Prediction of reversibly oxidized protein cysteine thiols using protein structure properties

RICARDO SANCHEZ,¹ MEGAN RIDDLE,¹ JONGWOOK WOO,² AND JAMIL MOMAND¹

¹Department of Chemistry and Biochemistry, California State University, Los Angeles, California 90032, USA

²Department of Computer Information Systems, California State University, Los Angeles, California 90032, USA

(RECEIVED September 29, 2007; FINAL REVISION November 10, 2007; ACCEPTED November 15, 2007)

Abstract

Protein cysteine thiols can be divided into four groups based on their reactivities: those that form permanent structural disulfide bonds, those that coordinate with metals, those that remain in the reduced state, and those that are susceptible to reversible oxidation. Physicochemical parameters of oxidation-susceptible protein thiols were organized into a database named the Balanced Oxidation Susceptible Cysteine Thiol Database (BALOSCTdb). BALOSCTdb contains 161 cysteine thiols that undergo reversible oxidation and 161 cysteine thiols that are not susceptible to oxidation. Each cysteine was represented by a set of 12 parameters, one of which was a label (1/0) to indicate whether its thiol moiety is susceptible to oxidation. A computer program (the C4.5 decision tree classifier re-implemented as the J48 classifier) segregated cysteines into oxidation-susceptible and oxidation-non-susceptible classes. The classifier selected three parameters critical for prediction of thiol oxidation susceptibility: (1) distance to the nearest cysteine sulfur atom, (2) solvent accessibility, and (3) pKa. The classifier was optimized to correctly predict 136 of the 161 cysteine thiols susceptible to oxidation. Leave-one-out cross-validation analysis showed that the percent of correctly classified cysteines was 80.1% and that 16.1% of the oxidation-susceptible cysteine thiols were incorrectly classified. The algorithm developed from these parameters, named the Cysteine Oxidation Prediction Algorithm (COPA), is presented here. COPA prediction of oxidation-susceptible sites can be utilized to locate protein cysteines susceptible to redox-mediated regulation and identify possible enzyme catalytic sites with reactive cysteine thiols.

Keywords: cysteine; thiol; oxidation; prediction; C4.5; J48; redox; classifier; decision tree

Protein cysteine thiols can be divided into four broad categories: those that form permanent structural disulfide bonds, those that coordinate metals, those that are permanently in the reduced state, and those that are reversibly oxidized. Permanent structural disulfide bonds are

formed during the folding process by oxidizing enzymes (for example, DsbA in bacteria and protein disulfide isomerase in eukaryotes) (Kadokura et al. 2003; Maattanen et al. 2006). Permanent structural disulfide bonds are typically observed in oxidizing environments such as extracellular spaces and the endoplasmic reticulum. Protein cysteine thiols can be coordinated to metal ions, typically iron, copper, or zinc. Metal coordinated thiols are found in oxidizing environments and in the cytosolic compartment of the cell. The remaining protein cysteine thiols in the cytosol are either permanently reduced or are susceptible to reversible oxidation (Thomas et al. 1995). The reversibly oxidized protein thiols (ROPTs) in the cytosol are often required for enzyme catalysis or for regulation of protein activity (Finkel 2003; Linke and Jakob 2003).

Reprint requests to: Jamil Momand, Department of Chemistry and Biochemistry, California State University, 5151 State University Drive, Los Angeles, CA 90032, USA; e-mail: jmomand@calstatela.edu; fax: (323) 343-2499.

Abbreviations: BALOSCTdb, balanced oxidation susceptible cysteine thiol database; COPA, cysteine oxidation prediction algorithm; DOS, disk operating system; OSCTdb, oxidation susceptible cysteine thiol database; Prx, peroxiredoxin; PTEN, phosphatase and tensin homolog mutated in multiple advanced cancers 1; ROPT, reversibly oxidized protein thiol.

Article and publication are at <http://www.proteinscience.org/cgi/doi/10.1110/ps.073252408>.

Agents that mediate oxidation of ROPTs include hydroxyl radicals (Furuta et al. 2002), peroxides (Aslund et al. 1999; Fuangthong and Helmann 2002; Panmanee et al. 2002), S-nitrosoglutathione (Ji et al. 1999), and glutathione (Collison and Thomas 1987). Cysteine thiol oxidation frequently results in the formation of an intramolecular disulfide bond (S–S) with a second protein thiol. But in the absence of a nearby second cysteine, thiol oxidation typically leads to the formation of sulfenic acid (S–OH), sulfinic acid (SO₂H), S-nitrosothiols (S–NO), or S-glutathione (S–SG). Sulfenic acid formation is often an intermediate step in creating disulfide bonded cysteine thiols (Poole et al. 2004). In the vast majority of cases, ROPTs can be reduced back to thiols by thioredoxin, glutaredoxin, or glutathione (Ghezzi 2005; Berndt et al. 2007). However, sulfinic acids may have longer half-lives (Mallis et al. 2002; Woo et al. 2003; Wood et al. 2003). Since these protein thiol oxidations are often reversible they constitute a facile switch for modulating protein activity.

ROPTs in enzyme active sites undergo oxidation during normal catalysis. These active site protein thiols can become “overoxidized,” leading to inhibition of catalysis. For example, some peroxiredoxins (Prxs) possess two cysteine thiols that transiently oxidize to disulfides as they reduce hydrogen peroxide to water during catalysis (Kang et al. 2005). One class of Prxs (called 2-Cys Prxs) can become overoxidized and form sulfinic acid at one of its catalytic cysteines, which inhibits activity (Woo et al. 2003; Wood et al. 2003). Some active site thiols do not undergo oxidation during normal catalysis but are nonetheless susceptible to oxidation by non-substrate oxidants. For example, the PTEN tumor suppressor possesses a catalytic thiol. Treatment with hydrogen peroxide causes the thiol to form an intramolecular disulfide bond with a second cysteine thiol, resulting in inhibition of phosphatase activity (Lee et al. 2002). Critical cysteines at enzyme active sites can be oxidized, resulting in catalysis inhibition.

