z]_41(41)_[A, B, G, J, K, O, P, T, V, X, Y, Z]_120(120)|	pka or asa file is not successfully generated
P05091	1NZW	D_1.00 E_1.00 F_1.00 G_1.00 A_1.00 B_1.00 C_1.00 H_1.00	A18_A517 B18_B517 C18_C517 D18_D517 E18_E517 F18_F517 G18_G517 H18_H517	18_517	[A, B, G]_301(318)_[A, B, G]_303(320)||[E]_302(319)_[E]_303(320)||[A]_301(318)_[A]_302(319)||[H]_302(319)_[H]_303(320)||[B]_301(318)_[B]_302(319)||[A, B, C, D, G]_302(319)_[A, B, C, D, G]_303(320)||[G]_301(318)_[G]_302(319)||[F]_302(319)_[F]_303(320)||[C, D, E, F, H]_301(318)_[C, D, E, F, H]_303(320)|	pka or asa file is not successfully generated

P00325	3HUD	A_1.00 B_1.00	A1_A374 B1_B374	2_375	[A, B]_100(100)_[A, B]_103(103)|[A]_97(97)_[A]_111(111)|[B]_46(46)_[B]_174(174)|[A]_97(97)_[A]_100(100)|[B]_97(97)_[B]_111(111)|[B]_97(97)_[B]_100(100)|[A]_46(46)_[A]_174(174)|[A]_103(103)_[A]_111(111)|[A]_97(97)_[A]_103(103)|[B]_103(103)_[B]_111(111)|[B]_97(97)_[B]_103(103)|[B]_195(195)_[B]_211(211)| [A, B]_100(100)_[A, B]_111(111)|[A]_195(195)_[A]_211(211)|[A]_281(281)_[A]_282(282)|	[B]_240(240)_CYS|[B]_286(286)_CYS|[A]_132(132)_CYS|
P45381	2Q51	A_1.00 B_1.00	A2_A313 B2_B313	1_313	[A, B]_124(124)_[A, B]_152(152)|	[A, B]_276(276)_CYS|[A, B]_310(310)_CYS|
Q7Z4W1	1PR9	A_1.00 B_1.00	A1_A244 B1_B244	1_244	[A, B]_138(138)_[A, B]_150(150)|	[B]_58(58)_CYS|[A, B]_51(51)_CYS|
P17735	3DYD	A_0.89 B_0.89	A41_A444 B41_B444	1_454	[A, B]_151(151)_[A, B]_275(275)|B_320(320)_A_320(320)|[A, B]_353(353)_[A, B]_431(431)|[A, B]_144(144)_[A, B]_275(275)|	[A, B]_444(444)_CYS|[B]_241(241)_CYS|[A, B]_258(258)_CYS|[A, B]_128(128)_CYS|
P11586	1A4I	A_0.32 B_0.32	A1_A300 B1_B300	2_935	[A, B]_152(151)_[A, B]_236(235)|	
O75880	2GGT	A_0.54 B_0.54	A135_A298 B135_B298	1_301	[A, B]_169(169)_[A, B]_173(173)|	
P00325	1U3V	A_1.00 B_1.00	A1_A374 B1_B374	2_375	[A, B]_195(195)_[A, B]_211(211)|[A]_103(103)_[A]_111(111)|[A, B]_97(97)_[A, B]_111(111)|[B]_103(103)_[B]_111(111)|[A, B]_100(100)_[A, B]_103(103)|[A, B]_97(97)_[A, B]_100(100)|[A, B]_97(97)_[A, B]_103(103)|[A, B]_100(100)_[A, B]_111(111)|[A, B]_46(46)_[A, B]_174(174)|	
P21695	1X0V	A_1.00 B_1.00	A1_A348 B1_B348	2_349	[A, B]_200(200)_[A, B]_256(256)|	[A, B]_168(168)_CYS|
P19367	1QHA	A_1.00 B_1.00	A1_A917 B1_B917	1_917	[A, B]_237(237)_[A, B]_256(256)|[A, B]_685(685)_[A, B]_704(704)|[A, B]_665(665)_[A, B]_672(672)|[A, B]_217(217)_[A, B]_224(224)|	[A, B]_581(581)_CYS|[B]_133(133)_CYS|[A]_717(717)_CYS|[A, B]_813(813)_CYS|
Q9NS18	2HT9	A_0.85 B_0.85	A41_A164 B41_B164	20_164	[A, B]_28(68)_[A, B]_113(153)|[A, B]_37(77)_[A, B]_40(80)|	
O15382	1EKV	A_1.00 B_1.00	A28_A392 B28_B392	28_392	[A, B]_315(342)_[A, B]_318(345)|	
P35558	2GMV	A_1.00 B_1.00	A1_A622 B1_B622	1_622	[A, B]_38(38)_[A, B]_133(133)|[A, B]_307(307)_[A, B]_413(413)|[A, B]_198(198)_[A, B]_212(212)|	[A, B]_606(606)_CYS|

P30405	2Z6W	A_0.92 B_0.92	A44_A207 B44_B207	30_207	[A, B]_40(82)_[A, B]_161(203)|	
P78310	1F5W	A_0.35 B_0.35	A15_A140 B15_B140	20_365	[A, B]_41(41)_[A, B]_120(120)|	
Q8IWL3	3BVO	A_1.00 B_1.00	A30_A235 B30_B235	30_235	[A, B]_44(44)_[A, B]_61(61)|[A]_58(58)_[A]_61(61)|[B]_44(44)_[B]_58(58)|[B]_41(41)_[B]_58(58)|[A]_44(44)_[A]_58(58)|[B]_58(58)_[B]_61(61)|[A]_41(41)_[A]_61(61)|[A]_41(41)_[A]_58(58)|[A]_41(41)_[A]_44(44)|[B]_41(41)_[B]_61(61)|[B]_41(41)_[B]_44(44)|	
P46063	2V1X	A_0.88 B_0.88	A49_A616 B49_B616	2_649	[A, B]_453(453)_[A, B]_478(478)|[A, B]_475(475)_[A, B]_478(478)|[A, B]_453(453)_[A, B]_471(471)|[A, B]_471(471)_[A, B]_478(478)|[A, B]_453(453)_[A, B]_475(475)|[A, B]_471(471)_[A, B]_475(475)|	[A, B]_270(270)_CYS|[A]_223(223)_CYS|[A, B]_479(479)_CYS|
P00325	1HDX	A_1.00 B_1.00	A1_A374 B1_B374	2_375	[A, B]_46(46)_[A, B]_174(174)|[B]_97(97)_[B]_103(103)|[A]_103(103)_[A]_111(111)|[A]_100(100)_[A]_103(103)|[B]_103(103)_[B]_111(111)|[A]_97(97)_[A]_103(103)|[B]_97(97)_[B]_111(111)|[B]_100(100)_[B]_103(103)|[B]_97(97)_[B]_100(100)|[A]_97(97)_[A]_111(111)|[A]_195(195)_[A]_211(211)|[A, B]_100(100)_[A, B]_111(111)|[A]_97(97)_[A]_100(100)|[B]_195(195)_[B]_211(211)|[A, B]_281(281)_[A, B]_282(282)|	
P22033	3BIC	A_1.00 B_1.00	A12_A750 B12_B750	33_750	[A, B]_533(533)_[A, B]_560(560)|	[A, B]_742(742)_CYS|
Q15149	3PE0	A_0.06 B_0.06	A750_A1028 B750_B1028	1_4684	[A, B]_740(850)_[A, B]_840(950)|[A, B]_840(950)_[A, B]_882(992)|[A, B]_740(850)_[A, B]_882(992)|	[A, B]_778(888)_CYS|[A, B]_739(849)_CYS|
P07203	2F8A	A_0.91 B_0.91	A12_A196 B12_B196	1_203	[A, B]_76(76)_[A, B]_113(113)|	[A, B]_154(154)_CYS|
P49327	3HHD	D_0.38 A_0.38 B_0.38 C_0.38	A2_A963 B2_B963 C2_C963 D2_D963	1_2511	[A, C, D]_496(496)_[A, C, D]_587(587)|[D]_630(630)_[D]_634(634)|[B]_496(496)_[B]_587(587)|[A, B, C]_630(630)_[A, B, C]_634(634)|[A, B, C, D]_212(212)_[A, B, C, D]_223(223)|	pka or asa file is not successfully generated
Q14790	1QDU	E_0.32 G_0.32 A_0.32 C_0.32 I_0.32 K_0.32	A222_A374 C222_C374 E222_E374 G222_G374 I222_I374 K222_K374	1_479	[A, C, E, G, I, K]_165(238)_[A, C, E, G, I, K]_233(306)|	
P07858	1HUC	A_0.14 C_0.14	A80_A126 C80_C126	18_339	[A, C]_14(93)_[A, C]_43(122)|	[A, C]_29(108)_CYS|
P21549	3R9A	A_1.00 C_1.00	A1_A392 C1_C392	1_392	[A, C]_173(173)_[A, C]_178(178)|	

Q14790	2Y1L	A_0.33 C_0.33	A218_A374 C218_C374	1_479	[A, C]_236(236)_[A, C]_313(313)|	
O15294	3TAX	A_0.69 C_0.69	A323_A1041 C323_C1041	2_1046	[A, C]_835(845)_[A, C]_911(921)|	[A, C]_610(620)_CYS|
P30101	3F8U	A_1.00 C_1.00	A25_A505 C25_C505	25_505	[A, C]_85(85)_[A, C]_92(92)|[A, C]_406(406)_[A, C]_409(409)|	pka or asa file is not successfully generated
P07858	1CSB	D_0.14 A_0.14	A80_A126 D80_D126	18_339	[A, D]_14(93)_[A, D]_43(122)|	
P08559	3EXF	E_1.00 G_1.00 A_1.00 C_1.00	A30_A390 C30_C390 E30_E390 G30_G390	30_390	[A, E, G]_65(94)_[A, E, G]_72(101)|[A, E, G]_161(190)_[A, E, G]_189(218)|[C]_65(94)_[C]_72(101)|[C]_161(190)_[C]_189(218)|[C, E]_12(41)_[C, E]_232(261)|[A, E, G]_152(181)_[A, E, G]_161(190)|	
Q9Y5J7	2BSK	E_1.00 A_1.00 C_1.00	A1_A89 C1_C89 E1_E89	2_89	[A, E]_32(32)_[A, E]_48(48)|[A, C, E]_28(28)_[A, C, E]_52(52)|[C]_32(32)_[C]_48(48)|[A, C, E]_32(32)_[A, C, E]_52(52)|[A, C]_28(28)_[A, C]_32(32)|[A]_48(48)_[A]_52(52)|[E]_28(28)_[E]_32(32)|	
P08559	3EXE	E_1.00 G_1.00 A_1.00 C_1.00	A30_A390 C30_C390 E30_E390 G30_G390	30_390	[A, G]_12(41)_[A, G]_232(261)|[A, C, E, G]_65(94)_[A, C, E, G]_72(101)|[A, C, E, G]_161(190)_[A, C, E, G]_189(218)|[A, E, G]_152(181)_[A, E, G]_161(190)|	[C]_152(181)_CYS|
P07858	1PBH	A_0.98	A18_A333	18_339	[A]_100(116)_[A]_132(148)|[A]_14(30)_[A]_43(59)|[A]_26(42)_[A]_71(87)|[A]_108(124)_[A]_119(135)|[A]_63(79)_[A]_67(83)|[A]_62(78)_[A]_128(144)|	[A]_240(256)_CYS|
P62995	2RRB	A_-1.00	inconsis. annot. btw UniProt and PDB	1_288	[A]_118(118)_[A]_119(119)|	
P05981	1Z8G	A_0.89	A46_A417	1_417	[A]_119(119)_[A]_138(138)|[A]_322(322)_[A]_338(338)|[A]_77(77)_[A]_140(140)|[A]_90(90)_[A]_150(150)|[A]_188(188)_[A]_204(204)|[A]_349(349)_[A]_381(381)|[A]_291(291)_[A]_359(359)|[A]_153(153)_[A]_277(277)|[A]_359(359)_[A]_372(372)|[A]_77(77)_[A]_119(119)|[A]_77(77)_[A]_138(138)|[A]_138(138)_[A]_140(140)|[A]_119(119)_[A]_140(140)|[A]_291(291)_[A]_372(372)|	
P08559	2OZL	A_1.00 C_1.00	A30_A390 C30_C390	30_390	[A]_12(41)_[A]_232(261)|[A, C]_65(94)_[A, C]_72(101)|[A, C]_161(190)_[A, C]_189(218)|[C]_152(181)_[C]_161(190)|	[A]_152(181)_CYS|
Q6FGP5	3MK4	A_0.89	A41_A373	1_373	[A]_123(123)_[A]_283(283)|	

O43819	2RLI	A_0.74	A100_A266	42_266	[A]_133(133)_[A]_137(137)|	[A]_115(115)_CYS|
P35670	2ROP	A_0.14	A238_A439	1_1465	[A]_133(370)_[A]_136(373)|	[A]_31(268)_CYS|[A]_68(305)_CYS|[A]_121(358)_CYS|
P04075	1ALD	A_1.00	A1_A363	2_364	[A]_134(134)_[A]_177(177)|	[A]_338(338)_CYS|[A]_239(239)_CYS|[A]_72(72)_CYS|[A]_201(201)_CYS|
P07858	2IPP	A_0.14	A80_A126	18_339	[A]_14(93)_[A]_43(122)|	
Q15019	2QAG	A_1.00	A1_A361	1_361	[A]_148(148)_[A]_149(149)|	
P11586	1DIG	A_0.33 B_0.33	A1_A306 B1_B306	2_935	[A]_152(152)_[A]_236(236)|[B]_1152(152)_[B]_1236(236)|	
P50336	3NKS	A_1.00	A1_A477	1_477	[A]_167(167)_[A]_183(183)|	[A]_382(382)_CYS|[A]_26(26)_CYS|[A]_258(258)_CYS|
O95050	2A14	A_1.00	A1_A263	1_263	[A]_168(168)_[A]_213(213)|[A]_44(44)_[A]_254(254)|	[A]_181(181)_CYS|[A]_141(141)_CYS|[A]_171(171)_CYS|[A]_115(115)_CYS|
O75880	2GQM	A_0.56	A132_A301	1_301	[A]_169(169)_[A]_173(173)|	
P21549	1J04	A_1.00	A1_A392	1_392	[A]_173(173)_[A]_178(178)|	
P49916	1UW0	A_0.12	A1_A117	1_1009	[A]_18(18)_[A]_21(21)|[A]_21(21)_[A]_55(55)|[A]_18(18)_[A]_55(55)|	[A]_7(7)_CYS|
P00325	1HSZ	A_1.00 B_1.00	A1_A374 B1_B374	2_375	[A]_195(195)_[A]_211(211)|[B]_100(100)_[B]_103(103)|[A]_103(103)_[A]_111(111)|[B]_195(195)_[B]_211(211)|[B]_97(97)_[B]_100(100)|[A]_100(100)_[A]_103(103)|[B]_97(97)_[B]_111(111)|[B]_103(103)_[B]_111(111)|[A]_97(97)_[A]_111(111)|[A]_97(97)_[A]_103(103)|[A]_97(97)_[A]_100(100)|[B]_97(97)_[B]_103(103)|[B]_46(46)_[B]_174(174)|[A,B]_100(100)_[A,B]_111(111)|[A]_46(46)_[A]_174(174)|[B]_281(281)_[B]_282(282)|	
P21695	1X0X	A_1.00	A1_A349	2_349	[A]_200(200)_[A]_256(256)|	[A]_168(168)_CYS|
P35557	1GLK	A_1.00	A1_A465	1_465	[A]_213(213)_[A]_220(220)|[A]_233(233)_[A]_252(252)|[A]_230(230)_[A]_382(382)|[A]_457(457)_[A]_461(461)|	[A]_129(129)_CYS|
P60709	3BYH	A_-1.00	inconsis. annot. btw UniProt and PDB	1_375	[A]_217(217)_[A]_257(257)|	[A]_272(272)_CYS|
Q8IYB8	3RC3	A_0.88	A47_A722	23_786	[A]_230(230)_[A]_274(274)|[A]_230(230)_[A]_279(279)|[A]_274(274)_[A]_279(279)|	[A]_587(587)_CYS|

