Prediction of Protein Oxidation Sites

Francisco J. Veredas[1(✉)], Francisco R. Cantón[2], and Juan C. Aledo[2]

[1] Dpto. Lenguajes y Ciencias de la Computación, Universidad de Málaga, 29071 Málaga, Spain
franveredas@uma.es

[2] Departamento de Biología Molecular y Bioquímica, Facultad de Ciencias, Universidad de Málaga, 29071 Málaga, Spain
{frcanton,caledo}@uma.es

Abstract. Although reactive oxygen species are best known as damaging agents linked to aerobic metabolism, it is now clear that they can also function as messengers in cellular signalling processes. Methionine, one of the two sulphur containing amino acids in proteins, is liable to be oxidized by a well-known reactive oxygen species: hydrogen peroxide. The awareness that methionine oxidation may provide a mechanism to the modulation of a wide range of protein functions and cellular processes has recently encouraged proteomic approaches. However, these experimental studies are considerably time-consuming, labor-intensive and expensive, thus making the development of *in silico* methods for predicting methionine oxidation sites highly desirable. In the field of protein phosphorylation, computational prediction of phosphorylation sites has emerged as a popular alternative approach. On the other hand, very few *in-silico* studies for methionine oxidation prediction exist in the literature. In the current study we have addressed this issue by developing predictive models based on machine learning strategies and models—random forests, support vector machines, neural networks and flexible discriminant analysis—, aimed at accurate prediction of methionine oxidation sites.

1 Introduction

Although reactive oxygen species (ROS) are best known as damaging agents involved in aerobic metabolism [1], a more subtle approach has emerged in recent years. It is well-known that some ROS, such as hydrogen peroxide (H_2O_2), can work as effective cellular messengers [2,3] by bringing about post-traslational modifications (PTM) that produce reversible changes in the activity of proteins. Amino acids that often experience PTM are those that have a functional group that can serve as a nucleophile in the modification reaction. To this respect, cysteine and methionine, the two sulphur containing residues in proteins, are liable to be oxidized by H_2O_2. For its part, methionine is oxidized to methionine sulfoxide (MetO) by addition of oxygen to its sulphur atom. This oxidation reaction can be reverted by enzyme-catalyzed reduction reactions [4]: MetO is reduced back to methionine by methionine sulfoxide reductases, which are enzymes that

I. Rojas et al. (Eds.): IWANN 2017, Part II, LNCS 10306, pp. 3–14, 2017.
DOI: 10.1007/978-3-319-59147-6_1

are present in all aerobic cells [5]. However, the role of methionine residues in cellular redox regulation remain thoroughly unexplored [6].

Methionine oxidation is a reversible covalent modification. The addition of an oxygen atom to the sulphur atom of methionine residues is able to produce changes in the physico-chemical properties of the whole protein. This, in turn, can affect the activity and stability of the protein [7]. Thus, methionine oxidation has been shown to both down-regulate [8] and up-regulate [9] protein function, through direct oxidation of specific methionine sites in the protein. Furthermore, methionine oxidation can also affect protein function indirectly by coupling oxidative signals to other sorts of PTMs, such as protein phosphorylation [10].

Different proteomic approaches have recently been inspirited by the realisation that methionine oxidation may provide a mechanism to the redox-dependent modulation of protein activity and cellular mechanisms. In this way, proteome-wide studies of methionine oxidation have identified, in both Arabidopsis [11] and human [12], a large number of proteins as potential targets of oxidative signals. Moreover, these proteomic approaches have pointed out the precise sites of oxidation on the target proteins. However, these experimental efforts are considerably labor-intensive, time-consuming and expensive, thus making the development of *in silico* methods for predicting methionine oxidation sites highly desirable. In the field of protein phosphorylation, which can be considered as the most widely studied PTM, computational methods for prediction of phosphorylation sites in proteins have become very popular approaches [13]. Unfortunately, to the best of our knowledge, there are no such methods for methionine oxidation site prediction. Thus, in the current study we have addressed this issue by developing predictive models based on computational intelligence, aimed at accurate prediction of methionine oxidation sites.

