Prediction of thioredoxin and glutaredoxin target proteins by identifying reversibly oxidized cysteinyl residues

Hang-mao Lee[1*], Karl Josef Dietz[2], and Ralf Hofestädt[1]

[1] Technical Faculty, Bielefeld University, Germany

[2] Faculty of Biology, Bielefeld University, Germany

[*] Corresponding author - hmlee@techfak.uni-bielefeld.de

Summary

A significant part of cellular proteins undergo reversible thiol-dependent redox transitions which often control or switch protein functions. Thioredoxins and glutaredoxins constitute two key players in this redox regulatory protein network. Both interact with various categories of proteins containing reversibly oxidized cysteinyl residues. The identification of thioredoxin/glutaredoxin target proteins is a critical step in constructing the redox regulatory network of cells or subcellular compartments. Due to the scarcity of thioredoxin/glutaredoxin target protein records in the public database, a tool called Reversibly Oxidized Cysteine Detector (ROCD) is implemented here to identify potential thioredoxin/glutaredoxin target proteins computationally, so that the in silico construction of redox regulatory network may become feasible. ROCD was tested on 46 thioredoxin target proteins in plant mitochondrion, and the recall rate was 66.7% when 50% sequence identity was chosen for structural model selection. ROCD will be used to predict the thioredoxin/glutaredoxin target proteins in human liver mitochondrion for our redox regulatory network construction project. The ROCD will be developed further to provide prediction with more reliability and incorporated into biological network visualization tools as a node prediction component. This work will advance the capability of traditional database- or text mining-based method in the network construction.

1 Introduction

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 [1,2,3,4]. On the molecular scale, oxidation will change the structure of biomolecules and often switches or tunes enzyme activity or causes enzyme 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 antioxidant enzymes [5]. 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 dithiol-disulfide transition. Cells have developed two rescue systems that involve thioredoxins and glutaredoxins [6,7], respectively, to re-reduce the oxidized proteins. Because these two proteins 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, they are termed ‘transmitters’ in the redox regulatory network [4].

The mitochondrion is a subcellular compartment where respiratory electron transport proceeds at high rate and with strong reduction potential differences. If NADH availability is high, electron carriers such as ubiquinone may become over-reduced during electron transport

at the mitochondrial inner membrane. In this case, electrons may be transferred to the oxygen molecule to produce superoxide, which is a strong oxidant [8]. Thus the thioredoxin and glutaredoxin systems in the mitochondrion play an important role in preventing this organelle from over-oxidized. Several researches have shown that oxidized mitochondrial proteins accumulate with aging and neural degenerative disease [1-3]. In order to ultimately simulate electron flow through the redox regulatory system, the thiol-disulfide protein network of human mitochondrion shall be constructed. This will allow testing of its capability to maintain redox homeostasis and to infer the biological outcome by the downstream regulated metabolic network in case of oxidative stress. The relationship between the redox regulatory network and the affected metabolic network is depicted in Fig. 1.

The proteins connecting the upstream regulatory network and the downstream metabolic network are the target proteins of thioredoxin/glutaredoxin. The critical step in expanding the redox regulatory network is to identify the thioredoxin/glutaredoxin target proteins in order to complete the network. The thioredoxin/glutaredoxin target proteins contain reversibly oxidized cysteines which form disulfide bonds with another cysteine bearing the same property when oxidized and are present in free thiol form when reduced. But unfortunately only few thioredoxin/glutaredoxin target proteins could be found in the public protein-protein interaction databases when human liver mitochondrion was chosen as our modeling environment. The lack of database and experimental information thus becomes an obstacle in the network construction process.