P35557	4DHY	A_0.98	A12_A465	1_465	[A]_230(230)_[A]_382(382)|[A]_233(233)_[A]_252(252)|[A]_213(213)_[A]_220(220)|	[A]_457(457)_CYS|[A]_129(129)_CYS|[A]_364(364)_CYS|
O00411	3SPA	A_0.95	A105_A1230	42_1230	[A]_232(232)_[A]_233(233)|	[A]_472(472)_CYS|[A]_446(446)_CYS|[A]_535(535)_CYS|[A]_726(726)_CYS|[A]_591(591)_CYS|[A]_413(413)_CYS|[A]_398(398)_CYS|
P35557	4DCH	A_1.00	A1_A465	1_465	[A]_233(233)_[A]_252(252)|[A]_230(230)_[A]_382(382)|[A]_213(213)_[A]_220(220)|	[A]_461(461)_CYS|[A]_129(129)_CYS|[A]_364(364)_CYS|
Q14790	3KJQ	A_0.34	A211_A374	1_479	[A]_236(236)_[A]_313(313)|	
P19367	1HKC	A_1.00	A1_A917	1_917	[A]_237(237)_[A]_256(256)|[A]_665(665)_[A]_672(672)|[A]_685(685)_[A]_704(704)|[A]_217(217)_[A]_224(224)|	[A]_581(581)_CYS|[A]_133(133)_CYS|[A]_813(813)_CYS|
P53370	3H95	A_0.56	A141_A316	1_316	[A]_243(243)_[A]_246(246)|	[A]_239(239)_CYS|
Q9NS18	2FLS	A_0.75	A56_A164	20_164	[A]_28(68)_[A]_113(153)|[A]_37(77)_[A]_40(80)|	
Q969S2	1VZP	A_0.38	A192_A319	2_332	[A]_291(291)_[A]_318(318)|[A]_291(291)_[A]_315(315)|[A]_315(315)_[A]_318(318)|	
P53396	3PFF	A_0.74	A1_A817	1_1101	[A]_293(293)_[A]_748(748)|[A]_742(742)_[A]_744(744)|	[A]_20(20)_CYS|[A]_633(633)_CYS|
Q7Z5U7	3PDF	A_1.00	A25_A463	25_463	[A]_297(321)_[A]_313(337)|[A]_30(54)_[A]_112(136)|[A]_231(255)_[A]_274(298)|[A]_6(30)_[A]_94(118)|[A]_267(291)_[A]_307(331)|	
O15382	1KTA	A_1.00 B_1.00	A28_A392 B28_B392	28_392	[A]_315(342)_[A]_318(345)|[B]_815(342)_[B]_818(345)|	
Q9H3N1	1X5E	A_0.44	A30_A142	27_280	[A]_34(56)_[A]_37(59)|	
Q96HE7	3AHQ	A_1.00	A22_A468	24_468	[A]_35(35)_[A]_48(48)|[A]_37(37)_[A]_46(46)|[A]_208(208)_[A]_241(241)|[A]_85(85)_[A]_391(391)|[A]_394(394)_[A]_397(397)|	
P26358	3EPZ	A_0.15 B_0.15	A351_A600 B351_B600	1_1616	[A]_353(353)_[A]_356(356)|[A]_356(356)_[A]_414(414)|[A, B]_353(353)_[A, B]_414(414)|[B]_356(356)_[B]_414(414)|[A]_356(356)_[A]_420(420)|[B]_353(353)_[B]_356(356)|[B]_356(356)_[B]_420(420)|[A, B]_414(414)_[A, B]_420(420)|	[A, B]_409(409)_CYS|[A]_580(580)_CYS|

Q14192	1X4K	A_0.21	A101_A159	1_279	[A]_36(129)_[A]_57(150)|[A]_36(129)_[A]_39(132)|[A]_39(132)_[A]_57(150)|[A]_8(101)_[A]_11(104)|[A]_11(104)_[A]_33(126)|[A]_8(101)_[A]_33(126)|[A]_39(132)_[A]_60(153)|[A]_36(129)_[A]_60(153)|[A]_57(150)_[A]_60(153)|	
P07237	1MEK	A_0.24	A18_A137	18_508	[A]_36(53)_[A]_39(56)|	
P35558	1KHF	A_1.00	A1_A622	1_622	[A]_38(38)_[A]_133(133)|[A]_307(307)_[A]_413(413)|[A]_198(198)_[A]_212(212)|	[A]_606(606)_CYS|
P41250	2PME	A_0.93	A55_A739	1_739	[A]_388(442)_[A]_390(444)|	[A]_126(180)_CYS|[A]_157(211)_CYS|[A]_177(231)_CYS|[A]_412(466)_CYS|[A]_101(155)_CYS|[A]_417(471)_CYS|
P07237	3UEM	A_0.70	A137_A479	18_508	[A]_397(397)_[A]_400(400)|	
Q14192	1X4L	A_0.21	A221_A279	1_279	[A]_41(254)_[A]_59(272)|[A]_8(221)_[A]_11(224)|[A]_11(224)_[A]_35(248)|[A]_59(272)_[A]_62(275)|[A]_38(251)_[A]_41(254)|[A]_38(251)_[A]_62(275)|[A]_8(221)_[A]_35(248)|[A]_41(254)_[A]_62(275)|[A]_38(251)_[A]_59(272)|	
P78310	1RSF	A_0.36	A21_A144	20_365	[A]_41(41)_[A]_120(120)|	
P54819	2C9Y	A_1.00	A1_A239	2_239	[A]_42(42)_[A]_92(92)|	[A]_40(40)_CYS|[A]_232(232)_CYS|
P05981	1P57	A_0.27	A46_A159	1_417	[A]_45(90)_[A]_105(150)|[A]_74(119)_[A]_93(138)|[A]_32(77)_[A]_95(140)|[A]_32(77)_[A]_74(119)|[A]_32(77)_[A]_93(138)|[A]_93(138)_[A]_95(140)|[A]_74(119)_[A]_95(140)|	
P35557	3VEY	A_0.97	A16_A465	1_465	[A]_457(457)_[A]_461(461)|[A]_233(233)_[A]_252(252)|[A]_230(230)_[A]_382(382)|[A]_213(213)_[A]_220(220)|	[A]_364(364)_CYS|
P07339	1LYA	A_0.24 C_0.24	A65_A161 C65_C161	19_412	[A]_46(110)_[A]_53(117)|[C]_27(91)_[C]_96(160)|[C]_46(110)_[C]_53(117)|[A]_27(91)_[A]_96(160)|	
Q05655	2YUU	A_0.10	A149_A218	1_676	[A]_48(189)_[A]_51(192)|[A]_34(175)_[A]_59(200)|[A]_31(172)_[A]_34(175)|[A]_31(172)_[A]_59(200)|[A]_48(189)_[A]_67(208)|[A]_51(192)_[A]_67(208)|	
O94925	3CZD	A_0.48	A221_A533	17_669	[A]_500(500)_[A]_518(518)|	
P22033	2XIJ	A_1.00	A12_A750	33_750	[A]_533(533)_[A]_560(560)|	[A]_471(471)_CYS|[A]_742(742)_CYS|

P49748	3B96	A_0.95	A69_A655	41_655	[A]_563(603)_[A]_567(607)|	[A]_116(156)_CYS|[A]_197(237)_CYS|
P30101	2DMM	A_0.27	A357_A485	25_505	[A]_57(406)_[A]_60(409)|	
P48147	3DDU	A_1.00	A2_A710	1_710	[A]_573(573)_[A]_703(703)|	[A]_57(57)_CYS|[A]_255(255)_CYS|[A]_25(25)_CYS|[A]_601(601)_CYS|
P48449	1W6K	A_1.00	A1_A732	2_732	[A]_584(584)_[A]_636(636)|	[A]_676(676)_CYS|[A]_6(6)_CYS|[A]_484(484)_CYS|
Q7Z5U7	1K3B	A_0.27	A25_A143	25_463	[A]_6(30)_[A]_94(118)|[A]_30(54)_[A]_112(136)|	
P49748	2UXW	A_0.95	A72_A655	41_655	[A]_603(603)_[A]_607(607)|	[A]_156(156)_CYS|[A]_237(237)_CYS|
P30101	2ALB	A_0.23	A25_A137	25_505	[A]_61(85)_[A]_68(92)|	[A]_33(57)_CYS|[A]_36(60)_CYS|
P49916	3L2P	A_0.57	A257_A833	1_1009	[A]_628(715)_[A]_637(724)|[A]_642(729)_[A]_720(807)|[A]_183(270)_[A]_324(411)|	[A]_174(261)_CYS|[A]_524(611)_CYS|
P07858	3PBH	A_0.98	A18_A333	18_339	[A]_63(79)_[A]_67(83)|[A]_108(124)_[A]_119(135)|[A]_26(42)_[A]_71(87)|[A]_100(116)_[A]_132(148)|[A]_62(78)_[A]_128(144)|[A]_14(30)_[A]_43(59)|	[A]_240(256)_CYS|
P08559	3EXI	A_1.00	A30_A390	30_390	[A]_65(94)_[A]_72(101)|[A]_12(41)_[A]_232(261)|[A]_161(190)_[A]_189(218)|[A]_152(181)_[A]_161(190)|	
P26358	3SWR	A_0.62	A601_A1600	1_1616	[A]_670(670)_[A]_686(686)|[A]_820(820)_[A]_896(896)|[A]_656(656)_[A]_659(659)|[A]_1476(1476)_[A]_1478(1478)|[A]_659(659)_[A]_691(691)|[A]_664(664)_[A]_686(686)|[A]_1478(1478)_[A]_1485(1485)|[A]_653(653)_[A]_691(691)|[A]_893(893)_[A]_896(896)|[A]_820(820)_[A]_893(893)|[A]_1476(1476)_[A]_1485(1485)|[A]_664(664)_[A]_667(667)|[A]_656(656)_[A]_691(691)|[A]_667(667)_[A]_686(686)|[A]_653(653)_[A]_656(656)|[A]_653(653)_[A]_659(659)|[A]_664(664)_[A]_670(670)|[A]_667(667)_[A]_670(670)|	[A]_762(762)_CYS|[A]_751(751)_CYS|
Q16775	2F50	A_0.55	A15_A177	14_308	[A]_71(63)_[A]_155(147)|	
Q14192	2D8Z	A_-1.00	inconsis. annot. btw UniProt and PDB	1_279	[A]_8(8)_[A]_11(11)|[A]_34(34)_[A]_37(37)|[A]_34(34)_[A]_58(58)|[A]_37(37)_[A]_55(55)|[A]_37(37)_[A]_58(58)|[A]_55(55)_[A]_58(58)|[A]_34(34)_[A]_55(55)|[A]_8(8)_[A]_31(31)|[A]_11(11)_[A]_31(31)|	[A]_60(60)_CYS|
P30405	3RDB	A_0.93	A43_A207	30_207	[A]_82(82)_[A]_203(203)|	

P35670	2EW9	A_0.10	A486_A633	1_1465	[A]_91(575)_[A]_94(578)|	[A]_18(502)_CYS|
P05091	1ZUM	D_1.00 E_1.00 F_1.00 G_1.00 A_1.00 B_1.00 C_1.00 L_1.00 H_1.00 I_1.00 J_1.00 K_1.00	A18_A517 B18_B517 C18_C517 D18_D517 E18_E517 F18_F517 G18_G517 H18_H517 I18_I517 J18_J517 K18_K517 L18_L517	18_517	[B, C, D, E, F, G, H, I, J, K, L]_301(318)_[B, C, D, E, F, G, H, I, J, K, L]_303(320)||[B, C, D, E, F, G, H, I, J, K, L]_302(319)_[B, C, D, E, F, G, H, I, J, K, L]_303(320)||[B, C, D, E, F, G, H, I, J, K, L]_301(318)_[B, C, D, E, F, G, H, I, J, K, L]_302(319)|	pka or asa file is not successfully generated
P07858	3K9M	A_0.79 B_0.79	A80_A333 B80_B333	18_339	[B]_100(179)_[B]_132(211)|[A]_108(187)_[A]_119(198)|[A, B]_26(105)_[A, B]_71(150)|[A, B]_63(142)_[A, B]_67(146)|[B]_108(187)_[B]_119(198)|[A]_100(179)_[A]_132(211)|[A, B]_62(141)_[A, B]_128(207)|[A, B]_14(93)_[A, B]_43(122)|	[B]_240(319)_CYS|
P00325	1DEH	A_1.00 B_1.00	A1_A374 B1_B374	2_375	[B]_103(103)_[B]_111(111)|[B]_97(97)_[B]_111(111)|[A]_103(103)_[A]_111(111)|[B]_100(100)_[B]_103(103)|[A]_97(97)_[A]_111(111)|[B]_46(46)_[B]_174(174)|[B]_195(195)_[B]_211(211)|[A]_100(100)_[A]_103(103)|[A]_195(195)_[A]_211(211)|[B]_97(97)_[B]_100(100)|[B]_97(97)_[B]_103(103)|[A]_100(100)_[A]_111(111)|[A]_97(97)_[A]_100(100)|[B]_100(100)_[B]_111(111)|[A]_97(97)_[A]_103(103)|[A]_46(46)_[A]_174(174)|[A]_281(281)_[A]_282(282)|	
P11586	1DIB	A_0.33 B_0.33	A1_A306 B1_B306	2_935	[B]_1152(152)_[B]_1236(236)|[A]_152(152)_[A]_236(236)|	
Q7Z4W1	3D3W	A_1.00 B_1.00	A1_A244 B1_B244	1_244	[B]_138(138)_[B]_150(150)|	[A, B]_51(51)_CYS|
P09668	1BZN	B_0.70	B116_B335	23_335	[B]_157(272)_[B]_207(322)|[B]_23(138)_[B]_66(181)|[B]_57(172)_[B]_99(214)|	

P05981	3T2N	A_0.89 B_0.89	A46_A417 B46_B417	1_417	[B]_188(188)_[B]_204(204)|[B]_77(77)_[B]_140(140)|[A]_119(119)_[A]_138(138)|[A, B]_322(322)_[A, B]_338(338)|[B]_119(119)_[B]_138(138)|[A]_153(153)_[A]_277(277)|[A]_291(291)_[A]_359(359)|[B]_153(153)_[B]_277(277)|[A]_90(90)_[A]_150(150)|[A]_188(188)_[A]_204(204)|[A]_77(77)_[A]_140(140)|[B]_291(291)_[B]_359(359)|[B]_90(90)_[B]_150(150)|[A, B]_359(359)_[A, B]_372(372)|[B]_138(138)_[B]_140(140)|[A, B]_77(77)_[A, B]_119(119)|[A, B]_77(77)_[A, B]_138(138)|[B]_119(119)_[B]_140(140)|[A, B]_291(291)_[A, B]_372(372)|[A]_138(138)_[A]_140(140)|[A]_119(119)_[A]_140(140)|	
Q14790	3H11	B_0.55	B217_B479	1_479	[B]_221(236)_[B]_298(313)|	[B]_394(409)_CYS|
P78310	1KAC	B_0.36	B21_B144	20_365	[B]_43(41)_[B]_122(120)|	
Q15027	3JUE	A_0.49 B_0.49	A378_A740 B378_B740	1_740	[B]_440(440)_[B]_443(443)|[A, B]_423(423)_[A, B]_440(440)|[A, B]_420(420)_[A, B]_443(443)|[A]_440(440)_[A]_443(443)|[B]_420(420)_[B]_440(440)|[B]_423(423)_[B]_443(443)|[A]_420(420)_[A]_440(440)|[B]_420(420)_[B]_423(423)|[A]_423(423)_[A]_443(443)|[A]_420(420)_[A]_423(423)|	[A, B]_657(657)_CYS|[A, B]_501(501)_CYS|
P00325	1HDZ	A_1.00 B_1.00	A1_A374 B1_B374	2_375	[B]_46(46)_[B]_174(174)|[A, B]_97(97)_[A, B]_111(111)|[A, B]_103(103)_[A, B]_111(111)|[B]_100(100)_[B]_103(103)|[A, B]_100(100)_[A, B]_111(111)|[A]_46(46)_[A]_174(174)|[A]_97(97)_[A]_100(100)|[A]_195(195)_[A]_211(211)|[B]_97(97)_[B]_103(103)|[B]_97(97)_[B]_100(100)|[A]_100(100)_[A]_103(103)|[A]_97(97)_[A]_103(103)|[B]_195(195)_[B]_211(211)|[A, B]_281(281)_[A, B]_282(282)|	
Q9Y2A7	3P8C	B_1.00	B1_B1128	1_1128	[B]_573(573)_[B]_599(599)|[B]_543(543)_[B]_579(579)|	pka or asa file is not successfully generated
P54886	2H5G	A_0.55 B_0.55	A362_A795 B362_B795	1_795	[B]_584(584)_[B]_612(612)|	[A]_612(612)_CYS|