Oxidation of ROPTs in non-enzyme proteins can also alter protein function. The p21ras oncoprotein is susceptible to oxidation at Cys118, located near the guanine nucleotide binding site (Mallis et al. 2001). Peroxynitrite (derived from nitric oxide) and glutathione (a natural intracellular reductant) are responsible for S-glutathionylation of Cys118, which increases the ability of p21ras to exchange GTP for GDP (Clavreul et al. 2006). The S-glutathionylated p21ras exhibits higher association with its downstream partner Raf-1, and mediates signaling through the MEK/ERK pathway. In another example, treatment of bacteria with hydrogen peroxide produces an intramolecular disulfide bond in the OxyR transcription factor (Zheng et al. 1998; Lee et al. 2004). The oxidized form of OxyR binds DNA and transactivates genes responsible for detoxifying H₂O₂. Oxidation of non-enzyme proteins can alter protein function.

Oxidation of ROPTs can increase or decrease protein activity. However, compared to other types of post-translational modifications (i.e., phosphorylation, methylation, acetylation, etc.) relatively few oxidation-susceptible thiols have been identified. No algorithm has been developed that predicts oxidation susceptibility of ROPTs. On the other hand, there has been an extensive effort aimed at developing algorithms (CysPRED, DiANNA, DISULFIND, GDAP) to predict protein cysteines that create permanent structural disulfide linkages (Fiser and Simon 2000; Martelli et al. 2002; O'Connor and Yeates 2004; Ceroni et al. 2006; Ferre and Clote 2006). Oxidation susceptibility of ROPTs likely depends on the degree of solvent exposure and the identity and distance of atoms surrounding the target thiol. Elucidating the molecular architecture of ROPTs will aid in the identification of novel oxidation-susceptible cysteine thiols.

We present an algorithm, called the Cysteine Oxidation Prediction Algorithm (COPA), that predicts whether protein cysteines are susceptible to reversible oxidation. Our approach to building COPA was to use physicochemical parameters derived from structure data to predict protein cysteine oxidation susceptibility. In this study we present two contributions: (1) a curated database of protein cysteine thiols with known oxidation susceptibility and (2) an algorithm that utilizes structure data to predict, with 80.1% accuracy, whether protein cysteine thiols are susceptible to reversible oxidation.

Results

Oxidation susceptible cysteine thiol database (OSCTdb)

The goal was to create a database containing physicochemical parameters of protein cysteine residues with known susceptibilities to oxidation. We sought to obtain structure information on both oxidation-susceptible cysteines oxidation and non-susceptible cysteines. Figure 1 shows the structure of the PTEN tumor suppressor in the reduced state. Upon treatment with hydrogen peroxide the oxidation-susceptible Cys71 and Cys124 form a disulfide bond (Lee et al. 2002). The six other cysteines remain reduced after hydrogen peroxide treatment. We collected data on all of the cysteines in PTEN and other proteins that contain ROPTs and placed them in a database named the oxidation susceptible cysteine thiol database (OSCTdb).

Each cysteine (called the target cysteine) was labeled with a 0 or 1—where 0 marks the cysteines that are not oxidation susceptible and 1 marks the oxidation-susceptible cysteines. The following parameters were obtained for each target cysteine sulfur in OSCTb: (1) distance to the nearest cysteine sulfur (S1), (2) solvent-accessible surface area of S1, (3) name of nearest basic amino acid,

Figure 1. Backbone structure of PTEN tumor suppressor with cysteine residues in space-filling format. All eight cysteines are shown, two of which are susceptible to intramolecular disulfide bond formation (Cys71 and Cys124). The remaining six cysteines remain in the reduced form (Cys83, Cys105, Cys136, Cys211, Cys218, Cys250). Data is from PDB record 1D5R viewed by WebLab Viewer (Accelrys)

(4) distance to the nearest side chain nitrogen atom on a basic amino acid (N1), (5) solvent-accessible surface area of N1, (6) name of nearest acidic amino acid, (7) distance to nearest side chain oxygen on an acidic amino acid (O1), (8) solvent-accessible surface area of O1, (9) solvent-accessible surface area of target cysteine thiol, (10) calculated electrostatic potential of target cysteine thiol, and (11) calculated pKa of target cysteine thiol.

The OSCTdb contains 161 oxidation-susceptible cysteines, 301 oxidation-non-susceptible cysteines, and a total of 100 polypeptides. To achieve balance between oxidation-susceptible and oxidation-non-susceptible cysteine thiols, the number of oxidation-non-susceptible cysteines was lowered to 161 in a randomized fashion to create a balanced OSCTdb (BALOSCTdb). Algorithm development and statistical analysis were performed on data in BALOSCTdb.

Analysis of BALOSCTdb

Table 1 presents the number of cysteines derived from each protein class within the BALOSCTdb. Proteins from the oxidoreductase class contributed the highest number of cysteines ($n = 119$). This is not surprising because in building the database there was a concentrated effort to obtain ROPTs from oxidoreductases (see Materials and Methods). Contributions to BALOSCTdb from oxidoreductases were followed by hydrolases ($n = 62$) and transferases ($n = 54$). Non-enzyme proteins contributed 64 cysteines, followed by lyases ($n = 15$) and combination enzymes (hydrolases/oxidoreductases, $n = 4$) and isomerases ($n = 4$). In the non-enzyme protein class structure proteins constituted the largest group ($n = 15$) followed by RNA/DNA binding proteins ($n = 11$).