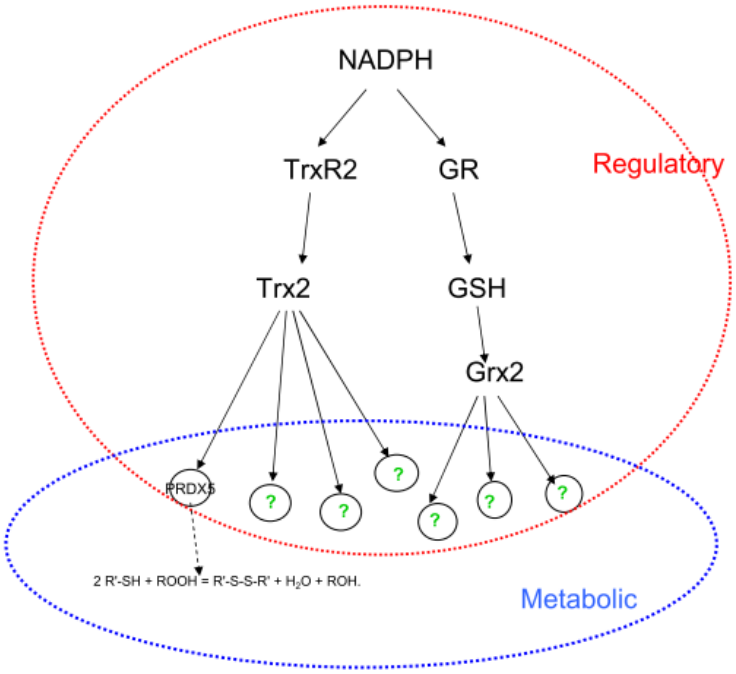

Fig. 1: Redox regulatory network and exemplarily affected metabolic network. Abbreviations: TrxR: thioredoxin reductase [9], GR: glutathione reductase, GSH: glutathione, Grx2: glutaredoxin.

With the advance of proteomics techniques, chromatography coupled with gel-based or mass spectrometry can experimentally identify thioredoxin/glutaredoxin target proteins or proteins that undergo thiol-disulfide transitions [10-13]. Due to technical limitation, further experiments are needed to eliminate the false positive proteins. In addition, 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. Therefore this work is aimed at deploying the capability of bioinformatics to identify additional potential target protein candidates to complete the network and also to support the experimental approach.

This tool is intended to serve as a node prediction module for the automatic construction of redox regulatory network and also as an experimental candidate predictor for identifying thioredoxin/glutaredoxin target proteins.

2 Related works

The thioredoxin- and glutaredoxin-dependent regulatory networks are two redox systems in the cell sustaining the normal protein structure and function in an oxidizing environment. Both networks are composed of proteins with reversibly oxidized cysteine residues. According to the redox potential of each player in the network, electrons are transferred through specific donor-acceptor pairs [4-7]. Besides experimental strategies such as affinity chromatography and gel based methods [10-13], Marino and Gladyshev have adapted an integrative methodology in bioinformatics to detect thiol oxidoreductases and their catalytic redox-active cysteine residues [14]. Thioredoxins and glutaredoxins regulate various categories of proteins, such as proteins involved in photorespiration, citric acid cycle-associated reactions, lipid metabolism, electron transport, etc.[11,15]. Due to the functional diversity of the target proteins, simple sequence analysis usually fails in predicting novel target proteins. The parameters that characterize thiol-disulfide transition proteins are the vicinity of two cysteinyl residues, the pK_a value of the thiol, and the accessible surface area (ASA) of the cysteinyl residues on the target protein. Sanchez *et al.*[16] provided an algorithm based on the exact values of these three parameters, cysteine-cysteine distance, pK_a, and accessible surface area, for predicting the reversibly oxidized cysteines. Implementation of this approach as computational methodology for identification of thioredoxin or glutaredoxin target proteins is lacking.