2 Materials and Methods

2.1 Datasets

Data regarding methionine peptides that were oxidized in Jurkat cells stressed by H_2O_2 were taken from Table S1 in the supplementary material from reference [12]. This set was further curated to exclude protein entries that have recently been deleted from UniProt (http://www.uniprot.org). The resulting data set was formed by 1646 different proteins accounting for 2616 methionine sulfoxides. A subset of this collection, composed of 774 proteins that exhibit extensive oxidation (degree of oxidation equal or greater than 20%), will be named 'highly H_2O_2-sensitive proteins'.

Using PDB cross-references from the UniProt, we collected a list of PDB identifiers for proteins belonging to the highly H_2O_2-sensitive group. In general, since many proteins were homooligomers, most crystal structures yielded a large number of duplicated observations, which were searched for and eliminated using a R script. Eventually, after removing redundancy and filtering out low quality structures (for instance, those where the target methionine did not appear

resolved), we assembled a collection of 127 unique polypeptides of known structure, containing 1118 methyonil residues, 136 of which were oxidation-prone. For each methionine, the distance of the sulfur atom to the geometric center of the aryl moiety of any aromatic residue was computed with the help of an ad doc R script that relies on the package bio3d [14]. Based on a previously established criterion [15], we considered any methionine sulfur atom within 7 Å of the aromatic ring, to be an S-aromatic motif. For each of the 1118 methionines, relevant information such as its redox status, the corresponding PDB identifier, the positions within the structure of both the analysed methionine and their closer aromatic residues, as well as their distances in Ångstroms, is provided.

2.2 Feature Extraction

For each methionine residue being analysed we have extracted several properties, both from the primary and tertiary structure of the protein.

54 independent variables (input):

- 40 protein's primary-structure features:
  - 40 distance variables:
    - `NT_X`: distance (number of positions in the primary structure) from the analysed methionine to the closest X residue toward the N-terminus.
    - `CT_X`: distance (number of positions in the primary structure) from the analysed methionine to the closest X residue toward the C-terminus.
- 14 protein's tertiary-structure features:
  - 9 inter-atomic distance variables, with X being either Y (tyrosine), F (phenylalanine), W1 (tryptophan ring #1) or W2 (tryptophan ring #2):
    - `Xd`: distance in Å between the S and the centre of the nearest ring of the X aromatic residue.
    - `Xn`: number of X residues at a distance <7 Å.
    - `nBonds`: number of S-aromatic bonds in which the analysed methionine takes part.
  - 2 accessibility properties:
    - `SASA`: solvent accessible surface area of the methionine residue.
    - `SASA_SD`: solvent accessible surface area of the sulphur atom of the methionine residue.
  - 2 entropy variables:
    - `H2`: Shannon base-2 entropy.
    - `H21`: Shannon base-21 entropy.
  - 1 frequency variable:
    - `fM`: relative frequency of methionine at the position of analysed methionine, after multiple sequence alignment (MSA).

Dependent variable (output):

- `oxidable`: binary variable indicating whether the methionine is oxidized (>20%) or not.

A more detailed explanation from some of the features in the list above needs to be given. Thus, to extract the `NT_X-CT_X` variables, given an amino acid, for instance alanine, we searched around the methionine of interest for the closest alanine residue toward the N-terminus and for the closest alanine toward the C-terminus. Once these alanine residues were found in the primary structure, we counted the positions away from the methionine being analysed (`NT_X` and `CT_X` variables, respectively). This operation was repeated for each of the 20 proteinogenic amino acids, accounting for a 40- dimensional vector. Missing values due to the absence of any particular amino acid either toward the N-terminus or the C-terminus, were imputed by using the protein length.

In a recent work we reported that methionyl residues forming part of an S-aromatic motif are less prone to be oxidized [16]. Therefore, 9 additional features related to this non covalent bond were used. Concretely, `Xd` was defined as the distance in ångströms between the sulfur atom from the analysed methionine and the nearest `X` aromatic residue. The variable `Xn` informs about the number of X aromatic residues at a distance <7 Å from the methionine. The feature `nBonds` was computed according to:

$$nBonds = \sum_{X \in \{Y,F,W_1,W_2\}} Xn.$$

Four additional features were related to the conservation of the considered methionine during evolution. To assess these features, besides the human sequence, the orthologous proteins from *Pan troglodytes*, *Gorilla gorilla*, *Rattus norvegicus*, *Bos taurus*, *Gallus gallus*, *Xenopus tropicalis* and *Danio rerio* were aligned. These alignments were used to compute the Shannon entropy according to the equation:

$$H2 = -\sum_{i=1}^{2} f_i log_2(f_i),$$

and

$$H21 = -\sum_{i=1}^{21} f_i log_{21}(f_i),$$

where f_i is the relative frequency of the symbol i at the analysed position across the alignment. Thus, for instance, f_M stands for the relative frequency of methionine. The logarithmic base for $H21$ was taken 21 because in addition to the 20 proteinogenic amino acids, the symbol `'-'` was considered when indels were present. For each analysed methionine, the variables `mean.entropy` and `sd.entropy` were computed as the mean and standard deviation, respectively, of the entropy determined at all the positions of the corresponding protein.

For its part, the solvent accessible features (`SASA` and `SASA_SD`) of each methionine residue was computed using the POPS program [17]. This software, which is freely available at http://mathbio.nimr.mrc.ac.uk, also provides the

accessibility, which is defined as the fraction of the residue surface that is exposed to the solvent.

For those features other than the 40 primary-structure characteristics (i.e. `NT_X` and `CT_X`) missing values were imputed by means of a machine learning approach: a bagging tree model for each predictor was fitted as a function of all the others input variables. Though this method has much higher computational cost than other imputation techniques, such as k-nearest neighbour imputation or imputation via medians, among others, it stands out for being accurate [18].

The data set supporting the results of this article can be downloaded from https://github.com/fveredas/PredictionOfMethionineOxidationSites.

2.3 Machine Learning Methods

Random forests (RF) are used in this study to design predictive models of protein oxidation. RFs are ensemble machine learning methods for classification that function by constructing a large pool of decision trees during the training phase, then giving an output that is the mode of the classes given by the individual trees in the pool. The method combines Breiman's 'bagging' idea and the random selection of features (i.e. predictor-set split) in order to construct a collection of decision trees with controlled variation [19].

The quantification of the variable importance is a crucial issue to interpret data and understand underlying phenomena under the methionine oxidation scenario. RFs use two different measures to estimate variable importance: the accuracy importance (AI) and the Gini-index Importance (GI). The AI of a variable is calculated as the average decrease in accuracy on the OOB samples when the values of the respective predictor are randomly permuted. The GI uses the decrease of Gini-index (impurity) after a node split as a measure of variable relevance. The average decrease in Gini-index over all trees in the RF defines the GI.

To account for the potential of RFs as efficient models for protein oxidation prediction, comparisons with other classification models are mandatory. For this purpose, two machine learning approaches, i.e. support vector machines (SVM) [20] and neural networks (NN) [21], as well as a more classical statistical approach, i.e. flexible discriminant analysis (FDA) [22], have been comparatively used.

Model Tuning. For RF model-fitting in our experiments for Met oxidation, the only sensible tuning parameter would be the number of variables (predictors) randomly sampled as candidates at each split (usually known as `mtry`), but it has been fixed to the optimal recommended value $\lfloor \sqrt{number\ of\ predictors} \rfloor = 7$ [23]. For its part, the RF parameter number of trees to grow has been fixed to 1000 trees to ensure that every input pattern gets predicted at least a few times [24].

For SVMs, a Gaussian radial basis function (RBF) kernel $k(x, x') = e^{-\sigma||x-x'||^2}$ was used (being k a function that calculates the inner product

$\langle \Phi(x), \Phi(x') \rangle$ of two vectors x, x' for a given projection $\Phi : X \rightarrow H$). The problem of model selection (parameter tuning) is partially addressed by an empirical observation for the Gaussian RBF kernel, where the optimal values of the hyperparameter σ are known to lie in between the 0.1 and 0.9 quantile of the $||x - x'||$ statistics [25]. A sample of the training set is used to estimate these quantiles, where any value within the quantile interval results in good performance. This way, σ parameter is automatically estimated. Additionally, the optimal hyperparameter *cost*, that represents the cost of constraints violation and stands for the 'C'-constant of the regularisation term in the Lagrange formulation, is tuned as the one of 12 incremental values in $\{2^i\}_{i=-2}^{9}$ that optimises the area under the ROC curve (AUC) of the SVM classifier.