Selection of classifier

We wanted to select a computer program classifier that would use the physicochemical data from BALOSCTdb to create a set of rules that predicts thiol oxidation susceptibility. Waikato Environment for Knowledge Analysis (WEKA) is a collection of machine learning algorithms with a built-in statistical package (Witten and Frank 2005). From the WEKA package we chose the J48 decision tree classifier, which is a DOS-implemented form of the C4.5 classifier (Quinlan 1993, 1996). A decision tree classifier is a hierarchically based machine learning program that compares data (in our case, known oxidation susceptibility data) with a range of attributes (here, physicochemical parameters). The selection of attributes is determined from a statistical assessment of the ability to separate data into classes (oxidation-susceptible and oxidation-non-susceptible thiols). Data can be separated along discrete values assigned to the attributes and the data separation is represented as a decision tree. A decision tree consists of decision nodes, connected by branches, extending downward from the root node until terminating in leaf nodes (Larose 2005). The decision tree can serve as the basis for creating rules that will constitute a cysteine oxidation prediction algorithm.

Under default conditions, J48 gave the percent of correctly classified thiols of 76.7%, as assessed by leave-one-out (also known as jackknife) cross-validation. In leave-one-out cross-validation one cysteine thiol is removed from the data set and a decision tree is generated from the remaining data set. The removed cysteine thiol is judged for correct classification by the newly generated decision tree. This process is repeated for all thiols. The results of all cysteine thiol judgments, one for each member of the data set, are averaged, and the average represents the final error estimate. The size of the decision tree produced by J48 was 44 decision nodes. We next concentrated on optimizing the J48 classifier.

Table 1. The protein class, number of cysteine thiols, number of correctly predicted cysteine thiols, and percent of correctly predicted cysteine thiols in the Balanced Oxidation Susceptible Cysteine Thiol Database

Protein class	Number of cysteines	Number correctly predicted	Percent correctly classified
Oxidoreductase	119	104	87.4
Hydrolase	62	46	74.2
Transferase	54	42	77.8
Lyase	15	10	66.7
Hydrolase/Oxidoreductase	4	4	100.0
Isomerase	4	3	75.0
Non-enzyme proteins	64	50	78.1
Totals	322	259	80.4

Optimization of *M* value setting on the J48 classifier

J48 produced a high percentage of correctly classified reactive cysteine thiols, but the number of decision nodes was high. This was due to the fact that, under default conditions, the threshold for the number of thiols that could form a leaf node was only 2. This threshold is called the *M* value (in this case, M2 for 2 thiols). Information from the decision tree will be used to create rules for an algorithm that classifies cysteine thiols into either oxidation-susceptible or oxidation-non-susceptible classes. Rules created from M2 run the risk of not being sufficiently general to successfully classify data outside of BALOSCTdb. To minimize this possibility the *M* value was increased so that rules derived from the decision tree could generate an algorithm that could be applied to a wide data set. Figure 2 shows that, as the *M* value was incrementally increased from 2 to 70, the percent of correctly classified thiols fluctuated from 72.1% to 80.1%. The highest percentage of correctly classified thiols was achieved in the M49–M52 range, where 80.1% was maintained. In this range, the tree size was only 7, indicating that there were relatively few decision nodes in the decision tree. Overall, as the *M* value was increased from M2 to M70, the tree size was lowered from 43 to 5.

We concentrated on the M49–M52 range since the highest percentage of correctly classified thiols was achieved in this range and the tree size was relatively small. The decision trees generated from M19 through M51 were identical, indicating that rules derived from these trees would likely be applicable over a wide data set. The tree generated from M52 was slightly different due to a small change in the value assigned to target thiol

Figure 2. Dependence of percent of correctly classified reactive cysteine thiols and decision tree size on *M* value. The *M* value of the J48 classifier was varied from 2 to 70 and allowed to segregate the reactive cysteine thiols into oxidation-susceptible and oxidation-non-susceptible classes. Percent correctly classified cysteine thiols were generated by leave-one-out cross-validation (filled squares). The tree size generated at each *M* value was plotted (filled circles).

calculated pKa (in M52 this value was 9 and in M19–M51 this value was 9.05). The percent of correctly classified thiols in the M19–M51-constrained J48 classifier ranged from 75.5% to 80.1%. It was somewhat surprising that the M19–M51 gave such a percentage range given that the same decision tree was produced. In the leave-one-out cross-validation calculation a new decision tree was created 322 times, one for each thiol. At the low end of the *M* value range a few decision trees were generated that did not match the consensus final decision tree output. These “rogue” decision trees resulted in lower percentages of correctly classified thiols. As the *M* values increased, fewer rogue decision trees were created because the threshold number of thiols for each leaf was large. Thus M49–M52 elicited the highest percent of correctly classified thiols.

To verify that the decision trees generated from J48 were robust, a Receiver Operating Characteristic (ROC) curve was generated. In this curve, the true positive rate is plotted against the false positive rate. The longer the distance of the curve from the 45 degree diagonal that divides ROC space the more accurate the test. Figure 3 shows a comparison of ROC curves generated from each *M*-value-constrained outcome used to classify the oxygen-susceptible thiols. *M* values that produced identical percentages of correctly classified cysteine thiols were plotted as one curve. The curve generated when the *M* value range was M49–M52 was one of the furthest from the diagonal. The M49–M52 constrained J48 outcome resulted in a true positive rate of 0.839 and a false positive rate of 0.236. One other curve was far from the diagonal. This ROC curve was generated by the M59–M60 constrained J48. It produced a true positive rate (TP = 0.789) that was 6% lower than the true positive rate generated from M49–M52, but it also produced a false positive rate (0.199) that was 16% lower.

To decide whether to use the decision tree generated from the M49–M52 constrained J48 or from the M59–M60 constrained J48 we determined the true positive and false positive rates generated for classification of oxidation-non-susceptible thiols. For M49–M52 the true positive rate (TP = 0.764) was 8.1% lower than that of M59–M60. On the other hand, the false negative rate (FP = 0.161) was 24% lower than that of M59–M60. Overall, it appears that M49–M52 has a significant advantage in the false negative rate when classifying oxidation-non-susceptible cysteines. The Kappa statistic is an index that compares the classification outcome against that which might be expected by chance. It can range from –1 (complete disagreement) to +1 (complete agreement). The Kappa statistic generated from the M49–M52 constrained J48 (0.6025) was slightly higher than the Kappa statistic from the M59–M60 constrained J48 (0.5901).