Compared to redox regulatory networks, present day knowledge of metabolic networks is more advanced in the public domain. The physiological influence resulting from the activity change of the metabolic enzyme could be inferred by exploring the downstream reactions in the metabolic network. There are databases devoted to storing metabolic network information, such as KEGG [17], Reactome [18], MetaCyc [19], and Brenda [20]. Some web-based or stand-alone network visualization tools have been developed and are suitable for easy examination of the metabolic network [21-23]. The metabolic network content displayed by the visualization tools can be inputted by users or through the built-in network retrieval module which queries the metabolic network databases. Besides the network structure, visualization tools can depict the annotation for each node of enzyme and metabolite. Network visualization tools provide browsing, querying, editing, and analyzing functions for the metabolic network. Some visualization tools allow the third-party plug-in to interact with the host application, so that the functionality of such visualization tools may be extended.

Novel sequencing technologies have shortened the time needed to sequence complete genomes. However, subsequent to sequence acquisition, the genome must be annotated for its genes, proteins, and encoded biological pathway. Thus, the construction of biological pathways often relies on the quality of information stored in the database and the use of homology concept. Biological databases often store information only on general and well-known biological pathways. If the focus is placed on different and more specialized networks, the researcher must usually explore literature and molecular databases and construct the network from primary information. In addition to literature search, text mining tools are another choice. Construction of the network will be hindered if the necessary data is still lacking, or if the information from the database or text-mining tools proves to be too rich to handle, since reliable target selection may not be possible.

3 Methods

3.1 ROCD algorithm

ROCD follows the strategy of Sanchez *et al.* [16]. The algorithms implemented by ROCD are depicted in Fig. 2. Users are requested to input the file containing the SwissProt accession numbers, the criteria for the cysteine-cysteine distance, accessible surface area (ASA) and pK_a screening, and to define the file name to save the output. Users can also choose the desired tissue and organelle instead of providing the protein list for scanning tissue and organelle specific proteins. The generation of tissue and organelle specific protein set is relied on Human Protein Reference Database (HPRD) [29]. HPRD is a highly expert-curated human protein database.

Since the PDB file is essential for PropKa [26] and Naccess [25] prediction, the PDB ID is obtained by querying iProClass [27] database, and the yet structure-unsolved entry is queried against Swiss Model Repository. iProClass provides the ID mapping function which ROCD used to obtain PDB ID and the residing chains for each protein entry, and SWISS MODEL Repository [24] stores automated modeled protein structure by homology. To comply with the maximum 5000 atoms limitation from Naccess, the PDB file is tailored to contain the first model in the original PDB file.

The tailored PDB is used to calculate the distance between any pair of cysteines which are located on the chains specified in iProClass. If there is any calculated distance falling in the user-defined range, the protein entry and the qualified residues are written to the result file.