Single-hidden-layer feed-forward NNs are also constructed and trained with different combinations of parameters to search for the best performance rates in the prediction of methionine oxidation. Optimisation of the NNs is done via the quasi-Newton method BFGS (also known as a variable metric algorithm) [26, 27]. The network `size` (i.e., number of *hidden units*) and *weight* `decay` are the parameters being tuned, selecting that combination of values giving the highest AUC. All the trained NNs have a number of outputs that is equal to the number of classes (i.e. $n = 2$), and a *softmax* output stage. Weights are randomly initialised, and maximum number of epochs was fixed to 100 [27].

Many classification models, such as ridge regression, the lasso, or adaptive regression splines (MARS) [22], can be extended to create discriminant variables. In particular, MARS can be used to create a set of discriminant functions that are non-linear combinations of the original predictors. This conceptual paradigm is referred to as flexible discriminant analysis (FDA) [18]. In this study we have followed a bagging approach for FDA, which uses MARS basis functions to compute a FDA model for each bootstrap sample. The only parameter to be tuned for model fitting was the maximum number of terms (including intercept) in the pruned model [22] (usually known as `nprune`), which is used to enforce an upper bound on the model size. The optimal `nprune` parameter was chosen as that in the range $\{1, \ldots, 25\}$ that gave the highest AUC rate. The maximum degree of interaction (Friedman's *mi*) was fixed to 1, thus an additive model (i.e., no interaction terms) was used.

Resampling Methods for Model Fitting. The data set has been divided into three independent sets, 70% (96 'positive'; 688 'control') patterns for training, 10% (14 'positive'; 98 'control') patterns for evaluation (this pattern set is used to compute the optimal threshold for the ROC curves) and, finally, 20% (26 'positive'; 196 'control') for testing. In order to preserve the unbalanced nature of the original class distribution within the splits a stratified random sampling strategy was used. To estimate the efficacy of the prediction model across the training set, performance measures (AUC, accuracy, sensitivity and specificity) of the out-of-bag (OOB) samples for 10-fold cross-validation with 5 repetitions (50 re-samplings) were calculated and the mean and standard deviation of those rates are summarised. The entire training set is used to fit a final model and its performance was finally measured on the testing set.

In our study class imbalance is inherent to the procedure being followed for data acquisition (see Sect. 2.1): of the complete set of methionine residues found in the 127 polypeptides analysed, only 136 out of 1118 appeared as oxidised, i.e. a mere 12%. Different approaches to counteracting the negative effects of class imbalance have been proposed in the literature [18], with model tuning (using metrics alternative to accuracy such as ROC, Cohen's Kappa or sensitivity), adjusting of prior probabilities, cost-sensitive training, ROC-curve alternative cutoffs, or sampling methods, among others. In this study a combination of the two latter has given the best results. Prior to model fitting, we have used the synthetic minority over-sampling technique (SMOTE) [28] to get a more balanced training dataset. The general idea of this method is to artificially over-sample the minority class (i.e. 'oxidised' class) patterns by generating new samples using the k-nearest neighbours (KNNs) of these cases ($k = 5$ in our experiments). Furthermore, the majority class (i.e. 'not oxidised' class) cases are also under-sampled.

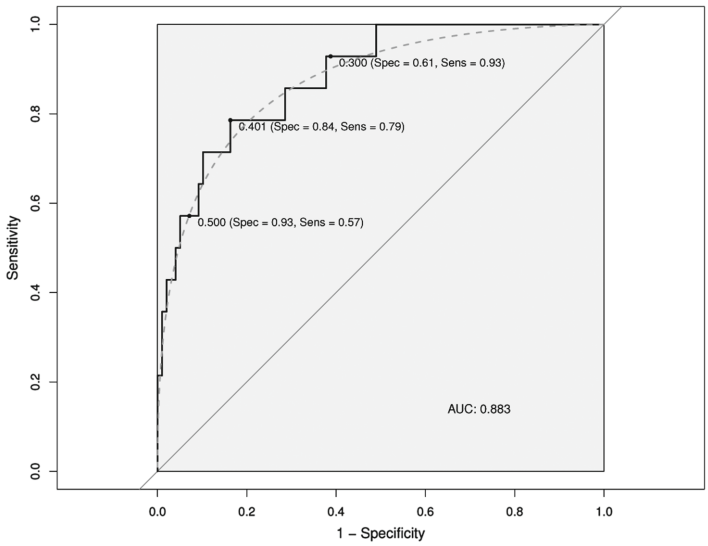

Fig. 1. ROC curve with different thresholds. ROC curve of the RF-Smote classifier on the evaluation data set. Different thresholds have been highlighted in the curve, along with their corresponding specificity and sensibility rates: 0.5 (original), 0.401 (alternative) and 0.3. The theoretical maximal area of reference (i.e. AUC = 1) has been also coloured grey. Dotted grey line represents the smoothed ROC curve.