Figure 3. ROC curves generated by leave-one-out analysis of J48-mediated segregated data in BALOSCTdb. Each curve was generated from a different M value setting. M values ranged from 2 to 70, and data were segregated into oxidation-susceptible and oxidation-non-susceptible classes. The true positive and false positive data were obtained by comparing the predicted oxidation-susceptible cysteines to the known oxidation-susceptible cysteines. Large filled orange circle denotes one point on the curve generated by M49–M52 constrained J48 classifier output. Diagonal line was placed in the plot for reference.

Thus, the algorithm from the M49–M52 constrained J48 exhibited a slightly higher percentage of correctly classified cysteine thiols (80.1% vs. 79.5%), had a lower FP rate in the classification of non-oxidation-susceptible cysteines, and a higher Kappa statistic than the M59–M60 constrained classifier. The decision trees generated from M49–M51 were identical to the decision trees generated from M19 through M48. For these reasons we used the decision tree from M49–M51 constrained J48 to develop a cysteine oxidation prediction algorithm.

Analysis of output from COPA

Rules derived from the M49–M51 decision tree were converted into a computer algorithm (Cysteine Oxidation Prediction Algorithm—COPA) that segregates protein cysteines into oxidation-susceptible and oxidation-non-susceptible classes. When this algorithm was applied to cysteines in the BALOSCTdb, it was able to properly classify the cysteines in 80.4% of the cases. (This differs slightly from the 80.1% accuracy derived from leave-one-out analysis because we are not removing any data for the calculation.) Table 1 shows the percent of correctly classified cysteines for each protein class in the database. The percentages of correctly classified cysteines for each protein class were: oxidoreductase, 87.4%; non-enzyme proteins, 78.1%; transferase, 77.8%; hydrolase, 74.2%; lyase, 66.7%. Within the hydrolase/oxidoreductase class and isomerase class the percentages were 100% and 75%,

respectively, but there were only four cysteines of each class represented in the BALOSCTdb.

The results demonstrate that COPA has some success at predicting oxidation susceptibility across different enzyme classes with the exception of lyase, where the accuracy was relatively low. It is possible that the molecular architecture surrounding the thiols in the lyase class is substantially different from that found in the other classes. It is also possible that the low number of thiols (only 15 cysteines from four protein chains) in the BALOSCTdb contributed to the low prediction accuracy. If the number of thiols from lyases was increased, J48 may be capable of generating a decision tree that would better predict the oxidation susceptibility of lyase cysteine thiols.

We next wanted to know if the accuracy of COPA prediction depended on the type of oxidizing ligand bound to the sulfur atom. We determined the percent of correctly classified thiols for each type of ligand bonded cysteine sulfur atom in BALOSCTdb. Table 2 shows that COPA performed best when predicting whether a cysteine thiol would form an intramolecular disulfide bridge (93.9%) and performed poorest when predicting whether the cysteine thiol would form a sulfenic acid (64.7%). The algorithm was able to predict when the cysteines were oxidized to sulfinic acid in 76.5% of the cases. Other ligands bound to the cysteine thiol were correctly predicted in the range of 33%–100%, but there were too few instances of each type of oxidizing ligand to draw

Table 2. Percent correctly classified cysteine state in relation to the ligand bound to the cysteine sulfur atom

Cysteine sulfur state ^a	Number of cysteines	Correctly classified (%)
SH	161	76.4
SS, +SOH (1), +SOOH (1), +SOOOH (2)	98	93.9
SOH	34	64.7
SOOH	17	76.5
SSG	3	66.7
SOOOH	3	100
Intermol SS	3	33.3
SNO, Intermol SS	1	100
Intermol SS, SSG, SOOH	1	100
Unknown	1	100
Total	322	80.4

^a SS is intramolecular cysteine disulfide. +SOH (1) signifies one case where a cysteine oxidizes to SS or SOH. +SOOH (1) signifies one case where a cysteine has been shown to oxidize to SS or SOOH. +SOOOH (2) signifies two cases where a cysteine oxidizes to either SS or SOOOH. SSG is S-glutathionylated cysteine. SOOOH is sulfonic acid. Intermol SS is intermolecular protein cysteine disulfide. SNO, Intermol SS is where the cysteine sulfur forms a covalent bond with NO or forms an intermolecular protein cysteine disulfide. Intermol SS, SSG, SOOH is where the cysteine sulfur forms an intermolecular protein disulfide, or is glutathionylated or forms sulfinic acid.

strong conclusions on the robustness of COPA in their outcome.

Analysis of the algorithm COPA generated by the decision tree produced by the J48 classifier

Analysis of the algorithm can be used to increase our understanding of the physicochemical parameters that increase the susceptibility of cysteine thiol groups to reversible oxidation. There are three rules that guide COPA (Fig. 4). One rule applies to the distance between the two thiols that form a disulfide. If the distance is ≤ 6.2 Å, there is a high probability that a disulfide will form. In our data set, 99 cysteines were correctly classified and 14 cysteines were incorrectly classified by this rule.

A second rule applies when the two cysteine thiols are >6.2 Å apart. If this is the case, and if the solvent-accessible area of the target cysteine thiol is <1.3 Å², there is a high probability that the target thiol is not susceptible to oxidation. In our data set 95 cysteines were correctly classified and eight cysteines were incorrectly classified by this rule.

The third rule applies when the two cysteines are >6.2 Å apart and the solvent accessible area of the target cysteine thiol is >1.3 Å². If the calculated pKa of the target thiol is less than or equal to 9.05 there is a moderately high probability that the thiol is susceptible to oxidation. In our data set 76 cysteines were correctly classified and 24 cysteines were incorrectly classified by this rule. There is a corollary to this rule. If the calculated pKa of the target thiol is greater than 9.05 there is a moderately high probability that the thiol is not susceptible to oxidation. In our data set, 52 cysteines were correctly classified and 16 cysteines were incorrectly classified by this corollary.