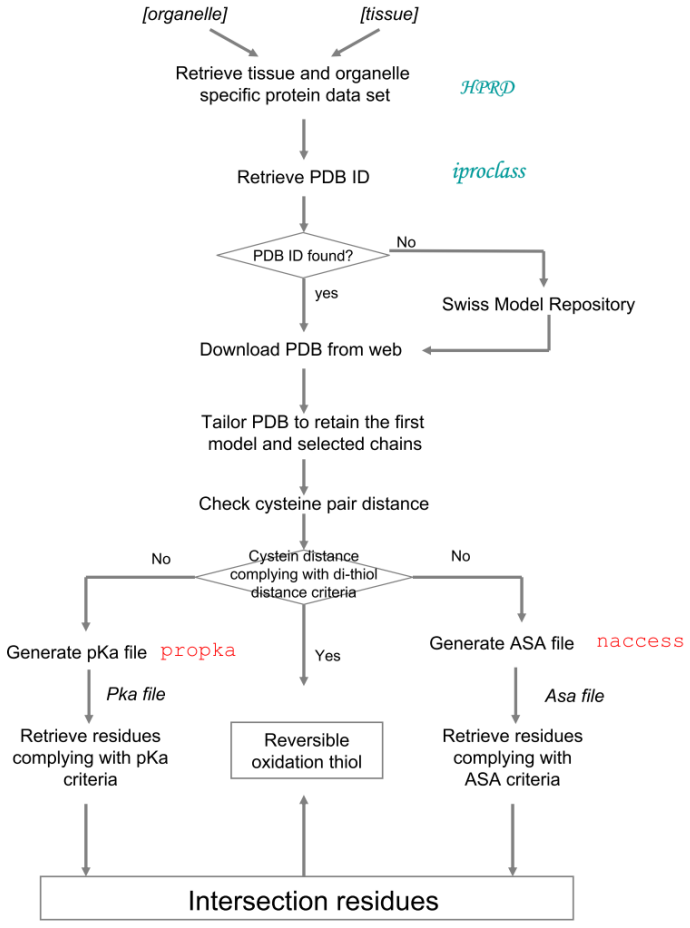

Fig. 2: The algorithm to predict protein with reversibly oxidized cysteinyl residues. The blue color denotes the database, and red denotes another standalone program.

If no cysteine-pair distance matches user's criteria, pK_a and ASA are calculated by PropKa and Naccess respectively with tailored PDB as the input. If the cysteine residue has the pK_a and ASA falling in the user-defined range and with the chain label as annotated in iProClass, the protein entry and the qualified residues are written to the result file.

During the execution of ROCD, the calculated cysteine-pair distance, pK_a, and ASA value for each cysteine residue are also written out to separate files for examination.

3.2 ROCD architecture

The architecture of ROCD is shown in Fig. 3. A command-line program and a web interface were created in this work. The program was implemented in Java, and the in-house HPRD and iProClass database [27] are stored in MySQL. The iProClass database is a collection of different molecular database accession numbers and makes the retrieval of corresponding molecular database accession numbers easier. The iProClass database flat file was downloaded from its website, and the data in each column of the flat file were parsed into different tables. In the web interface, users just provide: (i) Path to the file containing the SwissProt accession number, or to select tissue and organelle from the drop-down menu (ii) Criteria for the thiol-thiol distance, accessible surface area, and pK_a, (iii) file name to save the output files.