On the other hand, after model training using the SMOTE samples, we have used the ROC curve to determine alternative cutoffs for the probabilities predicted by the model. Using this ROC curve, an appropriate balance between sensitivity and specificity can be determined. Although several techniques do exist for determining a new cutoff, the more general approach is to find the

point on the ROC curve that is closest (i.e., the shortest distance) to the perfect model (with 100% sensitivity and 100% specificity), which is associated with the upper left corner of the plot [5]. To determine this cutoff point without distorting the results obtained from the final testing dataset, an independent evaluation dataset has been used (see above). In Fig. 1 the ROC curve obtained from the RF classifier (trained with the SMOTE samples) on the evaluation dataset is shown together with the computed alternative cutoff (0.401), as well as the original 0.5 and the 0.3 (shown for comparison purposes) cutoffs. As it can be observed in the figure, the alternative cutoff gives the best balance between sensitivity and specificity.

Table 1 shows the list of parameters being fitted. For each predictive model, the best values for the fitted parameters are computed as those giving the highest averaged AUC via 10-fold cross-validation on the training dataset. The ROC cutoffs obtained from the evaluation dataset after model fitting and training are also shown in the table.

Table 1. Model fitting.

	Fitted parameters	Best tune	ROC cutoff
RF	`mtry`	7	0.401
SVM	σ	0.032	0.138
	`cost`	64	
NN	`size`	17	0.181
	`decay`	0.001	
FDA	`nprune`	25	0.542

3 Results

In the following sections the comparative results from the four predictive models analysed in this study are presented.

3.1 Predicting Methionine Oxidation in Human Proteins

Table 2 shows the performances rates for the four predictive models analysed when applied on the testing dataset. While accuracy levels are above ∼70% for all the classifiers, it is the RF the model that gives the best results, with the highest accuracy rate as well as the best adjusted balance between sensitivity and specificity.

In Fig. 2 the 30 most important variables as estimated by the RF on the training set are shown along with their averaged decrease in Gini-index (see Sect. 2.3).

Table 2. Performance rates of predictive models.

	AUC	Accuracy	Sensitivity	Specificity
RF	0.8459	81.5315	76.9230	82.1429
SVM	0.7501	69.8198	76.9231	68.8776
NN	0.7388	72.9730	61.5385	74.4898
FDA	0.7329	76.5766	50.0000	80.1020

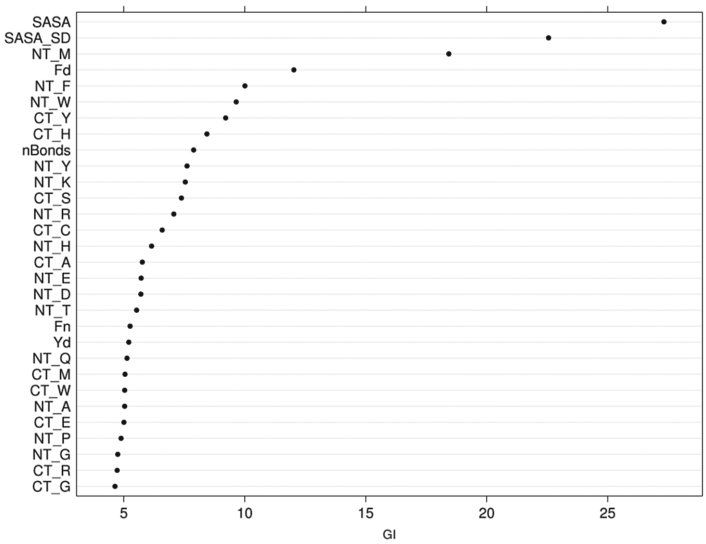

Fig. 2. Variable Importance. The 30 most important variables ordered by GI (averaged decrease in Gini-index) as estimated by the final RF model on patterns in the training dataset.