Figure 4. Cysteine Oxidation Prediction Algorithm created from the decision tree produced by the J48 classifier from data in BALOSCTdb.

One of the strongest predictors of cysteine oxidation susceptibility is distance of the thiol to another cysteine thiol. When the distance is ≤ 6.2 Å there is a high probability that the cysteine will become oxidized. It is likely that the oxidation product for cysteines that follow this rule would be cysteine disulfide. To determine if this is the case, we divided the oxidation-susceptible cysteines in BALOSCTdb into two groups. In one group were those that form cysteine disulfides and in the other group those that formed all other types of oxidizing ligands. We found that this rule correctly predicted 81% of the cysteines in the disulfide class. On the other hand, of the cysteines that do not create cysteine disulfides, only 3% were predicted by this rule to undergo oxidation. Distance between thiols strongly predicts intramolecular disulfide bond formation.

Another predictor of cysteine oxidation susceptibility is low calculated pKa value coupled with solvent accessibility. When the target cysteine thiol lies >6.2 Å from a second cysteine thiol, has a solvent accessible area of >1.32 Å², and possesses a calculated pKa lower than 9.05, there is a high probability that the cysteine will undergo oxidation. One would expect that cysteines that do not form cysteine disulfides would be affected by this rule. We observed that 66% of the cysteines that do not undergo intramolecular disulfide bond formation follow this rule. On the other hand, only 3% of the cysteines that create intramolecular disulfide bonds formation follow this rule. This indicates that there are two general classes of ROPTs, those that form cysteine disulfide bonds and those that do not, and although both classes are reversibly oxidized, different physicochemical parameters can predict their oxidation. Knowledge of these rules can help predict the type of oxidation reaction that can occur.

Discussion

A database, named BALOSCTdb, containing 161 oxidation-susceptible and 161 oxidation-non-susceptible protein cysteine thiols was created. The database contains 12 physicochemical parameters associated with each thiol. The J48 classifier used three of these parameters to develop a decision tree that predicts oxidation-susceptible and oxidation-non-susceptible cysteines with an accuracy of 80.1% according to leave-one-out cross-validation analysis. The three parameters that dictate oxidation susceptibility are distance between the target thiol and a second cysteine thiol, solvent accessibility of the target thiol, and target thiol calculated pKa. For the first time, we report the quantitative values for these three parameters that can be used to predict oxidation susceptibility of reversibly oxidized cysteine residues in proteins.

This study shows that two general classes of reversibly oxidized protein thiols can be predicted on the basis of

different parameters—those that form intramolecular disulfides and those that do not. When cysteines form intramolecular disulfides the overriding parameter is the distance to a second cysteine thiol. If that distance is 6.2 Å or less there is a high probability that disulfide bond formation will occur. This means that the range that predicts disulfide bond formation is typically between 3.0 Å and 6.2 Å since at a distance <3.0 Å a permanent disulfide bond is present. In the reduced human mitochondrial branched chain aminotransferase isoenzyme (hBCATm) where the thiol-to-thiol distance was measured to be ~3.1–3.5 Å, a hydrogen bond is created between a thiolate and thiol (Conway et al. 2002). This arrangement leads to increased propensity for oxidation and disulfide formation upon treatment with hydrogen peroxide. Thioredoxin and glutaredoxin are other examples containing ROPTs where there is a short distance between thiols. Our cysteine oxidation prediction algorithm was able to predict oxidation susceptibility of the disulfide bonding thiols in thioredoxin, glutaredoxin, and hBCATm.

There is an emerging group of proteins that form disulfides, but, in their reduced states, their thiols are not near each other. Some members of this group form intramolecular disulfides and others form intermolecular disulfides. OxyR forms an intramolecular disulfide, and, in its reduced state, the thiols are 17 Å apart (Choi et al. 2001). Our algorithm was able to correctly predict oxidation susceptibility of the two thiols in OxyR because their calculated pKas were below 9.05 (cys199 was 7.18 and cys208 was 8.3) and their solvent accessible surface areas were >1.3 Å² (cys199 was 2 Å² and cys208 was 68 Å²). It will be necessary to gather many examples of proteins that produce intermolecular disulfides by thiol travel over long distances to determine if these rules from COPA apply over a larger data set.

COPA was generally unsuccessful in predicting oxidation susceptibility of thiols that form intermolecular disulfides. COPA only predicted disulfide bond formation for 66.7% of the intermolecular disulfide bonders (see Table 2). Since there were only five of these thiols in our data set, it will be important to obtain more examples to derive rules to predict their oxidation susceptibilities. A case in point is 2-Cys peroxiredoxin, which possesses two thiols on each subunit that make intermolecular disulfide bonds with an identical subunit. Our algorithm failed to predict the oxidation-susceptible cysteine thiols in this enzyme (PDB record 2I81) (J.D. Artz, W. Qiu, A. Dong, J. Lew, H. Ren, Y. Zhao, I. Kozieradski, A.M. Edwards, C.H. Arrowsmith, J. Weigelt, et al., unpubl. data). For the disulfide bond-forming thiols the calculated pKa values were lower than 9.05 but the solvent accessibility areas were too low (0 Å² for one thiol and 0.4 Å² for the other thiol). This result may suggest that, for intermolecular

disulfide bonders, other non-thiol regions of two polypeptides first interact to create a conformation change that exposes the thiols to solvent. After exposure the thiols may then form disulfide bonds. COPA also had difficulty predicting oxidation susceptibility of lyase thiols (percent correctly predicted 66.7%, *n* = 15). Because we lacked sufficient data for intermolecular disulfide bonders and lyases, the rules derived from the decision tree generated misclassified their thiols without lowering overall accuracy. More representatives of ROPTs from these proteins in our database may alleviate this problem.