Q9NNW7	1W1E	A_1.00 B_0.94	A36_A524 B11_B497	37_524	[B]_61(61)_[B]_66(66)|[A]_61(86)_[A]_66(91)|	[A]_254(279)_CYS|[B]_143(143)_CYS|[A]_143(168)_CYS|[B]_497(497)_CYS|
O15382	2HHF	A_1.00 B_1.00	A28_A392 B28_B392	28_392	[B]_815(342)_[B]_818(345)|[A]_315(342)_[A]_318(345)|	
P05091	2ONM	D_1.00 E_1.00 F_1.00 G_1.00 A_1.00 B_1.00 C_1.00 L_1.00 H_1.00 I_1.00 J_1.00 K_1.00	A18_A517 B18_B517 C18_C517 D18_D517 E18_E517 F18_F517 G18_G517 H18_H517 I18_I517 J18_J517 K18_K517 L18_L517	18_517	[C, D, I, J]_301(318)_[C, D, I, J]_303(320)|[I]_302(319)_[I]_303(320)|[E, G, H]_301(318)_[E, G, H]_303(320)|[A, B]_302(319)_[A, B]_303(320)|[B]_301(318)_[B]_303(320)|[C, E, K, L]_302(319)_[C, E, K, L]_303(320)|[A]_301(318)_[A]_303(320)|[F, G, J]_302(319)_[F, G, J]_303(320)|[L]_301(318)_[L]_303(320)|[D, H]_302(319)_[D, H]_303(320)|[C, I, J]_301(318)_[C, I, J]_302(319)|[K]_301(318)_[K]_303(320)|[A, B, D, E, G, H]_301(318)_[A, B, D, E, G, H]_302(319)|[F]_301(318)_[F]_303(320)|[K, L]_301(318)_[K, L]_302(319)|	pka or asa file is not successfully generated
P10109	3N9Z	D_0.99 C_0.99	C62_C184 D62_D184	61_184	[C, D]_46(106)_[C, D]_52(112)|[C, D]_55(115)_[C, D]_92(152)|[C, D]_46(106)_[C, D]_92(152)|[C, D]_52(112)_[C, D]_55(115)|	
P67775	2NPP	F_1.00 C_1.00	C1_C309 F1_F309	1_309	[C, F]_266(266)_[C, F]_269(269)|	[C, F]_50(50)_CYS|
P08559	1NI4	A_1.00 C_1.00	A30_A390 C30_C390	30_390	[C]_161(190)_[C]_189(218)|[C]_65(94)_[C]_72(101)|[A]_161(190)_[A]_189(218)|[A]_65(94)_[A]_72(101)|[C]_152(181)_[C]_161(190)|	[A]_152(181)_CYS|
P07339	1LYB	A_0.24 C_0.24	A65_A161 C65_C161	19_412	[C]_46(110)_[C]_53(117)|[A, C]_27(91)_[A, C]_96(160)|[A]_46(110)_[A]_53(117)|	
P30044	2VL2	A_0.68 B_0.68 C_0.68	A2_A162 B2_B162 C2_C162	53_214	[C]_47(48)_[C]_151(152)|	[A, B]_47(48)_CYS|

P07858	1GMY	A_0.81 B_0.81 C_0.81	A79_A339 B79_B339 C79_C339	18_339	[C]_63(142)_[C]_67(146)|[A]_14(93)_[A]_43(122)|[A]_63(142)_[A]_67(146)|[C]_108(187)_[C]_119(198)|[B]_14(93)_[B]_43(122)|[A, B]_108(187)_[A, B]_119(198)|[B]_63(142)_[B]_67(146)|[B, C]_100(179)_[B, C]_132(211)|[A, C]_26(105)_[A, C]_71(150)|[B]_62(141)_[B]_128(207)|[B]_26(105)_[B]_71(150)|[A]_100(179)_[A]_132(211)|[C]_62(141)_[C]_128(207)|[C]_14(93)_[C]_43(122)|[A]_62(141)_[A]_128(207)|	[A, B, C]_240(319)_CYS|
P62072	2BSK	D_1.00 F_1.00 B_1.00	B1_B90 D1_D90 F1_F90	1_90	[D]_33(33)_[D]_50(50)|[F]_29(29)_[F]_54(54)|[B, F]_33(33)_[B, F]_50(50)|[B, D]_29(29)_[B, D]_54(54)|[B, D, F]_33(33)_[B, D, F]_54(54)|[B, D, F]_29(29)_[B, D, F]_33(33)|[F]_50(50)_[F]_54(54)|	
P08559	3EXH	E_1.00 G_1.00 A_1.00 C_1.00	A30_A390 C30_C390 E30_E390 G30_G390	30_390	[E]_12(41)_[E]_232(261)|[A, E]_65(94)_[A, E]_72(101)|[A]_12(41)_[A]_232(261)|[C, G]_65(94)_[C, G]_72(101)|[A, C, E, G]_161(190)_[A, C, E, G]_189(218)|[A, C, E, G]_152(181)_[A, C, E, G]_161(190)|	
P05091	3INL	D_1.00 E_1.00 F_1.00 G_1.00 A_1.00 B_1.00 C_1.00 H_1.00	A18_A517 B18_B517 C18_C517 D18_D517 E18_E517 F18_F517 G18_G517 H18_H517	18_517	[F]_301(318)_[F]_303(320)|[C, F]_302(319)_[C, F]_303(320)|[C, H]_301(318)_[C, H]_303(320)|[A]_302(319)_[A]_303(320)|[A, E, G]_301(318)_[A, E, G]_303(320)|[G]_302(319)_[G]_303(320)|[B]_301(318)_[B]_303(320)|[B, E, H]_302(319)_[B, E, H]_303(320)|[D]_301(318)_[D]_303(320)|	pka or asa file is not successfully generated
P05981	1O5E	H_0.61	H163_H417	1_417	[H]_136(283)_[H]_201(348)|[H]_42(189)_[H]_58(205)|[H]_168(315)_[H]_182(329)|[H]_191(338)_[H]_220(367)|[H]_201(348)_[H]_210(357)|[H]_136(283)_[H]_210(357)|	
P19367	1CZA	N_1.00	N1_N917	1_917	[N]_665(665)_[N]_672(672)|[N]_237(237)_[N]_256(256)|[N]_685(685)_[N]_704(704)|[N]_217(217)_[N]_224(224)|	[N]_581(581)_CYS|[N]_133(133)_CYS|
P33316	1Q5U	Y_0.52 X_0.52 Z_0.52	X24_X164 Y24_Y164 Z24_Z164	70_252	[X, Y, Z]_55(78)_[X, Y, Z]_111(134)|	
P78310	2NPL	X_0.27	X142_X235	20_365	[X]_23(162)_[X]_73(212)|[X]_7(146)_[X]_84(223)|[X]_7(146)_[X]_23(162)|[X]_7(146)_[X]_73(212)|[X]_23(162)_[X]_84(223)|[X]_73(212)_[X]_84(223)|	

P35557	3A0I	X_0.98	X12_X466	1_465	[X]_233(234)_[X]_252(253)|[X]_230(231)_[X]_382(383)|[X]_213(214)_[X]_220(221)|[X]_457(458)_[X]_461(462)|	[X]_129(130)_CYS|[X]_364(365)_CYS|
P30405	2BIU	X_0.93	X43_X207	30_207	[X]_40(82)_[X]_161(203)|	
P53621	3mkr	0.26	P53621_3mkr_98_905_1224_SwissModel	1_1224	1185_1201|	921_CYS|
Q9BUE6	2d2a	0.95	Q9BUE6_2d2a_36_19_129_SwissModel	13_129	121_123|	
O75600	3tqx	0.99	O75600_3tqx_55_24_418_SwissModel	22_419	134_233|	106_CYS|219_CYS|
P40925	5mdh	1	P40925_5mdh_95_2_334_SwissModel	2_334	137_154|	
P55084	1ulq	0.95	P55084_1ulq_33_53_472_SwissModel	34_474	138_458|	
Q9Y606	1vs3	0.61	Q9Y606_1vs3_22_84_342_SwissModel	1_427	142_196|	260_CYS|
P50990	3p9d	0.94	P50990_3p9d_44_6_518_SwissModel	2_548	148_149|	244_CYS|325_CYS|
P51553	3blv	0.97	P51553_3blv_48_45_386_SwissModel	40_393	148_284|	235_CYS|
O75251	3i9v	0.85	O75251_3i9v_46_59_207_SwissModel	39_213	153_183|88_89|	
P80404	1ohv	0.98	P80404_1ohv_95_39_499_SwissModel	29_500	163_166|197_205|	
Q9Y234	2e5a	0.99	Q9Y234_2e5a_85_31_373_SwissModel	26_373	182_224|	337_CYS|352_CYS|
P04731	4mt2	1	P04731_4mt2_85_1_61_SwissModel	1_61	19_29|34_37|7_26|34_36|13_26|50_59|37_50|36_50|15_24|7_13|34_50|37_57|41_60|5_7|50_60|37_60|33_48|24_29|44_50|15_29|24_26|13_15|5_21|19_24|7_24|37_41|7_21|5_24|44_60|50_57|15_26|41_44|57_60|7_15|15_19|33_34|59_60|19_21|33_44|36_37|34_48|36_57|57_59|37_44|21_24|34_44|44_48|48_50|5_26|7_19|13_24|41_50|34_57|	
Q7GXZ8	1v54	-1	Q7GXZ8_1v54_74_2_220_SwissModel	-1_-1	196_200|	35_CYS|
Q9UHK6	2gci	0.93	Q9UHK6_2gci_45_2_358_SwissModel	1_382	20_160|	117_CYS|284_CYS|293_CYS|
Q8N159	3s6k	0.83	Q8N159_3s6k_32_98_526_SwissModel	19_534	200_259|	113_CYS|130_CYS|150_CYS|186_CYS|320_CYS|

Q15842	3agw	0.43	Q15842_3agw_50_180_363_SwissModel	1_424	207_268|	353_CYS|
O95782	2jkr	0.63	O95782_2jkr_91_3_621_SwissModel	1_977	215_267|215_225|	129_CYS|213_CYS|354_CYS|
Q9UPN3	3f7p	0.03	Q9UPN3_3f7p_77_69_298_SwissModel	1_7388	216_222|	
O43491	2he7	0.28	O43491_2he7_78_216_498_SwissModel	2_1005	220_232|	424_CYS|
O43837	3blv	0.97	O43837_3blv_46_43_382_SwissModel	35_385	232_233|253_254|	
Q16526	3cvv	0.84	Q16526_3cvv_50_2_494_SwissModel	1_586	24_33|58_259|	178_CYS|412_CYS|
P00505	3pd6	1	P00505_3pd6_95_30_430_SwissModel	30_430	272_274|	187_CYS|
Q14410	2d4w	0.92	Q14410_2d4w_50_10_516_SwissModel	1_553	287_415|	220_CYS|293_CYS|381_CYS|
A6NJP5	2d4w	-1	A6NJP5_2d4w_50_10_523_SwissModel	-1_-1	293_421|	220_CYS|299_CYS|387_CYS|
Q15067	1is2	0.99	Q15067_1is2_82_1_654_SwissModel	1_660	297_392|	199_CYS|199_CYS|449_CYS|449_CYS|
P14854	1v54	0.93	P14854_1v54_87_8_86_SwissModel	2_86	30_65|40_54|	
Q96I99	2fp4	0.99	Q96I99_2fp4_96_38_429_SwissModel	38_432	305_317|	162_CYS|
P30837	1ag8	0.99	P30837_1ag8_75_25_517_SwissModel	18_517	318_320|318_319|319_320|	
Q53Y00	1cit	-1	Q53Y00_1cit_98_265_353_SwissModel	-1_-1	319_322|270_284|309_322|303_322|303_309|267_284|284_287|267_270|267_287|270_287|309_319|303_319|287_327|	276_CYS|
Q6YFQ2	1v54	0.86	Q6YFQ2_1v54_56_13_88_SwissModel	1_88	32_67|42_56|	
Q53YE7	2z5y	-1	Q53YE7_2z5y_99_12_524_SwissModel	-1_-1	321_323|	
Q8IZJ4	3qxl	0.54	Q8IZJ4_3qxl_27_214_470_SwissModel	1_473	323_349|	466_CYS|
O60488	3ni2	0.88	O60488_3ni2_20_71_699_SwissModel	1_711	333_338|	221_CYS|457_CYS|494_CYS|564_CYS|640_CYS|
Q969P6	1a35	0.76	Q969P6_1a35_73_51_471_SwissModel	51_601	340_341|	268_CYS|357_CYS|370_CYS|466_CYS|

Q96PN6	2w01	0.11	Q96PN6_2w01_21_287_468_SwissModel	1_1610	346_365|	373_CYS|
Q9Y5J9	3cjh	0.68	Q9Y5J9_3cjh_36_21_76_SwissModel	2_83	36_59|40_55|40_59|36_40|	
P49821	3i9v	0.96	P49821_3i9v_44_35_459_SwissModel	21_464	379_425|382_385|382_425|379_382|	206_CYS|238_CYS|286_CYS|332_CYS|
Q9HCN8	1t9f	0.91	Q9HCN8_1t9f_49_31_206_SwissModel	29_221	38_92|100_149|	
P55786	2yd0	0.94	P55786_2yd0_31_49_915_SwissModel	1_919	385_389|887_888|	265_CYS|339_CYS|527_CYS|537_CYS|
Q9NPH0	1nd6	0.94	Q9NPH0_1nd6_22_49_419_SwissModel	33_428	388_393|208_416|	183_CYS|267_CYS|
Q9HCC0	3u9r	0.99	Q9HCC0_3u9r_65_28_563_SwissModel	23_563	391_392|	
P50991	3p9d	0.96	P50991_3p9d_58_21_539_SwissModel	2_539	410_414|	252_CYS|337_CYS|
Q9UI32	3czd	0.53	Q9UI32_3czd_77_155_464_SwissModel	15_602	433_451|	282_CYS|282_CYS|
P31040	1zoy	0.99	P31040_1zoy_96_52_664_SwissModel	44_664	438_467|189_191|	238_CYS|266_CYS|305_CYS|475_CYS|
P23378	1wyu	0.43	P23378_1wyu_32_66_489_SwissModel	36_1020	442_444|210_291|	112_CYS|250_CYS|382_CYS|
Q9NR71	2zxc	0.87	Q9NR71_2zxc_35_100_779_SwissModel	1_780	448_498|	369_CYS|
Q9Y5L4	3cjh	0.6	Q9Y5L4_3cjh_36_35_91_SwissModel	1_95	46_69|50_65|50_69|46_50|	
O60220	3cjh	0.58	O60220_3cjh_41_28_84_SwissModel	1_97	47_62|43_66|47_66|43_47|	
P45880	3emn	0.97	P45880_3emn_74_12_294_SwissModel	2_294	47_76|	103_CYS|133_CYS|138_CYS|13_CYS|210_CYS|227_CYS|
Q96RP9	2bm0	0.92	Q96RP9_2bm0_42_41_732_SwissModel	1_751	514_516|	146_CYS|153_CYS|425_CYS|706_CYS|
P05166	3n6r	0.98	P05166_3n6r_66_37_539_SwissModel	29_539	516_517|	269_CYS|291_CYS|365_CYS|448_CYS|90_CYS|
P07919	1bcc	0.84	P07919_1bcc_96_26_91_SwissModel	14_91	53_67|37_81|43_81|37_43|	
O00116	2uuu	0.96	O00116_2uuu_29_82_657_SwissModel	59_658	541_565|231_247|	214_CYS|349_CYS|413_CYS|

Q16822	3dtb	1	Q16822_3dtb_72_31_640_SwissModel	33_640	55_151|325_431|	306_CYS|559_CYS|559_CYS|
Q16134	2gmh	0.99	Q16134_2gmh_95_39_617_SwissModel	34_617	561_589|589_592|	502_CYS|586_CYS|
Q9NX47	2d8s	0.26	Q9NX47_2d8s_24_5_77_SwissModel	1_278	65_68|14_17|33_35|35_68|17_46|33_68|35_65|33_65|14_46|	
Q96F15	2xto	0.67	Q96F15_2xto_50_26_232_SwissModel	1_307	66_103|	171_CYS|
P30038	3v9j	1.01	P30038_3v9j_92_22_563_SwissModel	25_563	66_95|	421_CYS|
Q13472	2gai	0.66	Q13472_2gai_21_37_695_SwissModel	1_1001	661_685|679_685|661_679|38_190|	149_CYS|296_CYS|303_CYS|
O15235	1i94	0.99	O15235_1i94_47_31_138_SwissModel	30_138	81_93|	64_CYS|80_CYS|
P10606	1v54	1	P10606_1v54_85_32_129_SwissModel	32_129	93_116|113_116|91_113|91_116|93_113|91_93|	
P21912	1zoy	0.95	P21912_1zoy_97_37_275_SwissModel	29_280	93_98|101_113|189_191|93_113|98_101|93_101|189_253|	186_CYS|196_CYS|243_CYS|249_CYS|
Q9H1K1	1wfz	0.89	Q9H1K1_1wfz_98_44_162_SwissModel	35_167	95_138|69_95|69_138|	
P37059	1yde	0.49	P37059_1yde_29_80_271_SwissModel	1_387	99_128|	
P09211	5GSS	A_1.00 B_1.00	A1_A209 B1_B209	2_210	A_101(101)_B_101(101)|	
P24752	2IBW	D_1.00 A_1.00 B_1.00 C_1.00	A34_A427 B34_B427 C34_C427 D34_D427	34_427	A_142(142)_B_142(142)||C_142(142)_D_142(142)|	
P20700	3TYY	A_0.13 B_0.13	A311_A388 B311_B388	2_586	A_317(317)_B_317(317)|	
Q9NTZ6	2EK6	D_0.09 A_0.09 B_0.09 C_0.09	A848_A929 B848_B929 C848_C929 D848_D929	1_932	A_45(887)_B_45(887)||D_45(887)_C_45(887)|	
Q99497	1PE0	A_1.00 B_1.00	A1_A189 B1_B189	1_189	A_53(53)_B_253(53)|	
O95292	3IKK	A_0.51 B_0.51	A1_A125 B1_B125	2_243	A_53(53)_B_53(53)|	
Q99497	3SF8	A_1.00 B_1.00	A1_A189 B1_B189	1_189	A_53(53)_B_53(53)|	
P09211	1EOH	D_1.00 E_1.00 F_1.00 G_1.00 A_1.00 B_1.00	A2_A210 B2_B210 C2_C210 D2_D210 E2_E210 F2_F210 G2_G210 H2_H210	2_210	B_101(102)_A_101(102)||F_101(102)_E_101(102)|	
P63241	3CPF	A_0.89 B_0.89	A15_A151 B15_B151	2_154	B_129(129)_A_129(129)||[A, B]_22(22)_[A, B]_73(73)|	