Predicting oxidation with primary-structure characteristics. Table 3 shows the performance results obtained by training the predictive models with only those protein primary-structure characteristics, i.e. using the `NT_X-CT_X` distance features (see Sect. 2.2) as the only input variables for the classifiers. The same sampling, tuning and training procedures explained in Sect. 2 are followed again. As can observed in the table, the RF gives again the best performance results on the testing set, followed by far by the SVM. Comparing Tables 2 and 3, i.e. comparing the results obtained when using the complete enriched set of primary and tertiary variables to those that used the primary features only, we can observe a significant decrease of the efficacy rates when the tertiary-structure variables are omitted.

Table 3. Using primary-structure features.

	AUC	Accuracy	Sensitivity	Specificity
RF	0.7805	72.5225	65.3846	73.4694
SVM	0.5389	55.8559	53.8462	56.1225
NN	0.5104	41.8919	57.6923	39.7959
FDA	0.6303	52.7027	57.6923	52.0408

3.2 Predicting *In Vitro* Methionine Oxidation

To double-check our predictive model we tested it on a different dataset. For this purpose, we searched through the literature to collect data on the reactivity of methionyl residues from protein pharmaceuticals. In that way, we gathered data for 8 proteins that satisfy the following requirements: (i) the protein should contain at least two methionine with different reactivities, (ii) the kinetics of oxidation in vitro with H_2O_2 of these residues must be reported in the literature, and (iii) the structure of the protein should be known and its PDB must be available. For each protein the reactivities of its methionines were ordered from lower to higher. Those residues showing reactivities lower than the median were labelled as 'oxidised', all others were treated as 'not oxidised' residues. A total of 35 methionine residues were found, of which 15 were labelled as 'oxidised' and 20 as 'not oxidised'. When we dealt with an odd number of residues, the methionine with reactivity equal to the median was sorted into the group containing the residue whose reactivity was closest to the median. For this new protein dataset, our RF model predicts methionine oxidation with accuracy 74.29%, sensibility 46.67% and specificity 95%.

4 Conclusions

Predictive models of methionine oxidation have been designed and tested in this study. Our results show high accuracy rates, with balanced sensitivity and specificity. The best efficiency results were obtained with random forests, while support vector machines and networks behaved worse, in general.

From the 54 predictors used in the design of the predictive models, some tertiary-structure ones, such as *solvent-accessibility area* have been identified as those with the highest contribution to the predictive power of the random forest model. Moreover, counting on a reliable computational tool to predict methionine oxidation could stimulate further investigation to determine the role of sulphoxidation in cellular oxidative signalling.

Our predictive models also include another important set of characteristics that reinforces their predictive power: *methionine-aromatic motif characteristics*. The *distance in Å between the methionine and the nearest aromatic residue* stands out as one of the most important variables for the predictive mosdels. This result, together with the highest relevance as predictive variable shown by the *solvent accessible surface area of the methionine residue*, emphasises the complex correlation between structural properties and methionine oxidation.

Acknowledgments. This work was partially supported by the Universidad de Málaga and project TIN2014-58516-C2-1-R, MICINN, Plan Nacional de I+D+I.