In our second class of ROPTs, where disulfide bonds are not formed, we found that a low calculated pKa and solvent accessibility are necessary characteristics for oxidation susceptibility. The fact that the pKa of the cysteine thiol is a useful parameter for predicting oxidation susceptibility is consistent with other studies. Dixon's group experimentally measured a pKa of 4.7 for the active site cysteine in a protein tyrosine phosphatase due to the presence of the positively charged local environment (Zhang and Dixon 1993). However, COPA was only successful in predicting 66% of the oxidation-susceptible thiols that do not form disulfides. It was somewhat surprising that our threshold for triggering oxidation susceptibility was a calculated pKa of 9.05—higher than expected since the experimentally derived pKa of protein cysteine thiols in aqueous environments is ~8.3. This higher-than-expected calculated pKa may be the result of a discrepancy between the pKa calculations and experimentally determined pKa values. In another study, the pKa calculation program we used (PROPKA) was tested on 12 protein cysteine thiols with known experimentally determined pKa values (Li et al. 2005). The calculated pKa values deviated up to 1.5 pH units away (in both directions up and down) from experimentally derived values. This suggests that predictions of ROPTs in non-disulfide bond forming proteins may improve once pKa calculations become more accurate. However, it must be noted that, when compared to three other pKa computing programs, PROPKA compared favorably (Davies et al. 2006).

In our second class of ROPTs we discovered that solvent accessibility is necessary for oxidation. The threshold for oxidation is 1.3 Å², indicating that perhaps only a small surface of the thiol is necessary for proton removal and thiolate generation. Thomas found that the overriding characteristic of cysteines susceptible to S-glutathionylation is solvent accessibility (Thomas et al. 1995). We note that one type of ROPT was not addressed in this study. It has been demonstrated that metal-coordinated cysteines can release metal ions and form disulfide bonds (Zdanowski et al. 2006). Metal bound cysteines were excluded from our database because these are usually permanent structures.

COPA is designed to predict oxidation susceptibility of cysteine thiols in proteins with known structures. COPA has the potential to be useful because there will continue to be an increase in structure data in the foreseeable future. The rate of yearly depositions in the PDB between 2002 and 2006 has climbed from 3381 to 6911 (Research Collaboratory for Structural Bioinformatics 2003; 2007) and there is a push to solve novel protein structures. This study is the first effort to systematically measure and identify the physicochemical parameters that predict reversible cysteine oxidation susceptibility in proteins.

Materials and Methods

Construction of OSCTdb

Construction of OSCTdb was accomplished in two phases. In the first phase, a literature search of PubMed was conducted to extract papers that describe proteins with ROPTs. Cysteines that participate in oxidation reactions during catalysis were included in the OSCTdb. Cysteines that normally are reduced but undergo oxidation as result of exposure to oxidizing condition were included in the OSCTdb. Cysteines that remain permanently reduced were included in the database. Cysteines that form permanent disulfide bonds, coordinated to metals, or bound to organic cofactors were excluded from the database.

Oxidation susceptibilities were obtained from the literature. It should be noted that the literature often only describes the oxidation susceptibility of one or two “target” cysteines and the oxidation susceptibilities of other cysteine residues on the protein are not evaluated. Unless data from the literature indicated otherwise, it was assumed that nontarget cysteines were permanently reduced.

In the second phase of building OSCTdb, we concentrated on oxidoreductases in the PDB because there are many examples where cysteine thiols are located in active sites of these enzymes and undergo oxidation during catalysis. We wanted to capture PDB records that had oxidation susceptible cysteines. Distances between intramolecular cysteine thiols in all oxidoreductases were calculated. If distances were $<3.0 \text{ \AA}$ (indicating a disulfide bond), the protein sequences were obtained from the PDB records. Using BLAST, these protein sequences were used to capture proteins from oxidoreductase PDB records with identical sequences but in reduced form. The goal was to find the PDB records that were the reduced counterparts to the records containing disulfides. If the S–S distances of PDB records of the reduced counterparts were $>3.0 \text{ \AA}$ (indicating no disulfide bond), the records were analyzed further to ensure that no metal ions or organic cofactors were bonded to the cysteine thiols. If these criteria were met, parameters from the PDB records with the long S–S distances were added to the OSCTdb.

The majority of PDB records contain more than one polypeptide chain and each chain is assigned an alphabetical label starting with “A.” Parameters for inclusion into the OSCTdb were extracted from chain A with five exceptions. In these five exceptions, heterocomplexes of polypeptide chains were observed in the records and the oxidation-susceptible peptide was not the A chain. In these records, parameters were extracted from the oxidation-susceptible chain. In some PDB records we found that crystallographers replaced the oxidation-susceptible

cysteines with serines (to prevent oxidation during crystallization). For these records the oxygen atoms of the Ser side chains were replaced with a sulfur atom and parameters on these “pseudo cysteines” were extracted.

The BLASTALL feature (www.ncbi.nlm.nih.gov/staff/tao/URLAPI/) within the BLAST suite of software programs was used to preclude overrepresentation of similar sequences in the OSCTdb (Altschul et al. 1990). Duplicate PDB records were removed from the OSCTdb if the aligned local sequences containing cysteines had E values <1 and identities $>35\%$.

Name and distance parameters were obtained from the PDB records with a parameter extraction computer program we created. The Naccess software program (Lee and Richards 1971), with default settings, was utilized to calculate the water-accessible surface areas. Electrostatic potentials of cysteine thiols were calculated with the macroscopic electrostatics with atomic detail (MEAD) package (Bashford 2004) implemented by the protein continuum electrostatics (PCE) online service (Miteva et al. 2005). Cysteine thiol pKa values were calculated with protein pKa prediction software program, PROPKA (Li et al. 2005). PROPKA uses the following equation for cysteine thiol pKa determination from PDB data:

$$\text{pKa} = \text{pK}_{\text{Model}} + \Delta\text{pKa}$$

where the pK_{Model} for cys is 9.00 and the ΔpKa is based on hydrogen bonding, desolvation effects, and charge–charge interactions.