P13693	3EBM	D_1.00 A_1.00 B_1.00 C_1.00	A1_A172 B1_B172 C1_C172 D1_D172	1_172	B_172(172)_C_172(172)|	[A, D]_172(172)_CYS|
O95822	2YGW	A_0.91 B_0.91	A40_A490 B40_B490	1_493	B_243(243)_A_243(243)|	[A, B]_206(206)_CYS|[A]_68(68)_CYS|[A]_448(448)_CYS|
P09211	1GSS	A_1.00 B_1.00	A1_A209 B1_B209	2_210	B_99(99)_A_99(99)|	
Q13162	3TJB	D_1.00 E_1.00 A_1.00 B_1.00 C_1.00	A38_A271 B38_B271 C38_C271 D38_D271 E38_E271	38_271	C_124(124)_D_245(245)||B_124(124)_A_245(245)|	[A, D, E]_124(124)_CYS|
P24752	2IBY	D_1.00 A_1.00 B_1.00 C_1.00	A34_A427 B34_B427 C34_C427 D34_D427	34_427	C_142(142)_D_142(142)||A_142(142)_B_142(142)|	
Q99497	3BWE	D_1.00 E_1.00 F_1.00 G_1.00 A_1.00 B_1.00 C_1.00	A1_A189 B1_B189 C1_C189 D1_D189 E1_E189 F1_F189 G1_G189	1_189	C_453(53)_D_653(53)||E_853(53)_F_1053(53)||A_53(53)_B_253(53)|	[A]_106(106)_CYS|[E]_906(106)_CYS|[C]_506(106)_CYS|
Q9NTZ6	2EK1	D_0.09 E_0.09 F_0.09 G_0.09 A_0.09 B_0.09 C_0.09 H_0.09	A848_A929 B848_B929 C848_C929 D848_D929 E848_E929 F848_F929 G848_G929 H848_H929	1_932	G_907(887)_H_907(887)||A_907(887)_B_907(887)||D_907(887)_C_907(887)||E_907(887)_F_907(887)|	
P16219	2VIG	D_0.99 E_0.99 F_0.99 G_0.99 A_0.99 B_0.99 C_0.99 H_0.99	A30_A412 B30_B412 C30_C412 D30_D412 E30_E412 F30_F412 G30_G412 H30_H412	25_412		[A, B, C, D, E, F, G, H]_109(109)_CYS|
P78540	1PQ3	D_0.92 E_0.92 F_0.92 A_0.92 B_0.92 C_0.92	A24_A329 B24_B329 C24_C329 D24_D329 E24_E329 F24_F329	23_354		[A, B, C, D, E, F]_134(134)_CYS|[A, B, C, D, E, F]_187(187)_CYS|[C, F]_63(63)_CYS|
P00367	1L1F	D_1.00 E_1.00 F_1.00 A_1.00 B_1.00 C_1.00	A54_A558 B54_B558 C54_C558 D54_D558 E54_E558 F54_F558	54_558		[A, B, C, D, E, F]_59(112)_CYS|
P30084	2HW5	D_1.00 E_1.00 F_1.00 A_1.00 B_1.00 C_1.00	A28_A290 B28_B290 C28_C290 D28_D290 E28_E290 F28_F290	28_290		[A, B, C, D, E, F]_62(62)_CYS|
Q8IVH4	2WWW	D_0.98 A_0.98 B_0.98 C_0.98	A72_A418 B72_B418 C72_C418 D72_D418	66_418		[A, B, C, D]_100(100)_CYS|
P17174	3II0	D_0.97 A_0.97 B_0.97 C_0.97	A14_A412 B14_B412 C14_C412 D14_D412	2_413		[A, B, C, D]_83(83)_CYS|
Q13162	3TKQ	D_1.00 E_1.00 A_1.00 B_1.00 C_1.00	A38_A271 B38_B271 C38_C271 D38_D271 E38_E271	38_271		[A, B, C, E]_87(124)_CYS|

Q96EY8	2IDX	A_0.89 B_0.89 C_0.89	A56_A250 B56_B250 C56_C250	33_250	[A, B, C]_119(119)_CYS|[B, C]_132(132)_CYS|
P30044	1OC3	A_0.99 B_0.99 C_0.99	A54_A214 B54_B214 C54_C214	53_214	[A, B, C]_47(100)_CYS|[B, C]_151(204)_CYS|
P00480	1FVO	A_1.00 B_1.00	A34_A354 B34_B354	33_354	[A, B]_109(109)_CYS|
Q9Y3E5	1Q7S	A_1.00 B_1.00	A63_A179 B63_B179	63_179	[A, B]_111(111)_CYS|
Q16698	1W6U	D_1.00 A_1.00 B_1.00 C_1.00	A35_A335 B35_B335 C35_C335 D35_D335	35_335	[A, B]_116(116)_CYS|
Q9NVV4	3PQ1	A_0.09 B_0.09	A44_A134 A172_A452 A490_A538 B44_B134 B172_B452 B490_B538	38_582	[A, B]_181(181)_CYS|[A]_119(119)_CYS|[A, B]_299(299)_CYS|[B]_82(82)_CYS|
Q16654	2ZDY	A_0.95 B_0.95	A20_A411 B20_B411	1_411	[A, B]_210(210)_CYS|[A, B]_49(49)_CYS|
P16435	3QFR	A_0.91 B_0.91	A64_A677 B64_B677	2_677	[A, B]_231(228)_CYS|
O75874	3MAP	A_1.00 B_1.00	A1_A414 B1_B414	1_414	[A, B]_269(269)_CYS|
P32019	3N9V	A_0.31 B_0.31	A342_A647 B342_B647	1_993	[A, B]_282(362)_CYS|[A, B]_425(505)_CYS|
O15294	1W3B	A_0.37 B_0.37	A16_A400 B16_B400	2_1046	[A, B]_313(313)_CYS|[A, B]_148(148)_CYS|[A]_169(169)_CYS|[A, B]_287(287)_CYS|[A, B]_58(58)_CYS|[A]_257(257)_CYS|[A]_189(189)_CYS|
P48728	1WSR	A_1.00 B_1.00	A29_A403 B29_B403	29_403	[A, B]_339(367)_CYS|
Q9BV79	2VCY	A_1.00 B_1.00	A31_A373 B31_B373	54_373	[A, B]_345(345)_CYS|
Q8WWV3	2VN8	A_0.99 B_0.99	A45_A396 B45_B396	41_396	[A, B]_349(349)_CYS|
O00763	3GID	A_0.21 B_0.21	A238_A760 B238_B760	1_2458	[A, B]_365(365)_CYS|[B]_433(433)_CYS|
Q9P286	2F57	A_0.41 B_0.41	A425_A719 B425_B719	1_719	[A, B]_464(464)_CYS|
Q00059	3TQ6	A_1.00 B_1.00	A43_A246 B43_B246	43_246	[A, B]_49(49)_CYS|
Q15149	3PDY	A_0.04 B_0.04	A653_A858 B653_B858	1_4684	[A, B]_620(730)_CYS|[A, B]_740(850)_CYS|[A, B]_739(849)_CYS|
P04150	3K23	A_0.33 B_0.33 C_0.33	A521_A777 B521_B777 C521_C777	1_777	[A, B]_622(622)_CYS|[C]_643(643)_CYS|
Q16836	1F12	A_0.98 B_0.98	A7_A308 B7_B308	13_314	[A, B]_80(86)_CYS|
P04040	1F4J	D_1.00 A_1.00 B_1.00 C_1.00	A1_A527 B1_B527 C1_C527 D1_D527	2_527	[A, C]_393(393)_CYS|

Q05193	2X2F	D_0.03 A_0.03	A6_A320 A726_A750 D6_D320 D726_D750	1_864	[A, D]_86(86)_CYS|
Q16181	3TW4	A_0.62 B_0.62	A48_A317 B48_B317	1_437	[A]_107(126)_CYS|
P00480	1EP9	A_1.00	A34_A354	33_354	[A]_109(109)_CYS|
P34896	1BJ4	A_0.97	A11_A480	1_483	[A]_110(110)_CYS|[A]_68(68)_CYS|
P62995	2CQC	A_0.28	A110_A191	1_288	[A]_118(118)_CYS|[A]_119(119)_CYS|
P41250	2ZT7	A_0.93	A55_A739	1_739	[A]_126(180)_CYS|[A]_157(211)_CYS|[A]_101(155)_CYS|[A]_417(471)_CYS|
P21796	2K4T	A_1.00	A1_A283	2_283	[A]_127(127)_CYS|[A]_232(232)_CYS|
Q8N4E7	1R03	A_0.94	A61_A242	50_242	[A]_130(190)_CYS|[A]_102(162)_CYS|
P14550	2ALR	A_1.00	A1_A324	2_325	[A]_133(133)_CYS|[A]_4(4)_CYS|
P16435	3FJO	A_0.66	A232_A677	2_677	[A]_135(155)_CYS|
Q16836	1M76	A_1.00 B_1.00	A13_A314 B13_B314	13_314	[A]_137(149)_CYS|
P43155	1S5O	A_0.95	A14_A605	1_626	[A]_148(148)_CYS|[A]_574(574)_CYS|
Q14318	2AWG	A_0.28	A33_A148	1_412	[A]_150(93)_CYS|
P30044	1URM	A_0.99	A54_A214	53_214	[A]_151(204)_CYS|
P20700	2KPW	A_0.19	A439_A549	2_586	[A]_16(443)_CYS|
Q16595	3T3X	A_0.76 B_0.76	A82_A210 B82_B210	42_210	[A]_165(165)_CYS|
P16435	1B1C	A_0.27	A61_A241	2_677	[A]_168(228)_CYS|
O75880	2GT5	A_0.56	A132_A301	1_301	[A]_169(169)_CYS|
P07237	2K18	A_0.45	A135_A357	18_508	[A]_183(312)_CYS|
Q92947	2R0N	A_1.00	A45_A438	45_438	[A]_188(232)_CYS|
P53396	3MWD	A_0.39	A1_A425	1_1101	[A]_20(20)_CYS|
Q9NRK6	4AA3	A_0.97	A1_A5 A126_A738	106_738	[A]_215(215)_CYS|[A]_582(582)_CYS|
O60333	2EH0	A_0.06	A531_A647	1_1816	[A]_23(546)_CYS|[A]_77(600)_CYS|
P21796	2JK4	A_1.00	A2_A283	2_283	[A]_235(232)_CYS|[A]_130(127)_CYS|
O15527	2XHI	A_1.00	A1_A345	1_345	[A]_241(241)_CYS|

Q9NZ01	2DZJ	A_0.26	A1_A81	1_308	[A]_25(18)_CYS|
O15527	1KO9	A_1.00	A1_A345	1_345	[A]_253(253)_CYS|[A]_241(241)_CYS|
Q6NVY1	3BPT	A_1.00	A32_A386	33_386	[A]_271(271)_CYS|[A]_45(45)_CYS|
Q14790	2K7Z	A_0.55	A217_A479	1_479	[A]_287(287)_CYS|[A]_345(345)_CYS|[A]_426(426)_CYS|
O15527	3IH7	A_0.91	A12_A325	1_345	[A]_292(292)_CYS|[A]_241(241)_CYS|
Q00610	2XZG	A_0.22	A1_A364	2_1675	[A]_328(328)_CYS|
Q9BW91	1QVJ	A_0.96	A59_A350	47_350	[A]_347(347)_CYS|[A]_207(207)_CYS|
O00763	3JRX	A_0.23	A217_A775	1_2458	[A]_365(365)_CYS|
Q8TD30	3IHJ	A_0.91	A49_A523	1_523	[A]_376(376)_CYS|[A]_281(281)_CYS|[A]_347(347)_CYS|
Q15119	2BU2	A_0.96	A16_A407	1_407	[A]_384(392)_CYS|[A]_37(45)_CYS|
Q05193	3SNH	A_0.86	A6_A746	1_864	[A]_427(427)_CYS|[A]_708(708)_CYS|
P20700	3UMN	A_0.21 B_0.21 C_0.21	A428_A550 B428_B550 C428_C550	2_586	[A]_443(443)_CYS|
Q9Y237	1EQ3	A_0.73	A36_A131	1_131	[A]_45(45)_CYS|
P30044	3MNG	A_0.99	A54_A214	53_214	[A]_47(100)_CYS|[A]_151(204)_CYS|
Q9NTZ6	1WEL	A_0.12	A412_A522	1_932	[A]_489(489)_CYS|
Q13636	2FG5	A_0.89	A2_A174	1_194	[A]_49(49)_CYS|[A]_23(23)_CYS|
Q02318	1MFX	A_1.00	A34_A531	34_531	[A]_498(531)_CYS|
Q99714	1F67	A_1.00	A1_A261	2_261	[A]_5(5)_CYS|[A]_214(214)_CYS|
P32019	3MTC	A_0.31	A339_A643	1_993	[A]_517(597)_CYS|[A]_282(362)_CYS|[A]_425(505)_CYS|
Q99497	3B36	A_1.00	A1_A189	1_189	[A]_53(53)_CYS|
P10515	3B8K	A_0.42	A376_A614	87_647	[A]_532(585)_CYS|
P46940	2RR8	A_-1.00	inconsis. annot. btw UniProt and PDB	2_1657	[A]_57(57)_CYS|
P43155	1NM8	A_0.95	A35_A626	1_626	[A]_574(595)_CYS|[A]_148(169)_CYS|
P10515	2DNE	A_0.12	A59_A153	87_647	[A]_59(110)_CYS|

P35670	2KOY	A_0.11	A1036_A1196	1_1465	[A]_60(1091)_CYS|[A]_134(1165)_CYS|[A]_48(1079)_CYS|[A]_73(1104)_CYS|
P04150	1NHZ	A_0.36	A500_A777	1_777	[A]_622(622)_CYS|[A]_736(736)_CYS|
P50440	8JDW	A_1.00	A38_A423	38_423	[A]_64(64)_CYS|
Q9NP58	3NH9	A_0.34	A558_A842	1_842	[A]_646(646)_CYS|[A]_801(801)_CYS|
P05165	2JKU	A_0.10	A633_A703	53_728	[A]_663(663)_CYS|
Q14318	2D9F	A_-1.00	inconsis. annot. btw UniProt and PDB	1_412	[A]_67(67)_CYS|
O00763	2KCC	A_0.03	A891_A965	1_2458	[A]_67(956)_CYS|
Q00059	3TMM	A_1.00	A43_A246	43_246	[A]_7(49)_CYS|
P22307	1QND	A_0.22	A425_A547	1_547	[A]_71(495)_CYS|
Q4LE76	2GAQ	A_-1.00	inconsis. annot. btw UniProt and PDB	1_2549	[A]_71(71)_CYS|
Q15118	2Q8G	A_1.00	A30_A436	29_436	[A]_71(71)_CYS|[A]_421(421)_CYS|
P34897	3OU5	A_1.00	A17_A504	30_504	[A]_75(91)_CYS|[A]_52(68)_CYS|
P49916	3QVG	A_0.08 C_0.08	A924_A1008 C924_C1008	1_1009	[A]_842(929)_CYS|
O00763	2DN8	A_-1.00	inconsis. annot. btw UniProt and PDB	1_2458	[A]_9(9)_CYS|[A]_79(79)_CYS|
P49916	1IMO	A_0.09	A835_A922	1_1009	[A]_922(922)_CYS|
Q14318	3EY6	A_0.29	A92_A210	1_412	[A]_93(150)_CYS|[A]_138(195)_CYS|
P53370	3FXT	D_0.28 E_0.28 F_0.28 G_0.28 A_0.28 B_0.28 C_0.28 H_0.28	A45_A134 B45_B134 C45_C134 D45_D134 E45_E134 F45_F134 G45_G134 H45_H134	1_316	[B, C, D, E, F, G, H]_114(114)_CYS|
P30044	1H4O	D_0.99 E_0.99 F_0.99 G_0.99 A_0.99 B_0.99 C_0.99 H_0.99	A54_A214 B54_B214 C54_C214 D54_D214 E54_E214 F54_F214 G54_G214 H54_H214	53_214	[B, D, E, F]_47(100)_CYS|[D, H]_151(204)_CYS|
P04150	1P93	D_0.36 A_0.36 B_0.36 C_0.36	A500_A777 B500_B777 C500_C777 D500_D777	1_777	[B, D]_622(622)_CYS|
Q15019	2QNR	A_0.83 B_0.83	A22_A320 B22_B320	1_361	[B]_114(114)_CYS|[B]_111(111)_CYS|
P13010	1JEY	B_0.77	B0_B564	2_732	[B]_157(156)_CYS|[B]_249(248)_CYS|[B]_296(295)_CYS|
O75874	40971	A_1.00 B_1.00	A1_A414 B1_B414	1_414	[B]_269(269)_CYS|