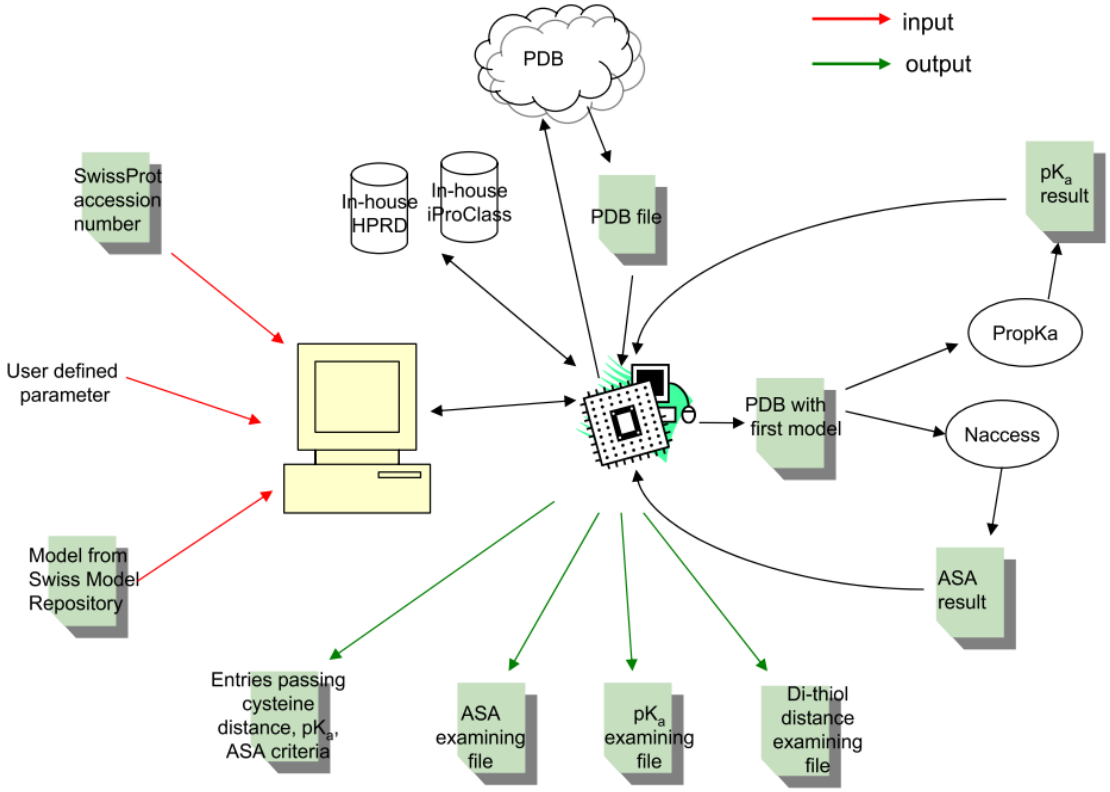

Fig. 3: The architecture of ROCD. Users are requested to provide three types of information to ROCD. ROCD utilizes three public databases (HPRD, iProClass, PDB) and two external tools (PropKa, Naccess) for the selection of reversibly oxidized cysteinyl residues.

In the back end, this program 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. Then the PDB file is downloaded from PDB database and is as the input for PropKa and Naccess. Finally, it generates four output files – one file for the distances of all cysteine pairs, one for all the calculated pK_a for cysteine residue, one for all the calculated ASA for the

SG atom, one for SwissProt accession numbers of potential proteins with reversibly oxidized cysteines.

3.3 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 undergo reversible oxidation and 161 cysteine thiols that are not susceptible to oxidation. Each PDB ID in BALOSCTdb is tested by ROCD. The parameters were chosen as specified in Sanchez *et al*.- 6.2 for cysteine-cysteine distance, 1.3 for ASA, and 9.05 for pK_a.

3.4 Examination of thioredoxin target proteins in plant mitochondrion

Balmer *et al.* collected 46 thioredoxin target proteins in plant mitochondrion. The target protein of thioredoxin should contain reversibly oxidized cysteines which form transient disulfide bond with thioredoxin during interaction. We use ROCD to test the existence of reversibly oxidized cysteines for these 46 proteins.

4 Results

ROCD implements the algorithms suggested by Sanchez *et al*. [16] for reversibly oxidized cysteine detection. ROCD identifies proteins which fulfill the three criteria provided by the user: cysteine-cysteine distance, pK_a range, and solvent accessible area range, in a list of proteins provided by the user or with specific tissue and organelle localization.

Tab. 1: Result from testing ROCD on BALOSCTdb

		Actual condition	
		Non reversibly oxidized	**Reversibly oxidized**
Prediction result	**Non reversibly oxidized**	137	61
	Reversibly oxidized	24	100

After the implementation, the compliance of our prediction was checked with BALOSCTdb (Table 1). Our prediction achieved 62.1% accuracy for the cysteine residues which are marked "reversibly oxidized cysteine" in BALOSCTdb and 85.1 % for "non-reversibly oxidized cysteines". We also applied ROCD on 46 thioredoxin target proteins in Balmer *et al.* [15] with the reported 3 parameter values: (i) cysteine-cysteine distance $\leq$ 6.2 Å, (ii) accessible solvent area $\geq$ 1.3 $Å^2$, (iii) $pK_a \leq 9.05$(Fig. 3). Only 3 proteins have corresponding PDB IDs, and 36 unresolved ones have homology entries in SWISS MODEL Repository, and 3 could be modeled manually by Swiss Model. The modeled PDB structure from a template with less than 50% similarity to the query sequence was abandoned due to low reliability. The remaining proteins were inspected by ROCD, and the recall rate was 66.7% (Table 2).