Electronic supplemental material

The OSCTdb and BALOSCTdb are implemented in HTML format. Aside from the parameters described in the text, the following data are included in these databases: (1) PDB ID, PDB record number; (2) CHAIN, protein chain; (3) SEQUENCE POSITION, cysteine amino acid sequence number; (4) SEQUENCE, partial amino acid sequence displaying five amino acids upstream and five amino acids downstream of target cysteine; (5) S1 POSITION, the amino acid sequence number of the cysteine nearest to the target thiol; (6) N1 ID, the nitrogen atom on the nearest basic amino acid; (7) N1 POSITION, the amino acid sequence number of the basic amino acid nearest to the target thiol; (8) O1 ID, the oxygen atom on the nearest acidic amino acid; (9) O1 POSITION, the amino acid sequence number of the acid amino acid nearest to the target thiol; (10) REF STRUCTURE, the PDB record of the protein in the oxidized state or the literature reference for the oxidation-susceptible cysteine; (11) PUBMED ID1, identification number of literature citation in PubMed; (12) PUBMED ID2, identification number of a second literature citation in PubMed; (13) MOLECULE, name of protein; (14) OXIDANT, name of oxidant (if known); (15) LIGAND, status of the cysteine thiol after oxidation (see Table 2); (16) SOURCE, species source of protein sequence; (17) CLASSIFICATION, enzyme class name; (18) EC #, enzyme class number.

Supplemental material pages: 1 webpage (http://www.calstatela.edu/faculty/jmomand/OSCTdb_BALOSCTdb.htm?Submit24=Oxidation+Susceptibility+Cysteine+Thiol+Database).

Description of Supplemental material pages: An online database of parameters associated with reversibly oxidized protein cysteine thiols and cysteine thiols that are not oxidation susceptible.

Acknowledgments

We gratefully acknowledge funding from the following: NIH (1-R-15-CA101757-01), NSF-NIH (EEC-0609454), and the NIH Biomed Prep Program (GM 064105).