P20700	3JT0	A_0.23 B_0.23	A426_A558 B426_B558	2_586	[B]_29(443)_CYS|
P13010	1JEQ	B_0.77	B0_B564	2_732	[B]_339(338)_CYS|[B]_157(156)_CYS|
Q16654	2ZDX	A_0.95 B_0.95	A20_A411 B20_B411	1_411	[B]_396(396)_CYS|[A, B]_210(210)_CYS|[A, B]_49(49)_CYS|
Q99714	2O23	A_-1.00 B_-1.00	inconsis. annot. btw UniProt and PDB	2_261	[B]_5(5)_CYS|
P38117	2A1U	B_1.00	B1_B255	2_255	[B]_71(71)_CYS|[B]_42(42)_CYS|[B]_66(66)_CYS|
Q96PE7	3RMU	D_0.94 A_0.94 B_0.94 C_0.94	A45_A176 B45_B176 C45_C176 D45_D176	37_176	[C, D]_166(166)_CYS|
P54868	2WYA	D_0.97 A_0.97 B_0.97 C_0.97	A51_A508 B51_B508 C51_C508 D51_D508	38_508	[C, D]_454(454)_CYS|
Q15831	2WTK	F_0.70 C_0.70	C43_C347 F43_F347	1_433	[C, F]_73(73)_CYS|[C]_278(278)_CYS|[C, F]_158(158)_CYS|[C, F]_151(151)_CYS|
P67775	3C5W	C_1.00	C1_C309	1_309	[C]_50(50)_CYS|
P04150	3H52	D_0.32 A_0.32 B_0.32 C_0.32	A528_A777 B528_B777 C528_C777 D528_D777	1_777	[C]_736(736)_CYS|
Q68CS0	2BYL	A_1.00 B_1.00 C_1.00	A1_A439 B1_B439 C1_C439	26_439	[C]_93(93)_CYS|
P11177	3EXH	D_1.00 F_1.00 B_1.00 H_1.00	B31_B359 D31_D359 F31_F359 H31_H359	31_359	[D, H]_131(161)_CYS|
Q13162	3TKS	D_1.00 E_1.00 A_1.00 B_1.00 C_1.00	A38_A271 B38_B271 C38_C271 D38_D271 E38_E271	38_271	[E]_87(124)_CYS|
P67775	2IAE	F_1.00 C_1.00	C1_C309 F1_F309	1_309	[F]_251(251)_CYS|[F]_20(20)_CYS|[C]_266(266)_CYS|
P04406	2FEH	Q_1.00 P_1.00 R_1.00 O_1.00	O0_O334 P0_P334 Q0_Q334 R0_R334	2_335	[P, Q, R]_152(151)_CYS|
P38117	2A1T	S_1.00	S1_S255	2_255	[S]_71(71)_CYS|[S]_42(42)_CYS|[S]_66(66)_CYS|
Q15388	2v1t	0.47	Q15388_2v1t_98_59_126_SwissModel	1_145	100_CYS|
Q9NYK5	3hvz	0.17	Q9NYK5_3hvz_24_73_130_SwissModel	1_338	100_CYS|123_CYS|
Q49AM1	3n7q	0.84	Q49AM1_3n7q_29_74_367_SwissModel	36_385	101_CYS|259_CYS|276_CYS|288_CYS|

Q9H5Q4	3r0q	0.17	Q9H5Q4_3r0q_29_93_157_SwissModel	20_396	102_CYS|
Q96KJ9	1v54	0.85	Q96KJ9_1v54_49_26_171_SwissModel	1_171	108_CYS|
Q9BSK2	2lck	0.66	Q9BSK2_2lck_22_6_218_SwissModel	1_321	111_CYS|173_CYS|203_CYS|20_CYS|30_CYS|
Q15181	2ik0	0.99	Q15181_2ik0_50_3_288_SwissModel	1_289	114_CYS|254_CYS|274_CYS|
Q53FZ2	3b7w	0.96	Q53FZ2_3b7w_58_51_584_SwissModel	28_586	115_CYS|
P24298	3ihj	0.96	P24298_3ihj_65_21_495_SwissModel	2_496	115_CYS|254_CYS|320_CYS|349_CYS|450_CYS|
Q9Y277	2jk4	0.99	Q9Y277_2jk4_68_5_283_SwissModel	2_283	122_CYS|229_CYS|36_CYS|65_CYS|8_CYS|
P55916	2lck	0.96	P55916_2lck_73_13_311_SwissModel	1_312	124_CYS|133_CYS|194_CYS|230_CYS|259_CYS|25_CYS|75_CYS|
Q9Y619	2lck	0.63	Q9Y619_2lck_21_108_298_SwissModel	1_301	125_CYS|132_CYS|222_CYS|
P53007	2lck	0.96	P53007_2lck_24_26_310_SwissModel	14_311	127_CYS|287_CYS|41_CYS|70_CYS|
P05141	2c3e	0.99	P05141_2c3e_89_2_294_SwissModel	2_298	129_CYS|
P12236	2c3e	0.99	P12236_2c3e_89_2_294_SwissModel	2_298	129_CYS|
O95398	1o7f	0.28	O95398_1o7f_60_89_351_SwissModel	1_923	129_CYS|130_CYS|166_CYS|
Q3ZCQ8	3qle	0.6	Q3ZCQ8_3qle_37_132_317_SwissModel	45_353	133_CYS|236_CYS|
Q9UFN0	1vqs	0.43	Q9UFN0_1vqs_30_30_137_SwissModel	1_247	135_CYS|
P49406	3bbo	0.38	P49406_3bbo_34_90_200_SwissModel	1_292	137_CYS|163_CYS|
Q9BSE5	3nio	0.87	Q9BSE5_3nio_61_42_349_SwissModel	1_352	138_CYS|153_CYS|216_CYS|57_CYS|
O14521	1zoy	0.98	O14521_1zoy_96_59_159_SwissModel	57_159	140_CYS|150_CYS|88_CYS|
P48047	2bo5	0.63	P48047_2bo5_89_24_143_SwissModel	24_213	141_CYS|
Q9H0C2	1okc	0.91	Q9H0C2_1okc_73_17_304_SwissModel	1_315	141_CYS|235_CYS|
Q13637	2y8e	0.75	Q13637_2y8e_36_24_192_SwissModel	2_225	145_CYS|162_CYS|

Q02252	4e4g	0.96	Q02252_4e4g_44_38_519_SwissModel	34_535	149_CYS|368_CYS|
Q96A46	2lck	0.52	Q96A46_2lck_23_70_258_SwissModel	1_364	151_CYS|178_CYS|211_CYS|88_CYS|95_CYS|
O95239	2xt3	0.27	O95239_2xt3_45_6_334_SwissModel	1_1232	153_CYS|28_CYS|
O43772	2lck	0.65	O43772_2lck_25_6_202_SwissModel	1_301	155_CYS|23_CYS|58_CYS|89_CYS|
Q9H1D0	2rfa	0.31	Q9H1D0_2rfa_89_44_265_SwissModel	1_725	157_CYS|213_CYS|70_CYS|
Q8WVM0	3tqs	0.86	Q8WVM0_3tqs_27_35_307_SwissModel	28_346	160_CYS|260_CYS|
Q9UJZ1	3bk6	0.41	Q9UJZ1_3bk6_36_69_215_SwissModel	1_356	167_CYS|
O95571	2gcu	0.92	O95571_2gcu_60_23_245_SwissModel	13_254	170_CYS|219_CYS|34_CYS|
O75879	3h0l	0.82	O75879_3h0l_41_61_492_SwissModel	31_557	170_CYS|228_CYS|322_CYS|
Q05932	1o5z	0.73	Q05932_1o5z_24_41_440_SwissModel	43_587	171_CYS|
Q9UBX3	2lck	0.98	Q9UBX3_2lck_35_8_287_SwissModel	1_287	181_CYS|18_CYS|212_CYS|217_CYS|22_CYS|23_CYS|240_CYS|69_CYS|
P51451	1blj	0.22	P51451_1blj_83_113_224_SwissModel	2_505	181_CYS|213_CYS|
P15880	2zkq	0.51	P15880_2zkq_96_101_249_SwissModel	2_293	182_CYS|229_CYS|
Q9BWU0	2jpe	0.16	Q9BWU0_2jpe_31_149_275_SwissModel	1_796	189_CYS|197_CYS|224_CYS|
O95470	3mau	0.79	O95470_3mau_44_104_552_SwissModel	1_568	193_CYS|
Q49AN0	3cvv	0.86	Q49AN0_3cvv_51_6_513_SwissModel	1_593	197_CYS|431_CYS|433_CYS|49_CYS|
Q9BZJ4	2lck	0.53	Q9BZJ4_2lck_22_162_353_SwissModel	1_359	202_CYS|334_CYS|
Q6IB77	3ec4	0.26	Q6IB77_3ec4_20_206_284_SwissModel	1_296	207_CYS|208_CYS|
Q96NN9	3he3	0.06	Q96NN9_3he3_38_193_231_SwissModel	1_605	209_CYS|225_CYS|
Q9UMS0	1veh	0.3	Q9UMS0_1veh_95_169_241_SwissModel	10_254	210_CYS|213_CYS|

Q5HYM3	1v88	-1	Q5HYM3_1v88_97_147_266_SwissModel	-1_-1	210_CYS|256_CYS|
Q9Y6N1	1so9	0.45	Q9Y6N1_1so9_44_144_268_SwissModel	1_276	217_CYS|219_CYS|
Q92581	2l0e	0.04	Q92581_2l0e_55_218_246_SwissModel	1_669	220_CYS|
Q9UJ68	1fva	0.95	Q9UJ68_1fva_91_30_230_SwissModel	24_235	220_CYS|
Q9BXI2	2lck	0.29	Q9BXI2_2lck_27_211_299_SwissModel	1_301	222_CYS|232_CYS|
P35218	1dmx	0.88	P35218_1dmx_80_61_296_SwissModel	39_305	224_CYS|
Q9H936	2c3e	0.3	Q9H936_2c3e_26_220_317_SwissModel	1_323	233_CYS|246_CYS|290_CYS|
Q9Y276	3r8c	0.06	Q9Y276_3r8c_39_226_253_SwissModel	1_419	234_CYS|252_CYS|
Q14032	3hlk	0.98	Q14032_3hlk_43_1_410_SwissModel	1_418	235_CYS|372_CYS|373_CYS|
Q16762	1rhs	0.99	Q16762_1rhs_89_2_293_SwissModel	2_297	248_CYS|
O15229	3rqs	0.07	O15229_3rqs_37_6_42_SwissModel	1_486	25_CYS|
Q9H1K4	2c3e	0.27	Q9H1K4_2c3e_24_4_89_SwissModel	1_315	25_CYS|
O43822	2omz	0.22	O43822_2omz_34_17_74_SwissModel	1_256	25_CYS|36_CYS|
P49662	2fqq	0.39	P49662_2fqq_63_124_270_SwissModel	1_377	258_CYS|
P40763	1bg1	0.75	P40763_1bg1_96_136_715_SwissModel	1_770	259_CYS|367_CYS|418_CYS|426_CYS|468_CYS|542_CYS|712_CYS|
Q86WU2	3pm9	0.95	Q86WU2_3pm9_27_27_507_SwissModel	1_507	28_CYS|384_CYS|63_CYS|
Q99797	2o36	0.92	Q99797_2o36_23_72_697_SwissModel	36_713	281_CYS|568_CYS|687_CYS|
P78549	1orn	0.59	P78549_1orn_29_126_310_SwissModel	1_312	290_CYS|297_CYS|300_CYS|306_CYS|
Q5JWV7	2k21	-1	Q5JWV7_2k21_23_1_109_SwissModel	-1_-1	3_CYS|72_CYS|98_CYS|
Q9Y371	1x43	0.2	Q9Y371_1x43_89_293_365_SwissModel	1_365	304_CYS|

O43236	3ftq	0.56	O43236_3ftq_62_141_410_SwissModel	1_478	323_CYS|
Q9BXM7	2x4f	0.5	Q9BXM7_2x4f_20_262_514_SwissModel	78_581	323_CYS|
P55060	1wa5	0.99	P55060_1wa5_33_7_963_SwissModel	1_971	325_CYS|61_CYS|630_CYS|85_CYS|939_CYS|
Q9H845	2uxw	0.94	Q9H845_2uxw_45_35_618_SwissModel	1_621	327_CYS|507_CYS|613_CYS|
Q9P258	4d9s	0.33	Q9P258_4d9s_28_330_504_SwissModel	1_522	337_CYS|437_CYS|
P28330	2pg0	0.93	P28330_2pg0_48_53_426_SwissModel	31_430	342_CYS|366_CYS|
Q8N490	1qh5	0.69	Q8N490_1qh5_38_119_384_SwissModel	1_385	347_CYS|
O60313	2x2e	0.37	O60313_2x2e_26_262_588_SwissModel	88_960	375_CYS|
Q9Y697	3lvm	0.85	Q9Y697_3lvm_58_57_446_SwissModel	1_457	381_CYS|
Q969S9	2bm0	0.91	Q969S9_2bm0_38_65_773_SwissModel	1_779	406_CYS|
Q96TC7	1fch	0.11	Q96TC7_1fch_28_400_451_SwissModel	1_470	427_CYS|
Q08AH1	3b7w	0.98	Q08AH1_3b7w_57_39_574_SwissModel	32_577	431_CYS|
Q96RR1	3zq6	0.08	Q96RR1_3zq6_30_396_447_SwissModel	32_684	433_CYS|
Q14008	2qk2	0.12	Q14008_2qk2_54_267_502_SwissModel	1_2032	441_CYS|
Q15139	2d9z	0.14	Q15139_2d9z_59_419_545_SwissModel	1_912	450_CYS|
Q53XN1	1v54	-1	Q53XN1_1v54_41_26_68_SwissModel	-1_-1	47_CYS|
P28288	1g41	0.04	P28288_1g41_37_465_491_SwissModel	1_659	472_CYS|477_CYS|
P24311	1v54	0.85	P24311_1v54_85_31_78_SwissModel	25_80	49_CYS|
P26439	3hsk	0.09	P26439_3hsk_42_3_35_SwissModel	2_372	5_CYS|
Q9UDR5	2axq	0.49	Q9UDR5_2axq_38_481_922_SwissModel	33_926	534_CYS|765_CYS|

P46199	3izy	0.77	P46199_3izy_87_178_713_SwissModel	30_727	544_CYS|616_CYS|
P40939	3had	0.39	P40939_3had_32_359_641_SwissModel	37_763	550_CYS|
Q9Y2Z2	2z5y	0.05	Q9Y2Z2_2z5y_44_35_68_SwissModel	26_717	58_CYS|
Q8IWA4	1t3j	0.08	Q8IWA4_1t3j_90_674_734_SwissModel	1_741	681_CYS|681_CYS|
P46776	2zkr	0.97	P46776_2zkr_83_6_147_SwissModel	2_148	70_CYS|
O95140	1t3j	0.08	O95140_1t3j_53_693_754_SwissModel	1_757	700_CYS|
Q93084	2voy	0.03	Q93084_2voy_96_749_779_SwissModel	1_1043	774_CYS|
Q9UL12	3dk9	0.04	Q9UL12_3dk9_36_66_106_SwissModel	1_918	79_CYS|
Q8N573	4acj	0.19	Q8N573_4acj_77_709_874_SwissModel	1_874	801_CYS|849_CYS|
P43304	2hqm	0.06	P43304_2hqm_45_68_107_SwissModel	43_727	84_CYS|
Q96DT0	2yv8	0.44	Q96DT0_2yv8_40_38_186_SwissModel	1_336	84_CYS|87_CYS|
Q9BUI6	1c4z	-1	Q9BUI6_1c4z_99_517_866_SwissModel	-1_-1	840_CYS|
P11177	3EXG	D_1.00 F_1.00 B_1.00 L_1.00 N_1.00 H_1.00 J_1.00 2_1.00 T_1.00 V_1.00 6_1.00 P_1.00 4_1.00 R_1.00 X_1.00 Z_1.00	B31_B359 D31_D359 F31_F359 H31_H359 J31_J359 L31_L359 N31_N359 P31_P359 R31_R359 T31_T359 V31_V359 X31_X359 Z31_Z359 231_2359 431_4359 631_6359	31_359	pka or asa file is not successfully generated
P51659	1S9C	D_0.40 E_0.40 F_0.40 G_0.40 A_0.40 B_0.40 C_0.40 L_0.40 H_0.40 I_0.40 J_0.40 K_0.40	A318_A615 B318_B615 C318_C615 D318_D615 E318_E615 F318_F615 G318_G615 H318_H615 I318_I615 J318_J615 K318_K615 L318_L615	2_736	pka or asa file is not successfully generated
Q05193	3ZYS	D_0.03 A_0.03	A6_A320 A726_A750 D6_D320 D726_D750	1_864	pka or asa file is not successfully generated

A0FGR8	2DMG	A_-1.00	inconsis. annot. btw UniProt and PDB	1_921
B0ZBD0	3iz6	-1	B0ZBD0_3iz6_56_3_141_SwissModel	-1_-1
O00746	1EHW	A_0.92 B_0.92	A14_A175 B14_B175	34_187
O00763	3FF6	D_0.31 A_0.31 B_0.31 C_0.31	A1693_A2450 B1693_B2450 C1693_C2450 D1693_D2450	1_2458
O14957	2fyu	0.95	O14957_2fyu_88_1_53_SwissModel	1_56
O14979	1hd0	0.18	O14979_1hd0_78_149_223_SwissModel	1_420
O15217	3IK7	D_1.00 A_1.00 B_1.00 C_1.00	A1_A222 B1_B222 C1_C222 D1_D222	1_222
O15382	2HGW	A_1.00 B_1.00	A28_A392 B28_B392	28_392
O43169	3NER	A_0.63 B_0.63	A12_A103 B12_B103	1_146
O43464	1LCY	A_0.76	A134_A458	32_458
O43521	3D7V	B_0.13	B141_B166	1_198
O43852	2vrg	0.08	O43852_2vrg_34_192_217_SwissModel	20_315
O60346	1wwl	0.02	O60346_1wwl_35_1106_1139_SwissModel	1_1717
O60610	1v9d	0.25	O60610_1v9d_96_847_1167_SwissModel	1_1272
O60825	1fbt	0.38	O60825_1fbt_80_250_439_SwissModel	1_505
O60884	2ctp	0.15	O60884_2ctp_61_10_70_SwissModel	1_412
O60936	2dbd	0.36	O60936_2dbd_26_5_84_SwissModel	1_219
O75323	1vqy	0.36	O75323_1vqy_28_182_286_SwissModel	1_286
O75414	1k44	0.73	O75414_1k44_40_12_147_SwissModel	1_186
O75489	3mcr	0.63	O75489_3mcr_28_43_185_SwissModel	37_264

O75643	2Q0Z	X_-1.00	inconsis. annot. btw UniProt and PDB	1_2136
O75874	3INM	A_1.00 B_1.00 C_1.00	A1_A414 B1_B414 C1_C414	1_414
O76031	3hws	0.06	O76031_3hws_67_167_203_SwissModel	57_633
O94826	2kc7	0.06	O94826_2kc7_33_112_147_SwissModel	2_608
O95153	2csq	0.04	O95153_2csq_73_1623_1694_SwissModel	1_1857
O95202	3skq	0.31	O95202_3skq_30_252_446_SwissModel	116_739
O95831	1M6I	A_0.88	A121_A613	55_613
P00167	2I96	A_0.80	A0_A107	2_134
P00387	1M91	A_1.00	A1_A300	2_301
P00414	1v54	0.99	P00414_1v54_87_3_261_SwissModel	1_261
P01116	3GFT	D_0.89 E_0.89 F_0.89 A_0.89 B_0.89 C_0.89	A1_A169 B1_B169 C1_C169 D1_D169 E1_E169 F1_F169	1_189
P02549	1OWA	A_0.06	A1_A156	1_2419
P04040	1QQW	D_1.00 A_1.00 B_1.00 C_1.00	A1_A527 B1_B527 C1_C527 D1_D527	2_527
P04150	3K22	A_0.33 B_0.33	A521_A777 B521_B777	1_777
P04406	1ZNQ	Q_1.00 P_1.00 R_1.00 O_1.00	O0_O334 P0_P334 Q0_Q334 R0_R334	2_335
P05165	2CQY	A_0.14	A151_A245	53_728
P06576	2ck3	0.97	P06576_2ck3_99_60_525_SwissModel	48_529
P07237	3BJ5	A_0.28	A230_A368	18_508
P09211	3KMN	A_1.00 B_1.00	A2_A210 B2_B210	2_210
P10515	1Y8N	B_0.19	B179_B286	87_647
P10809	1sx3	0.96	P10809_1sx3_50_26_551_SwissModel	27_573
P11142	3FZH	A_0.59	A4_A381	2_646
P11177	3EXI	B_1.00	B31_B359	31_359

P11182	1K8O	A_0.20	A62_A145	62_482
P13010	1RW2	A_0.20	A565_A709	2_732
P13051	1AKZ	A_0.70	A85_A304	1_313
P13073	1v54	0.98	P13073_1v54_81_26_169_SwissModel	23_169
P13693	1YZ1	D_1.00 A_1.00 B_1.00 C_1.00	A1_A172 B1_B172 C1_C172 D1_D172	1_172
P14406	1v54	0.97	P14406_1v54_60_24_81_SwissModel	24_83
P16435	3QE2	A_0.91 B_0.91	A64_A677 B64_B677	2_677
P20020	2kne	0.02	P20020_2kne_85_1100_1126_SwissModel	1_1258
P20674	2y69	0.95	P20674_2y69_96_47_150_SwissModel	42_150
P20815	2v0m	0.93	P20815_2v0m_84_30_495_SwissModel	1_502
P22307	2C0L	B_0.22	B426_B547	1_547
P23434	3klr	1	P23434_3klr_96_49_173_SwissModel	49_173
P26038	1E5W	A_0.60	A1_A345	2_577
P27144	2AR7	A_1.00 B_1.00	A1_A223 B1_B223	1_223
P30042	1oy1	0.99	P30042_1oy1_44_43_267_SwissModel	42_268
P30101	2H8L	A_0.50 B_0.50 C_0.50	A134_A376 B134_B376 C134_C376	25_505
P31942	1wez	0.22	P31942_1wez_85_194_270_SwissModel	1_346
P32969	2CQL	A_0.45	A1_A87	1_192
P35579	3zwh	0.02	P35579_3zwh_97_1893_1934_SwissModel	2_1960
P35637	2LA6	A_0.17	A282_A370	1_526
P35670	2ARF	A_0.11	A1032_A1196	1_1465
P36776	2X36	D_0.23 E_0.23 F_0.23 A_0.23 B_0.23 C_0.23	A753_A959 B753_B959 C753_C959 D753_D959 E753_E959 F753_F959	68_959
P42166	1GJJ	A_0.24	A1_A168	2_694

P43897	2CP9	A_0.18	A45_A95	46_325
P46940	3FAY	A_0.23	A962_A1345	2_1657
P49327	2PX6	A_0.12 B_0.12	A2200_A2511 B2200_B2511	1_2511
P49411	1d2e	0.97	P49411_1d2e_95_55_451_SwissModel	44_452
P49916	3PC8	D_0.08 C_0.08	C924_C1008 D924_D1008	1_1009
P50416	2LE3	A_0.05	A1_A42	2_773
P50440	6JDW	A_1.00	A38_A423	38_423
P51149	1YHN	A_1.00	A1_A207	1_207
P51659	1ZBQ	D_0.41 E_0.41 F_0.41 A_0.41 B_0.41 C_0.41	A1_A304 B1_B304 C1_C304 D1_D304 E1_E304 F1_F304	2_736
P52566	1DS6	B_0.89	B23_B201	2_201
P52815	1dd3	0.88	P52815_1dd3_28_64_198_SwissModel	46_198
P60709	3D2U	G_0.02 C_0.02	C170_C178 G170_G178	1_375
P62995	2KXN	B_0.33	B106_B200	1_288
P63241	1FH4	A_-1.00	inconsis. annot. btw UniProt and PDB	2_154
P78310	1JEW	R_0.34	R21_R140	20_365
P83111	2p74	0.08	P83111_2p74_27_155_190_SwissModel	116_547
P99999	3NWV	D_1.00 A_1.00 B_1.00 C_1.00	A2_A105 B2_B105 C2_C105 D2_D105	2_105
Q00059	3FGH	A_0.33	A153_A219	43_246
Q05193	1DYN	A_0.15 B_0.15	A506_A633 B506_B633	1_864
Q05655	1YRK	A_0.18	A1_A123	1_676
Q08257	1YB5	A_1.00 B_1.00	A1_A329 B1_B329	1_329
Q12906	2L33	A_0.09	A521_A600	1_894
Q12931	3rlq	0.33	Q12931_3rlq_42_82_292_SwissModel	60_704

Q12983	2J5D	A_0.23 B_0.23	A146_A190 B146_B190	1_194
Q13011	2VRE	A_0.93 B_0.93 C_0.93	A50_A322 B50_B322 C50_C322	34_328
Q13162	3TJK	D_1.00 E_1.00 A_1.00 B_1.00	A38_A271 B38_B271 C38_C271 D38_D271	38_271
Q15119	2BU8	A_0.96	A16_A407	1_407
Q15149	3F7P	A_0.06 B_0.06	A1_A293 B1_B293	1_4684
Q15233	3SDE	B_0.55	B53_B312	1_471
Q15717	3HI9	D_0.25 A_0.25 B_0.25 C_0.25	A18_A99 B18_B99 C18_C99 D18_D99	1_326
Q16181	2QAG	C_0.96	C20_C437	1_437
Q16595	3T3J	A_0.76	A82_A210	42_210
Q16698	1W73	D_1.00 A_1.00 B_1.00 C_1.00	A35_A335 B35_B335 C35_C335 D35_D335	35_335
Q16740	1TG6	D_1.00 E_1.00 F_1.00 G_1.00 A_1.00 B_1.00 C_1.00	A1_A277 B1_B277 C1_C277 D1_D277 E1_E277 F1_F277 G1_G277	57_277
Q16775	1QH5	A_0.84 B_0.84	A1_A260 B1_B260	14_308
Q16836	1LSO	A_0.98 B_0.98	A7_A308 B7_B308	13_314
Q16854	2OCP	D_1.00 E_1.00 F_1.00 G_1.00 A_1.00 B_1.00 C_1.00 H_1.00	A37_A277 B37_B277 C37_C277 D37_D277 E37_E277 F37_F277 G37_G277 H37_H277	40_277
Q1L6K6	1M6I	A_-1.00	inconsis. annot. btw UniProt and PDB	55_613
Q4LE76	4FAP	B_-1.00	inconsis. annot. btw UniProt and PDB	1_2549
Q53GQ0	3d4o	0.13	Q53GQ0_3d4o_32_50_89_SwissModel	1_312
Q5JTZ9	3hy0	0.47	Q5JTZ9_3hy0_40_36_486_SwissModel	24_985
Q5ST30	2ajg	0.13	Q5ST30_2ajg_21_318_454_SwissModel	27_1063
Q5T9A4	3iij	0.05	Q5T9A4_3iij_48_347_377_SwissModel	1_648
Q68CS0	2BYJ	A_1.00 B_1.00 C_1.00	A1_A439 B1_B439 C1_C439	26_439

Q6I9V2	2w6g	-1	Q6I9V2_2w6g_97_227_297_SwissModel	-1_-1
Q6IBG8	1lvg	-1	Q6IBG8_1lvg_88_5_193_SwissModel	-1_-1
Q6IBV1	2ka1	-1	Q6IBV1_2ka1_82_178_212_SwissModel	-1_-1
Q6IMN6	1o91	0.11	Q6IMN6_1o91_33_997_1126_SwissModel	1_1127
Q6KCM7	1y1x	0.14	Q6KCM7_1y1x_29_79_145_SwissModel	1_469
Q6UB35	2eo2	0.07	Q6UB35_2eo2_92_512_574_SwissModel	32_978
Q70HW3	2c3e	0.59	Q70HW3_2c3e_22_4_165_SwissModel	1_274
Q7L2E3	2DB2	A_-1.00	inconsis. annot. btw UniProt and PDB	1_1194
Q8IXI1	2y8e	0.26	Q8IXI1_2y8e_26_3_165_SwissModel	1_618
Q8IXI2	2atv	0.26	Q8IXI2_2atv_25_4_167_SwissModel	1_618
Q8N0X7	2DL1	A_0.15	A8_A111	1_666
Q92522	1hst	0.35	Q92522_1hst_33_44_118_SwissModel	2_213
Q92947	1SIR	A_1.00	A45_A440	45_438
Q969X6	3v7d	0.05	Q969X6_3v7d_25_284_315_SwissModel	1_686
Q969Y2	3qf4	0.07	Q969Y2_3qf4_39_246_273_SwissModel	82_492
Q96RQ3	2EJM	A_0.12	A640_A725	42_725
Q96TA2	1lv7	0.32	Q96TA2_1lv7_56_332_581_SwissModel	1_773
Q99417	2YY0	D_0.51 A_0.51 B_0.51 C_0.51	A42_A94 B42_B94 C42_C94 D42_D94	2_103
Q99497	2R1V	A_0.99 B_0.99	A2_A188 B2_B188	1_189
Q99714	1SO8	A_1.00	A1_A261	2_261
Q9BPW8	1vqs	0.37	Q9BPW8_1vqs_30_178_283_SwissModel	1_284

Q9BRQ8	2wet	0.09	Q9BRQ8_2wet_38_12_44_SwissModel	2_373
Q9BV35	2b1u	0.15	Q9BV35_2b1u_31_9_77_SwissModel	1_468
Q9BV79	1ZSY	A_1.00	A40_A373	54_373
Q9BVL2	3t98	0.14	Q9BVL2_3t98_97_341_426_SwissModel	1_599
Q9BW91	1Q33	A_0.96	A59_A350	47_350
Q9BX68	3o1x	0.79	Q9BX68_3o1x_60_48_163_SwissModel	18_163
Q9BXF6	2gzh	0.09	Q9BXF6_2gzh_42_596_652_SwissModel	1_653
Q9BXK5	2IPD	@_-1.00	inconsis. annot. btw UniProt and PDB	1_485
Q9H9P8	3lov	0.11	Q9H9P8_3lov_42_47_91_SwissModel	52_463
Q9HC38	1knd	0.09	Q9HC38_1knd_34_4_32_SwissModel	2_313
Q9NQY0	2fic	0.81	Q9NQY0_2fic_20_21_225_SwissModel	1_253
Q9NR28	1FEW	A_1.00	A56_A239	56_239
Q9NS23	2KZU	B_-1.00	inconsis. annot. btw UniProt and PDB	1_344
Q9NTZ6	2CPY	A_0.11	A536_A636	1_932
Q9NUL7	2gxs	0.42	Q9NUL7_2gxs_30_127_353_SwissModel	1_540
Q9P0L0	2RR3	A_0.50	A11_A135	2_249
Q9P2R7	2fp4	0.95	Q9P2R7_2fp4_53_53_444_SwissModel	53_463
Q9UBS4	2dn9	0.19	Q9UBS4_2dn9_68_25_90_SwissModel	23_358
Q9UIJ7	1ZD8	A_1.00	A0_A226	2_227
Q9UKM9	1WF1	A_0.31	A1_A97	1_306

Q9UQ90	2QZ4	A_-1.00	inconsis. annot. btw UniProt and PDB	1_795
Q9Y237	1FJD	A_0.79	A28_A131	1_131
Q9Y3D7	2guz	0.48	Q9Y3D7_2guz_45_54_113_SwissModel	1_125
Q9Y4L1	3iuc	0.41	Q9Y4L1_3iuc_31_32_428_SwissModel	33_999

Bibliography

[AAA+10] Aranda, B., Achuthan, P., Alam-Faruque, Y., Armean, I., *et al.* (2010) The IntAct molecular interaction database in 2010, *Nucleic acids research*, 38, D525-531.

[AAA+05] Alfarano, C., Andrade, C.E., Anthony, K., Bahroos, N., *et al.* (2005) The Biomolecular Interaction Network Database and related tools 2005 update, *Nucleic acids research*, 33, D418-424.

[ABK+11] Aranda, B., Blankenburg, H., Kerrien, S., Brinkman, F.S., *et al.* (2011) PSICQUIC and PSISCORE: accessing and scoring molecular interactions, *Nature methods*, 8, 528-529.

[ABW+04] Apweiler, R., Bairoch, A., Wu, C.H., Barker, W.C., *et al.* (2004) UniProt: the Universal Protein knowledgebase, *Nucleic acids research*, 32, D115-119.

[AVB01] Achard, F., Vaysseix, G. and Barillot, E. (2001) XML, bioinformatics and data integration, *Bioinformatics (Oxford, England)*, 17, 115-125.

[Bai00] Bairoch, A. (2000) The ENZYME database in 2000, *Nucleic acids research*, 28, 304-305.

[BB05] Buchanan, B.B. and Balmer, Y. (2005) Redox regulation: a broadening horizon, *Annu Rev Plant Biol*, 56, 187-220.

[BBH03] Bader, G.D., Betel, D. and Hogue, C.W. (2003) BIND: the Biomolecular Interaction Network Database, *Nucleic acids research*, 31, 248-250.

[BG91] Blass, J.P. and Gibson, G.E. (1991) The role of oxidative abnormalities in the pathophysiology of Alzheimer's disease, *Revue neurologique*, 147, 513-525.

[BHTH07] Bernstam, E.V., Herskovic, J.R., Tantaka, L.Y. and Hersh, W. (2007) A day in the life of PubMed: Analysis of a typical day's query log, *Journal of the American Medical Informatics Association*, 14, 212-220.

[BKL+03] Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L. (2003) GenBank, *Nucleic acids research*, 31, 23-27.