References

- Altschul, S.F., Gish, W., Miller, W., Myers, E.W., and Lipman, D.J. 1990. Basic local alignment search tool. *J. Mol. Biol.* **215**: 403–410.
- Aslund, F., Zheng, M., Beckwith, J., and Storz, G. 1999. Regulation of the OxyR transcription factor by hydrogen peroxide and the cellular thiol-disulfide status. *Proc. Natl. Acad. Sci.* **96**: 6161–6165.
- Bashford, D. 2004. Macroscopic electrostatic models for protonation states in proteins. *Front. Biosci.* **9**: 1082–1099.
- Berndt, C., Lillig II, C., and Holmgren, A. 2006. Thiol-based mechanisms of the thioredoxin and glutaredoxin systems: Implications for diseases in the cardiovascular system. *Am. J. Physiol. Heart Circ. Physiol.* **292**: H1227–H1136. doi: 10.1152/ajpheart.01162.2006.
- Ceroni, A., Passerini, A., Vullo, A., and Frasconi, P. 2006. DISULFIND: A disulfide bonding state and cysteine connectivity prediction server. *Nucleic Acids Res.* **34**: W177–W181. doi: 10.1093/nar/gkl266.
- Choi, H., Kim, S., Mukhopadhyay, P., Cho, S., Woo, J., Storz, G., and Ryu, S. 2001. Structural basis of redox switch in OxyR transcription factor. *Cell* **105**: 103–113.
- Clavreul, N., Adachi, T., Pimental, D.R., Ido, Y., Schoneich, C., and Cohen, R.A. 2006. S-glutathiolation by peroxynitrite of p21ras at cysteine-118 mediates its direct activation and downstream signaling in endothelial cells. *FASEB J.* **20**: 518–520.
- Collison, M.W. and Thomas, J.A. 1987. S-thiolation of cytoplasmic cardiac creatine kinase in heart cells treated with diamide. *Biochim. Biophys. Acta* **928**: 121–129.
- Conway, M.E., Yennawar, N., Wallin, R., Poole, L.B., and Hutson, S.M. 2002. Identification of a peroxide-sensitive redox switch at the CXXC motif in the human mitochondrial branched chain aminotransferase. *Biochemistry* **41**: 9070–9078.
- Davies, M.N., Toseland, C.P., Moss, D.S., and Flower, D.R. 2006. Benchmarking pK_a prediction. *BMC Biochem.* **7**: 18. doi: 10.1186/1471-2091-7-18.
- Ferre, F. and Clote, P. 2006. DiANNA 1.1: An extension of the DiANNA web server for ternary cysteine classification. *Nucleic Acids Res.* **34**: W182–W185. doi: 10.1093/nar/gkl189.
- Finkel, T. 2003. Oxidant signals and oxidative stress. *Curr. Opin. Cell Biol.* **15**: 247–254.
- Fiser, A. and Simon, I. 2000. Predicting the oxidation state of cysteines by multiple sequence alignment. *Bioinformatics* **16**: 251–256.
- Fuangthong, M. and Helmann, J.D. 2002. The OhrR repressor senses organic hydroperoxides by reversible formation of a cysteine-sulfenic acid derivative. *Proc. Natl. Acad. Sci.* **99**: 6690–6695.
- Furuta, S., Ortiz, F., Sun, X.Z., Wu, H.H., Mason, A., and Momand, J. 2002. Copper uptake is required for pyrrolidine dithiocarbamate-mediated oxidation and protein level increase of p53 in cells. *Biochem. J.* **365**: 639–648.
- Ghezzi, P. 2005. Oxidoreduction of protein thiols in redox regulation. *Biochem. Soc. Trans.* **33**: 1378–1381.
- Ji, Y., Akerboom, T.P., Sies, H., and Thomas, J.A. 1999. S-nitrosylation and S-glutathiolation of protein sulfhydryls by S-nitroso glutathione. *Arch. Biochem. Biophys.* **362**: 67–78.
- Kadokura, H., Katzen, F., and Beckwith, J. 2003. Protein disulfide bond formation in prokaryotes. *Annu. Rev. Biochem.* **72**: 111–135.
- Kang, S.W., Rhee, S.G., Chang, T.S., Jeong, W., and Choi, M.H. 2005. 2-Cys peroxiredoxin function in intracellular signal transduction: Therapeutic implications. *Trends Mol. Med.* **11**: 571–578.
- Larose, D.T. 2005. *Discovering knowledge in data: An introduction to data mining*. John Wiley & Sons, Hoboken, NJ.
- Lee, B. and Richards, F.M. 1971. The interpretation of protein structures: Estimation of static accessibility. *J. Mol. Biol.* **55**: 379–400.
- Lee, S.R., Yang, K.S., Kwon, J., Lee, C., Jeong, W., and Rhee, S.G. 2002. Reversible inactivation of the tumor suppressor PTEN by H₂O₂. *J. Biol. Chem.* **277**: 20336–20342.
- Lee, C., Lee, S.M., Mukhopadhyay, P., Kim, S.J., Lee, S.C., Ahn, W.S., Yu, M.H., Storz, G., and Ryu, S.E. 2004. Redox regulation of OxyR requires specific disulfide bond formation involving a rapid kinetic reaction path. *Nat. Struct. Mol. Biol.* **11**: 571–578.
- Li, H., Robertson, A.D., and Jensen, J.H. 2005. Very fast empirical prediction and rationalization of protein pK_a values. *Proteins* **61**: 704–721.
- Linke, K. and Jakob, U. 2003. Not every disulfide lasts forever: Disulfide bond formation as a redox switch. *Antioxid. Redox Signal.* **5**: 425–434.
- Maattanen, P., Kozlov, G., Gehring, K., and Thomas, D.Y. 2006. ERp57 and PDI: Multifunctional protein disulfide isomerases with similar domain architectures but differing substrate-partner associations. *Biochem. Cell Biol.* **84**: 881–889.
- Mallis, R.J., Buss, J.E., and Thomas, J.A. 2001. Oxidative modification of H-ras: S-thiolation and S-nitrosylation of reactive cysteines. *Biochem. J.* **355**: 145–153.
- Mallis, R.J., Hamann, M.J., Zhao, W., Zhang, T., Hendrich, S., and Thomas, J.A. 2002. Irreversible thiol oxidation in carbonic anhydrase III: Protection by S-glutathiolation and detection in aging rats. *Biol. Chem.* **383**: 649–662.
- Martelli, P.L., Fariselli, P., Malaguti, L., and Casadio, R. 2002. Prediction of the disulfide-bonding state of cysteines in proteins at 88% accuracy. *Protein Sci.* **11**: 2735–2739.
- Miteva, M.A., Tufféry, P., and Villoutreix, B.O. 2005. PCE: Web tools to compute protein continuum electrostatics. *Nucleic Acids Res.* **33**: W372–W375. doi: 10.1093/nar/gki365.
- O'Connor, B.D. and Yeates, T.O. 2004. GDAP: A web tool for genome-wide protein disulfide bond prediction. *Nucleic Acids Res.* **32**: W360–W364. doi: 10.1093/nar/gkh376.
- Panmanee, W., Vattanaviboon, P., Eiamphungporn, W., Whangsuk, W., Sallabhan, R., and Mongkolsuk, S. 2002. OhrR, a transcription repressor that senses and responds to changes in organic peroxide levels in *Xanthomonas campestris* pv. *phaseoli*. *Mol. Microbiol.* **45**: 1647–1654.
- Poole, L.B., Karplus, P.A., and Claiborne, A. 2004. Protein sulfenic acids in redox signaling. *Annu. Rev. Pharmacol. Toxicol.* **44**: 325–347.
- Quinlan, J.R. 1993. *C4.5: Programs for machine learning*. Morgan Kaufmann Publishers, San Francisco, CA.
- Quinlan, J.R. 1996. Improved use of continuous attributes in c4.5. *J. Artif. Intell. Res.* **4**: 77–90.
- Research Collaboratory for Structural Bioinformatics. 2003. RCSB PDB Newsletter. http://www.resb.org/pdb/static.do?p=general_information/news_publications/newsletters/2002q4/index.html
- Research Collaboratory for Structural Bioinformatics. 2007. RCSB PDB Newsletter. http://www.rcsb.org/pdb/general_information/news_publications/newsletters/2006q4/deposition.html#statistics
- Thomas, J.A., Poland, B., and Honzatko, R. 1995. Protein sulfhydryls and their role in the antioxidant function of protein S-thiolation. *Arch. Biochem. Biophys.* **319**: 1–9.
- Witten, I.A. and Frank, E. 2005. *Data mining: Practical machine learning tools and techniques*, 2nd ed. Morgan Kaufmann, San Francisco, CA.
- Woo, H.A., Chae, H.Z., Hwang, S.C., Yang, K.S., Kang, S.W., Kim, K., and Rhee, S.G. 2003. Reversing the inactivation of peroxiredoxins caused by cysteine sulfenic acid formation. *Science* **300**: 653–656.
- Wood, Z.A., Poole, L.B., and Karplus, P.A. 2003. Peroxiredoxin evolution and the regulation of hydrogen peroxide signaling. *Science* **300**: 650–653.
- Zdanowski, K., Dougherty, P., Jakimowicz, P., O'Hara, L., Buttner, M.J., Paget, M.S., and Kleanthous, C. 2006. Assignment of the zinc ligands in RsrA, a redox-sensing ZAS protein from *Streptomyces coelicolor*. *Biochemistry* **45**: 8294–8300.
- Zhang, Z.Y. and Dixon, J.E. 1993. Active site labeling of the *Yersinia* protein tyrosine phosphatase: The determination of the pK_a of the active site cysteine and the function of the conserved histidine 402. *Biochemistry* **32**: 9340–9345.
- Zheng, M., Aslund, F., and Storz, G. 1998. Activation of the OxyR transcription factor by reversible disulfide bond formation. *Science* **279**: 1718–1721.