[BP10] Baitaluk, M. and Ponomarenko, J. (2010) Semantic integration of data on transcriptional regulation, *Bioinformatics (Oxford, England)*, 26, 1651-1661.

[BTW+07] Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., *et al.* (2007) NCBI GEO: mining tens of millions of expression profiles--database and tools update, *Nucleic acids research*, 35, D760-765.

[BTW+10] Barrett, T., Troup, D.B., Wilhite, S.E., Ledoux, P., *et al.* NCBI GEO: archive for functional genomics data sets--10 years on, *Nucleic acids research*, 39, D1005-1010.

[BVT+04] Balmer, Y., Vensel, W.H., Tanaka, C.K., Hurkman, W.J., *et al.* (2004) Thioredoxin links redox to the regulation of fundamental processes of plant mitochondria, *Proc Natl Acad Sci U S A*, 101, 2642-2647.

[CAD+09] Caspi, R., Altman, T., Dale, J.M., Dreher, K., *et al.* (2010) The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases, *Nucleic acids research*, 38, D473-479.

[CAL+10] Ceol, A., Chatr Aryamontri, A., Licata, L., Peluso, D., Briganti, L., Perfetto, L., Castagnoli, L. and Cesareni, G. MINT, the molecular interaction database: 2009 update, *Nucleic acids research*, 38, D532-539

[CGG04] Conour, J.E., Graham, W.V. and Gaskins, H.R. (2004) A combined in vitro/bioinformatic investigation of redox regulatory mechanisms governing cell cycle progression, *Physiol Genomics*, 18, 196-205.

[CH08] Cohen, K. and Hunter, L. (2008) Getting started in text mining, *PLoS Comput Biol*, 4.

[CLN+03] Cuellar, A.A., Lloyd, C.M., Nielsen, P.F., Bullivant, D.P., Nickerson, D.P. and Hunter, P.J. (2003) An overview of CellML 1.1, a biological model description language, *Simulation-Transactions of the Society for Modeling and Simulation International*, 79, 740-747.

[CM10] Collet, J.F. and Messens, J. Structure, function, and mechanism of thioredoxin proteins, *Antioxid Redox Signal*, 13, 1205-1216.

[CSC+07] Cline, M. S., Smoot, M., Cerami, E., Kuchinsky, A., *et al.* (2007) Integration of biological networks and gene expression data using Cytoscape, *Nat Protoc*, 2, 2366-2382.

[CSG+09] Chang, A., Scheer, M., Grote, A., Schomburg, I. and Schomburg, D. (2009) BRENDA, AMENDA and FRENDA the enzyme information system: new content and tools in 2009, *Nucleic acids research*, 37, D588-592.

[CWH+02] Cheung, K.H., White, K., Hager, J., Gerstein, M., *et al.* (2002) YMD: a microarray database for large-scale gene expression analysis, *Proceedings AMIA 2002 Annual Symposium*, 140-144.

[DBD+02] Demir, E., Babur, O., Dogrusoz, U., Gursoy, A., *et al.* (2002) PATIKA: an integrated visual environment for collaborative construction and analysis of cellular pathways, *Bioinformatics (Oxford, England)*, 18, 996-1003.

[DBJ+07] Duarte, N.C., Becker, S.A., Jamshidi, N., Thiele, I., Mo, M.L., Vo, T.D., Srivas, R. and Palsson, B.O. (2007) Global reconstruction of the human metabolic network based on genomic and bibliomic data, *Proc Natl Acad Sci U S A*, 104, 1777-1782.

[DCP+10] Demir, E., Cary, M.P., Paley, S., Fukuda, K., *et al.* (2010) The BioPAX community standard for pathway data sharing, *Nat Biotechnol*, 28, 935-942.

[DF10] De Las Rivas, J. and Fontanillo, C. (2010) Protein-Protein Interactions Essentials: Key Concepts to Building and Analyzing Interactome Networks, *Plos Computational Biology*, 6.

[Diet03] Dietz, K.J. (2003) Redox control, redox signaling, and redox homeostasis in plant cells, *International review of cytology*, 228, 141-193.

[Diet08] Dietz, K.J. (2008) Redox signal integration: from stimulus to networks and genes, *Physiologia plantarum*, 133, 459-468.

[DKS89] DiGiacomo, R.A., Kremer, J.M. and Shah, D.M. (1989) Fish-oil dietary supplementation in patients with Raynaud's phenomenon: a double-blind, controlled, prospective study, *The American journal of medicine*, 86, 158-164.

[DKZ+12] Dreszer, T.R., Karolchik, D., Zweig, A.S., Hinrichs, A.S., *et al.* The UCSC Genome Browser database: extensions and updates 2011, *Nucleic acids research*, 40, D918-923.

[DMNL09] Dogan, I.D., Murray, G.C., Neveol, A, Lu, Z. (2009) Understanding PubMed user search behavior through log analysis, *Database (Oxford)*, 2009, baq018.

[DP11] Dietz, K.J. and Pfannschmidt, T. Novel regulators in photosynthetic redox control of plant metabolism and gene expression, *Plant physiology*, 155, 1477-1485.

[DTMF06] Davies, M.N., Toseland, C.P., Moss, D.S. and Flower, D.R. (2006) Benchmarking pK(a) prediction, *BMC Biochem*, 7, 18.

[DWH10] Divoli, A., Wooldridge, M.A. and Hearst, M.A. Full text and figure display improves bioscience literature search, *PLoS One*, 5, e9619.

[Ferr09] Ferrer, I. (2009) Early involvement of the cerebral cortex in Parkinson's disease: convergence of multiple metabolic defects, *Progress in neurobiology*, 88, 89-103.

[FMG08] Fomenko, D.E., Marino, S.M. and Gladyshev, V.N. (2008) Functional diversity of cysteine residues in proteins and unique features of catalytic redox-active cysteines in thiol oxidoreductases, *Mol Cells*, 26, 228-235.

[FMJ+08] Funahashi, A., Matsuoka Y., Jouraku A., Morohashi M., Kikushi N., Kitano H.(2008) ,CellDesigner 3.5: A versatile modeling tool for biochemical networks, *Proc. IEEE*, 96(8), 1254-65

[FWL+09] Fu, C., Wu, C., Liu, T., Ago, T., Zhai, P., Sadoshima, J. and Li, H. (2009) Elucidation of thioredoxin target protein networks in mouse, *Mol Cell Proteomics*, 8, 1674-1687.

[FXA+07] Fomenko, D.E., Xing, W., Adair, B.M., Thomas, D.J. and Gladyshev, V.N. (2007) High-throughput identification of catalytic redox-active cysteine residues, *Science*, 315, 387-389.

[GGM11] Gnad, F., Gunawardena, J. and Mann, M. (2011) PHOSIDA 2011: the posttranslational modification database, *Nucleic acids research*, 39, D253-260.

[GHM+10] Gibson, F., Hoogland, C., Martinez-Bartolome, S., Medina-Aunon, J.A., *et al.* (2010) The gel electrophoresis markup language (GelML) from the Proteomics Standards Initiative, *Proteomics*, 10, 3073-3081.

[GNC+08] Garny, A., Nickerson, D.P., Cooper, J., dos Santos, R.W., Miller, A.K., McKeever, S., Nielsen, P.M.F. and Hunter, P.J. (2008) CellML and associated tools and techniques, *Philosophical Transactions of the Royal Society a-Mathematical Physical and Engineering Sciences*, 366, 3017-3043.

[Go10] Go, E.P. (2010) Database Resources in Metabolomics: An Overview, *Journal of Neuroimmune Pharmacology*, 5, 18-30.

[GOC12] Gene Ontology Consortium (2012) The Gene Ontology: enhancements for 2011, *Nucleic acids research*, 40, D559-564.

[GSZK09] Gopal, S., Srinivas, V., Zameer, F. and Kreft, J. (2009) Prediction of proteins putatively involved in the thiol:disulfide redox metabolism of a bacterium (Listeria): the CXXC motif as query sequence, *In Silico Biology*, 9:0032

[HBH+10] Hucka, M., Bergmann, F. T., Hoops, S., Keating, S. M. *et al.* (2010) The Systems Biology Markup Language (SBML): Language Specification for Level 3 Version 1 Core, *Natura Precedings*

[HDG+07] Hearst, M.A., Divoli, A., Guturu, H., Ksikes, A., Nakov, P., Wooldridge, M.A. and Ye, J. (2007) BioText Search Engine: beyond abstract search, *Bioinformatics (Oxford, England)*, 23, 2196-2197.

[HHF+05] Hisabori, T., Hara, S., Fujii, T., Yamazaki, D., Hosoya-Matsuda, N. and Motohashi, K. (2005) Thioredoxin affinity chromatography: a useful method for further understanding the thioredoxin network, *J Exp Bot*, 56, 1463-1468.

[HJB+05] Holmgren, A., Johansson, C., Berndt, C., Lonn, M.E., *et al.* (2005) Thiol redox control via thioredoxin and glutaredoxin systems, *Biochemical Society transactions*, 33, 1375-1377.

[HKT+10] Hippe, K., Kormeier, B., Toepel, T., Janowski, S., Hofestaedt, R. (2010) DAWIS-M.D. - a data warehouse system for metabolic data. *http://agbi.techfak.uni-bielefeld. de/DAWISMD/pdf/dawismd.pdf*

[HML+04] Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., *et al.* (2004) IntAct: an open source molecular interaction database, *Nucleic acids research*, 32, D452-455.

[HMS+11] Huang, H.H., H. Z., McGarvey, P.B., Suzek, B.E., Mazumder, R., Zhang, J.A., Chen, Y.X. and Wu, C.H. (2011) A comprehensive protein-centric ID mapping service for molecular data integration, *Bioinformatics (Oxford, England)*, 27, 1190-1191.

[HMZG10] Hao, T., Ma, H.W., Zhao, X.M. and Goryanin, I. (2010) Compartmentalization of the Edinburgh Human Metabolic Network, *BMC bioinformatics*, 11, 393.

[HNY+07] Hu, Z., Ng, D.M., Yamada, T., Chen, C., *et al.* (2007) VisANT 3.0: new modules for pathway visualization, editing, prediction and construction, *Nucleic acids research*, 35, W625-632.

[Hol89] Holmgren, A. (1989) Thioredoxin and glutaredoxin systems, *J Biol Chem*, 264, 13963-13966.

[HPMH05] Hristovski, D., Peterlin, B., Mitchell, J.A. and Humphrey, S.M. (2005) Using literature-based discovery to identify disease candidate genes, *Int J Med Inform*, 74, 289-298.
]
[HSA+05] Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., *et al.* (2005) Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders, *Nucleic Acids Res*, 33, D514-517.

[HSS+08] Huang, D.W., Sherman, B.T., Stephens, R., Baseler, M.W., Lane, H. C., Lempicki, R.A. (2008) DAVID gene ID conversion tool, *Bioinformatition*, 2, 428-430.

[HT93] Hubbard, S.J., Thornton, J.M. (1993) Naccess, Computer Program, Department of Biochemistry and Molecular Biology, University College London

[JSB06] Jensen, L.J., Saric, J. and Bork, P. (2006) Literature mining for the biologist: from information retrieval to biological discovery, *Nature reviews*, 7, 119-129.

[JKT+10] Janowski, S., Kormeier, B., Töpel T., Hippe, K., *et al*. (2010) Modeling of cell-to-cell communication processes with Petri nets using the example of quorum sensing, *In Silico Biology*, 10:0003

[KAA+06] Kulikova, T., Akhtar, R., Aldebert, P., Althorpe, N., *et al.* (2007) EMBL Nucleotide Sequence Database in 2006, *Nucleic acids research*, 35, D16-20.

[KAB+12] Kerrien, S., Aranda, B., Breuza, L., Bridge, A., *et al.* (2012) The IntAct molecular interaction database in 2012, *Nucleic acids research*, 40, D841-846.

[KAF+07] Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., *et al.* (2007) IntAct--open source resource for molecular interaction data, *Nucleic acids research*, 35, D561-565.

[KAG+08] Kanehisa, M., Araki, M., Goto, S., Hattori, M., *et al.* (2008) KEGG for linking genomes to life and the environment, *Nucleic acids research*, 36, D480-D484.

[KAK+09] Kiefer, F., Arnold, K., Kunzli, M., Bordoli, L. and Schwede, T. (2009) The SWISS-MODEL Repository and associated resources, *Nucleic acids research*, 37, D387-392.

[KAN+10] Katayama, T., Arakawa, K., Nakao, M., Ono, K., *et al.* (2010) The DBCLS BioHackathon: standardization and interoperability for bioinformatics web services and workflows. The DBCLS BioHackathon Consortium*, *Journal of biomedical semantics*, 1, 8.

[KB03] Kanehisa, M. and Bork, P. (2003) Bioinformatics in the post-sequence era, *Nature genetics*, 33 Suppl, 305-310.

[KBB05] Kiemer, L., Bendtsen, J.D. and Blom, N. (2005) NetAcet: prediction of N-terminal acetylation sites, *Bioinformatics (Oxford, England)*, 21, 1269-1270.

[KBB+07] Kuntzer, J., Backes, C., Blum, T., Gerasch, A., *et al.* (2007) BNDB - the Biochemical Network Database, *BMC bioinformatics*, 8, 367.

[KGH+06] Kanehisa, M., Goto, S., Hattori, M., Aoki-Kinoshita, K.F., Itoh, M., Kawashima, S., Katayama, T., Araki, M. and Hirakawa, M. (2006) From genomics to chemical genomics: new developments in KEGG, *Nucleic acids research*, 34, D354-D357.

[KH80] Kallis, G.B. and Holmgren, A. (1980) Differential reactivity of the functional sulfhydryl groups of cysteine-32 and cysteine-35 present in the reduced form of thioredoxin from Escherichia coli, *J Biol Chem*, 255, 10261-10265.

[KNT10] Katayama, T., Nakao, M. and Takagi, T. (2010) TogoWS: integrated SOAP and REST APIs for interoperable bioinformatics Web services, *Nucleic acids research*, 38, W706-W711.

[KOM+05] Karp, P.D., Ouzounis, C.A., Moore-Kochlacs, C., Goldovsky, L., *et al.* (2005) Expansion of the BioCyc collection of pathway/genome databases to 160 genomes, *Nucleic acids research*, 33, 6083-6089.

[KPK+10] Karp, P.D., Paley, S.M., Krummenacker, M., Latendresse, M., *et al.* (2010) Pathway Tools version 13.0: integrated software for pathway/genome informatics and systems biology, *Briefings in Bioinformatics*, 11, 40-79.

[KPV+06] Krull, M., Pistor, S., Voss, N., Kel, A., *et al.* (2006) TRANSPATH: an information resource for storing and visualizing signaling pathways and their pathological aberrations, *Nucleic acids research*, 34, D546-551.

[KSF+12] Kuhn, M., Szklarczyk, D., Franceschini, A., von Mering, C., *et al.* (2012) STITCH 3: zooming in on protein-chemical interactions, *Nucleic acids research*, 40, D876-880.

[KZM+04] Krieger, C.J., Zhang, P., Mueller, L.A., Wang, A., Paley, S., Arnaud, M., Pick, J., Rhee, S.Y. and Karp, P.D. (2004) MetaCyc: a multiorganism database of metabolic pathways and enzymes, *Nucleic acids research*, 32, D438-442.

[LAC+04] Lemer, C., Antezana, E., Couche, F., Fays, F., *et al.* (2004) The aMAZE LightBench: a web interface to a relational database of cellular processes, *Nucleic acids research*, 32, D443-448.

[LDH10] Lee, H.M., Dietz, K.J. and Hofestadt, R. (2010) Prediction of thioredoxin and glutaredoxin target proteins by identifying reversibly oxidized cysteinyl residues, *Journal of integrative bioinformatics*, 7.

[LDR+10] Laibe, C., Li, C., Donizelli, M., Rodriguez, N., *et al.* (2010) BioModels Database: An enhanced, curated and annotated resource for published quantitative kinetic models, *Bmc Systems Biology*, 4.

[Lidd05] Liddy, E. D. (2005) Automatic Document Retrieval, *In Encyclopedia of Language and Linguistics*, 2nd Edition, Elsevier Press

[LMZ+07] Lemaire, S.D., Michelet, L., Zaffagnini, M., Massot, V. and Issakidis-Bourguet, E. (2007) Thioredoxins in chloroplasts, *Curr Genet*, 51, 343-365.

[LPW+06] Lee, T.J., Pouliot, Y., Wagner, V., Gupta, P., Stringer-Calvert, D.W.J., Tenenbaum, J.D. and Karp, P.D. (2006) BioWarehouse: a bioinformatics database warehouse toolkit, *BMC bioinformatics*, 7.

[LR71] Lee, B. and Richards, F.M. (1971) The interpretation of protein structures: estimation of static accessibility, *J Mol Biol*, 55, 379-400.

[Lu11] Lu, Z. (2011) PubMed and beyond: a survey of web tools for searching biomedical literature, *Database (Oxford)*, 2011, baq036.

[MAA+05] Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., *et al.* (2005) InterPro, progress and status in 2005, *Nucleic acids research*, 33, D201-205.

[MBHC95] Murzin, A.G., Brenner, S.E., Hubbard, T. and Chothia, C. (1995) SCOP: a structural classification of proteins database for the investigation of sequences and structures, *J Mol Biol*, 247, 536-540.

[MBVR09] Meyer, Y., Buchanan, B.B., Vignols, F. and Reichheld, J.P. (2009) Thioredoxins and glutaredoxins: unifying elements in redox biology, *Annual review of genetics*, 43, 335-367.

[MCS+11] Martens, L., Chambers, M., Sturm, M., Kessner, D., *et al.* (2011) mzML--a community standard for mass spectrometry data, *Mol Cell Proteomics*, 10, R110 000133.

[MDM+10] Mi, H., Dong, Q., Muruganujan, A., Gaudet, P., *et al.* (2010) PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium, *Nucleic acids research*, 38, D204-210.

[MFG+03] Matys, V., Fricke, E., Geffers, R., Gossling, E., *et al.* (2003) TRANSFAC (R): transcriptional regulation, from patterns to profiles, *Nucleic acids research*, 31, 374-378.

[MG09] Marino, S.M. and Gladyshev, V.N. (2009) A structure-based approach for detection of thiol oxidoreductases and their catalytic redox-active cysteine residues, *PLoS Comput Biol*, 5, e1000383.

[MHF+06] Maeda, K., Hagglund, P., Finnie, C., Svensson, B. and Henriksen, A. (2006) Structural basis for target protein recognition by the protein disulfide reductase thioredoxin, *Structure*, 14, 1701-1710.

[MHMF96] Martinez, M., Hernandez, A.I., Martinez, N. and Ferrandiz, M.L. (1996) Age-related increase in oxidized proteins in mouse synaptic mitochondria, *Brain research*, 731, 246-248.

[MKF+06] Matys, V., Kel-Margoulis, O.V., Fricke, E., Liebich, I., *et al.* (2006) TRANSFAC and its module TRANSCompel: transcriptional gene regulation in eukaryotes, *Nucleic acids research*, 34, D108-110.

[MMB+11] Muhlberger, I., Moenks, K., Bernthaler, A., Jandrasits, C., *et al.* (2011) Integrative bioinformatics analysis of proteins associated with the cardiorenal syndrome, *International journal of nephrology*, 2011, 809378.

[MSM+07] Ma, H., Sorokin, A., Mazein, A., Selkov, A., *et al.* (2007) The Edinburgh human metabolic network reconstruction and its functional analysis, *Molecular systems biology*, 3, 135.

[MYC+02] Mellor, J.C., Yanai, I., Clodfelter, K.H., Mintseris, J. and DeLisi, C. (2002) Predictome: a database of putative functional links between proteins, *Nucleic acids research*, 30, 306-309.

[NBG+06] Ng, A., Bursteinas, B., Gao, Q., Mollison, E. and Zvelebil, M. (2006) Resources for integrative systems biology: from data through databases to networks and dynamic system models, *Brief Bioinform*, 7, 318-330.

[NDMM04] Nagasaki, M., Doi, A., Matsuno, H. and Miyano, S. (2004) A versatile petri net based architecture for modeling and simulation of complex biological processes, *Genome informatics*, 15, 180-197.

[NM01] Noy, N. F. and McGuinness, D. L. (*2001*). Ontology development 101: a guide to creating your first ontology. Technical Report KSL-01-05 and Stanford Medical Informatics Technical Report SMI-2001-0880. Stanford Knowledge Systems Laboratory. Available at *http://www.ksl.stanford.edu/people/dlm/papers/ontology-tutorial-noy-mcguinness-abstract.html*

[NSJ+10] Nagasaki, M., Saito, A., Jeong, E., Li, C., *et al.* (2010) Cell illustrator 4.0: a computational platform for systems biology, *In silico biology*, 10, 5-26.

[OH07] Orchard, S. and Hermjakob, H. (2007) The HUPO proteomics standards initiative–easing communication and minimizing data loss in a changing world, *Briefings in Bioinformatics*, 9, 166-173.

[OMM+07] Okumura, T., Makiguchi, H., Makita, Y., Yamashita, R., *et al.* (2007) Melina II: a web tool for comparisons among several predictive algorithms to find potential motifs from promoter regions, *Nucleic acids research*, 35, W227-231.

[PGK+09] Keshava Prasad, T.S., Goel, R., Kandasamy, K., Keerthikumar, S., *et al.* (2009) Human Protein Reference Database--2009 update, *Nucleic acids research*, 37, D767-772.

[PNA+03] Peri, S., Navarro, J.D., Amanchy, R., Kristiansen, T.Z., *et al.* (2003) Development of human protein reference database as an initial platform for approaching systems biology in humans, *Genome Res*, 13, 2363-2371.

[PNK+04] Peri, S., Navarro, J.D., Kristiansen, T.Z., Amanchy, R., *et al.* (2004) Human protein reference database as a discovery resource for proteomics, *Nucleic acids research*, 32, D497-501.

[PSK+11] Parkinson, H., Sarkans, U., Kolesnikov, N., Abeygunawardena, N., *et al.* (2011) ArrayExpress update--an archive of microarray and high-throughput sequencing-based functional genomics experiments, *Nucleic acids research*, 39, D1002-1004.

[PT90] Pollock, R. and Treisman, R. (1990) A sensitive method for the determination of protein-DNA binding specificities, *Nucleic acids research*, 18, 6197-6204.

[PTK+09] Portales-Casamar, E., Thongjuea, S., Kwon, A.T., Arenillas, D., *et al.* (2009) JASPAR 2010: the greatly expanded open-access database of transcription factor binding profiles, *Nucleic acids research*, 38, D105-110.

[PTM05] Pruitt, K.D., Tatusova, T. and Maglott, D.R. (2005) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins, *Nucleic acids research*, 33, D501-504.

[PUD+01] Peterson, J.D., Umayam, L.A., Dickinson, T., Hickey, E.K. and White, O. (2001) The Comprehensive Microbial Resource, *Nucleic acids research*, 29, 123-125.

[PYD+11] Podkolodnaya, O.A., Yarkova, E.E., Demenkov, P.S., Konovalova, O.S., Ivanisenko, V.A. and Kolchanov, N.A. (2011) Application of the ANDCellComputer System to Reconstruction and Analysis of Associative Networks Describing Potential Relationships between Myopia and Glaucoma, *Russian Journal of Genetics: Applied Research*, 1, 21-28.

[RAG+08] Rebholz-Schuhmann, D., Arregui, M., Gaudan, S., Kirsch, H. and Jimeno, A. (2008) Text processing through Web services: calling Whatizit, *Bioinformatics (Oxford, England)*, 24, 296-298.

[RBB+11] Rose, P.W., Beran, B., Bi, C., Bluhm, W.F., *et al.* The RCSB Protein Data Bank: redesigned web site and web services, *Nucleic acids research*, 39, D392-401.

[Rodr09] Rodriguez-Esteban, R. (2009) Biomedical Text Mining and Its Applications, *Plos Computational Biology*, 5.

[RSN07] Rubin, D.L., Shah, N.H. and Noy, N.F. (2008) Biomedical ontologies: a functional perspective, *Briefings in Bioinformatics*, 9, 75-90.

[RTD+10] Radrich, K., Tsuruoka, Y., Dobson, P., Gevorgyan, A., Swainston, N., Baart, G. and Schwartz, J.M. (2010) Integration of metabolic databases for the reconstruction of genome-scale metabolic networks, *BMC Syst Biol*, 4, 114.

[RVS+05] Rouhier, N., Villarejo, A., Srivastava, M., Gelhaye, E., *et al.* (2005) Identification of plant glutaredoxin targets, *Antioxidants & redox signaling*, 7, 919-929.

[RWG+04] Romero, P., Wagg, J., Green, M.L., Kaiser, D., *et al.* (2005) Computational prediction of human metabolic pathways from the complete human genome, *Genome biology*, 6, R2.

[SAC10] Sarac, O.S., Atalay, V. and Cetin-Atalay, R. (2010) GOPred: GO molecular function prediction by combined classifiers, *PloS one*, 5, e12382.

[SAC+08] Stockinger, H., Attwood, T., Chohan, S.N., Cote, R., *et al.* (2008) Experience using web services for biological sequence analysis, *Briefings in Bioinformatics*, 9, 493-505.

[SAE+04] Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W. and Lenhard, B. (2004) JASPAR: an open-access database for eukaryotic transcription factor binding profiles, *Nucleic acids research*, 32, D91-94.

[SAK+09] Mi, H., Dong, Q., Muruganujan, A., Gaudet, P., *et al.* (2010) PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium, *Nucleic acids research*, 38, D204-210.

[SAR+07] Smith, B., Ashburner, M., Rosse, C., Bard, J., *et al.* (2007) The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration, *Nature Biotechnology*, 25, 1251-1255.

[SBR+06] Stark, C., Breitkreutz, B.J., Reguly, T., Boucher, L., Breitkreutz, A. and Tyers, M. (2006) BioGRID: a general repository for interaction datasets, *Nucleic acids research*, 34, D535-539.

[SD06] Stroher, E. and Dietz, K.J. (2006) Concepts and approaches towards understanding the cellular redox proteome, *Plant biology (Stuttgart, Germany)*, 8, 407-418.

[SD08] Stroher, E. and Dietz, K.J. (2008) The dynamic thiol-disulphide redox proteome of the Arabidopsis thaliana chloroplast as revealed by differential electrophoretic mobility, *Physiologia plantarum*, 133, 566-583.

[SMS+04] Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U. and Eisenberg, D. (2004) The Database of Interacting Proteins: 2004 update, *Nucleic acids research*, 32, D449-451.

[SMS+02] Spellman, P.T., Miller, M., Stewart, J., Troup, C., *et al.* (2002) Design and implementation of microarray gene expression markup language (MAGE-ML), *Genome Biology*, 3.

[SOR+11] Smoot, M.E., Ono, K., Ruscheinski, J., Wang, P.L., *et al.* (2011) Cytoscape 2.8: new features for data integration and network visualization, *Bioinformatics (Oxford, England)*, 27, 431-432.

[SOS96] Sanner, M.F., Olson, A.J. and Spehner, J.C. (1996) Reduced surface: an efficient way to compute molecular surfaces, *Biopolymers*, 38, 305-320.

[SRWM08] Sanchez, R., Riddle, M., Woo, J. and Momand, J. (2008) Prediction of reversibly oxidized protein cysteine thiols using protein structure properties, *Protein Sci*, 17, 473-481.

[SST06] Swanson, D.R., Smalheiser, N.R. and Torvik, V.I. (2006) Ranking indirect connections in literature-based discovery: The role of medical subject headings, *Journal of the American Society for Information Science and Technology*, 57, 1427-1439.

[Stei03] Stein, L.D. (2003) Integrating biological databases, *Nature Reviews Genetics*, 4, 337-345.

[Swan86] Swanson, D. R. (1986). Fish iol, Raynaud's syndrome, and undiscovered public knowledge, *Perspect Biol Med.*, 30(1):7-18.

[SWL+05] Stelzl, U., Worm, U., Lalowski, M., Haenig, C., *et al.* (2005) A human protein-protein interaction network: a resource for annotating the proteome, *Cell*, 122, 957-968.

[TGOC00] The Gene Ontology Consortium (2000) Gene ontology: tool for the unification of biology, Nat. Genet., 25(1), 25-29.

[TMH+11] Tsuruoka, Y., Miwa, M., Hamamoto, K., Tsujii, J., *et al.* (2011) Discovering and visualizing indirect associations between biomedical concepts, *Bioinformatics*, 27, i111-119.

[TUC10] The UniProt Consortium. (2010) The Universal Protein Resource (UniProt) in 2010, *Nucleic Acids Res*, 38, D142-148.

[VDS+07] Vastrik, I., D'Eustachio, P., Schmidt, E., Gopinath, G., *et al.* (2007) Reactome: a knowledge base of biologic pathways and processes, *Genome Biol*, 8, R39.

[VLM+07] Valko, M., Leibfritz, D., Moncol, J., Cronin, M. T., *et al.* (2007) Free radicals and antioxidants in normal physiological functions and human disease, *Int J Biochem Cell Biol*, 39(1): 44-84

[VSC+92] VanBogelen, R.A., Sankar, P., Clark, R.L., Bogan, J.A. and Neidhardt, F.C. (1992) The gene-protein database of Escherichia coli: edition 5, *Electrophoresis*, 13, 1014-1054.

[VSP+08] Viswanathan, G.A., Seto, J., Patil, S., Nudelman, G. and Sealfon, S.C. (2008) Getting started in biological pathway construction and analysis, *PLoS Comput Biol*, 4, e16.

[WAB+06] Wu, C.H., Apweiler, R., Bairoch, A., Natale, D.A., *et al.* (2006) The Universal Protein Resource (UniProt): an expanding universe of protein information, *Nucleic acids research*, 34, D187-191.

[WCH+00] Wingender, E., Chen, X., Hehl, R., Karas, H., *et al.* (2000) TRANSFAC: an integrated system for gene expression regulation, *Nucleic acids research*, 28, 316-319.

[WCH+01] Wangensteen, O.S., Chueca, A., Hirasawa, M., Sahrawy, M., *et al.* (2001) Binding features of chloroplast fructose-1,6-bisphosphatase-thioredoxin interaction, *Biochimica et biophysica acta*, 1547, 156-166.

[WCL+00] Wheeler, D.L., Chappey, C., Lash, A.E., Leipe, D.D., *et al.* (2000) Database resources of the National Center for Biotechnology Information, *Nucleic acids research*, 28, 10-14.

[WFH10] Wouters, M.A., Fan, S.W. and Haworth, N.L. (2010) Disulfides as redox switches: From molecular mechanisms to functional significance, *Antioxidants & redox signaling*, 12, 53-91.

[WHN+04] Wu, C.H., Huang, H., Nikolskaya, A., Hu, Z. and Barker, W.C. (2004) The iProClass integrated database for protein functional analysis, *Comput Biol Chem*, 28, 87-96.

[WKG+09] Wishart, D.S., Knox, C., Guo, A.C., Eisner, R., *et al.* (2009) HMDB: a knowledgebase for the human metabolome, *Nucleic acids research*, 37, D603-610.

[WKM05] Weeber, M., Kors, J.A. and Mons, B. (2005) Online tools to support literature-based discovery in the life sciences, *Briefings in Bioinformatics*, 6, 277-286.

[XW10] Xia, J. and Wishart, D.S. (2010) MetPA: a web-based metabolomics tool for pathway analysis and visualization, *Bioinformatics (Oxford, England)*, 26, 2342-2344.

[YSN+11] Yamamoto, S., Sakai, N., Nakamura, H., Fukagawa, H., *et al.* (2011) INOH: ontology-based highly structured database of signal transduction pathways, *Database (Oxford)*, 2011, bar052.

[YWL+01] Yano, H., Wong, J.H., Lee, Y.M., Cho, M.J. and Buchanan, B.B. (2001) A strategy for the identification of proteins targeted by thioredoxin, *Proc Natl Acad Sci U S A*, 98, 4794-4799.

[ZMQ+02] Zanzoni, A., Montecchi-Palazzi, L., Quondam, M., Ausiello, G., Helmer-Citterich, M. and Cesareni, G. (2002) MINT: a Molecular INTeraction database, *FEBS Lett*, 513, 135-140.

[Url1] PubMed Tutorial, http://www.nlm.nih.gov/bsd/disted/pubmedtutorial/ index.html

[Url2] Medical Subject Headings, http://www.nlm.nih.gov/mesh/meshhome.html

[Url3] FAQ, The Systems Biology Markup Language, http://sbml.org/ Documents/ FAQ

[Url4] Gene Ontology Documentation, http://www.geneontology.org/GO. contents. doc.shtml

[Url5] BioPax documentation, http://biopax.hg.sourceforge.net/hgweb/biopax/ biopax/file/default/Level3/docs

[Url6] The Open Biological and Biomedical Ontologies, http://www.obofoundry. org/

[Url7] CSML, http://www.csml.org/csml/

[Url8] PubMed: The Bibliographic Database, The NCBI Handbook, http://www. ncbi.nlm.nih.gov/ books/NBK21101/

[Url9] Fact Sheet-Online Indexing System, http://www.nlm.nih.gov/pubs/ factsheets/online_indexing_system.html

[Url10] Frequently asked questions about indexing, Bibliographic Services Division, U.S. National Library of Medicine, http://www.nlm.nih.gov/bsd/indexfaq.html

[Url11] Apache Lucene project, http://lucene.apache.org/

[Url12] PubMed Central Help, http://www.ncbi.nlm.nih.gov/ books/NBK3826/

[Url13] Journal Publishing Tag Set, http://dtd.nlm.nih.gov/publishing/

[Url14] Java Architecture for XML Binding (JAXB), http://www.oracle.com/ technetwork/articles/javase/index-140168.html

[Url15] Unified Medical Language System, http://www.nlm.nih.gov/research/ umls/

[Url16] PubMed Tutorial, http://www.nlm.nih.gov/bsd/disted/pubmedtutorial/ glossary.html

[Url17] BN++, http://www.bnplusplus.org/database/links