References

1. Aledo, J.C.: Life-history constraints on the mechanisms that control the rate of ROS production. Curr. Genom. **15**, 217–230 (2014)
2. Collins, Y., Chouchani, E.T., James, A.M., Menger, K.E., Cochemé, H.M., Murphy, M.P.: Mitochondrial redox signalling at a glance. J. Cell Sci. **125**, 801–806 (2012)
3. Veredas, F.J., Cantón, F.R., Aledo, J.C.: Methionine residues around phosphorylation sites are preferentially oxidized in vivo under stress conditions. Sci. Rep. **7**, 40403 (2017)
4. Arnér, E.S., Holmgren, A.: Physiological functions of thioredoxin and thioredoxin reductase. Eur. J. Biochem. **267**, 6102–6109 (2000)
5. Kim, H.Y.: The methionine sulfoxide reduction system: selenium utilization and methionine sulfoxide reductase enzymes and their functions. Antioxid. Redox Sig. **19**, 958–969 (2013)
6. Kim, G., Weiss, S.J., Levine, R.L.: Methionine oxidation and reduction in proteins. BBA-Gen. Subj. **1840**, 901–905 (2014)
7. Jacques, S., Ghesquière, B., Breusegem, F., Gevaert, K.: Plant proteins under oxidative attack. Proteomics **13**, 932–940 (2013)
8. Härndahl, U., Kokke, B.P., Gustavsson, N., Linse, S., Berggren, K., Tjerneld, F., Boelens, W.C., Sundby, C.: The chaperone-like activity of a small heat shock protein is lost after sulfoxidation of conserved methionines in a surface-exposed amphipathic alpha-helix. Biochim. Biophys. Acta **1545**, 227–237 (2001)
9. Drazic, A., Miura, H., Peschek, J., Le, Y., Bach, N.C., Kriehuber, T., Winter, J.: Methionine oxidation activates a transcription factor in response to oxidative stress. Proc. Natl. Acad. Sci. USA **110**, 9493–9498 (2013)
10. Rao, R.S.P., Møller, I.M., Thelen, J.J., Miernyk, J.A.: Convergent signaling pathways—interaction between methionine oxidation and serine/threonine/tyrosine O-phosphorylation. Cell Stress Chaperones **20**, 15–21 (2014)
11. Jacques, S., Ghesquière, B., Bock, P.J., Demol, H., Wahni, K., Willemns, P., Messens, J., Breusegem, F., Gevaert, K.: Protein methionine sulfoxide dynamics in arabidopsis thaliana under oxidative stress. Mol. Cell. Proteomics **14**, 1217–1229 (2015)
12. Ghesquière, B., Jonckheere, V., Colaert, N., Van Durme, J., Timmerman, E., Goethals, M., Schymkowitz, J., Rousseau, F., Vandekerckhove, J., Gevaert, K.: Redox proteomics of protein-bound methionine oxidation. Mol. Cell. Proteomics **10**, M110.006866 (2011)
13. Datta, S., Mukhopadhyay, S.: A grammar inference approach for predicting kinase specific phosphorylation sites. PLoS One **10**, e0122294 (2015)
14. Grant, B.J., Rodrigues, A.P.C., ElSawy, K.M., McCammon, J.A., Caves, L.S.D.: Bio3D: an R package for the comparative analysis of protein structures. Bioinformatics **22**, 2695–2696 (2006)
15. Valley, C.C., Cembran, A., Perlmutter, J.D., Lewis, A.K., Labello, N.P., Gao, J., Sachs, J.N.: The methionine-aromatic motif plays a unique role in stabilizing protein structure. J. Biol. Chem. **287**, 34979–34991 (2012)
16. Aledo, J.C., Cantón, F.R., Veredas, F.J.: Sulphur atoms from methionines interacting with aromatic residues are less prone to oxidation. Sci. Rep. **5** (2015)

17. Cavallo, L.: POPS: a fast algorithm for solvent accessible surface areas at atomic and residue level. Nucleic Acids Res. **31**, 3364–3366 (2003)
18. Kuhn, M., Johnson, K.: Applied Predictive Modeling. Springer, New York (2013)
19. Breiman, L.: Random forests. Mach. Learn. **45**, 5–32 (2001)
20. Vapnik, V.: Statistical Learning Theory. Wiley, New York (1998)
21. Ripley, B.D.: Pattern Recognition and Neural Networks. Cambridge University Press, Cambridge (2007)
22. Friedman, J.H.: Multivariate adaptive regression splines. Ann. Stat. **19**, 1–67 (1991)
23. Díaz-Uriarte, R., Alvarez de Andrés, S.: Gene selection and classification of microarray data using random forest. BMC Bioinform. **7**, 3 (2006)
24. Liaw, A., Wiener, M.: Classification and regression by randomForest. R News **2**, 18–22 (2002)
25. Karatzoglou, A., Smola, A., Hornik, K., Zeileis, A.: kernlab - an {S4} package for kernel methods in {R}. J. Stat. Softw. **11**, 1–20 (2004)
26. Nash, J.C.: Compact Numerical Methods for Computers: Linear Algebra and Function Minimisation, 2nd edn. CRC Press, New York (1990)
27. Venables, W.N., Ripley, B.D.: Modern Applied Statistics with S, 4th edn. Springer, New York (2002)
28. Chawla, N.V., Bowyer, K.W., Hall, L.O., Kegelmeyer, W.P.: SMOTE: synthetic minority over-sampling technique. J. Artif. Intell. Res. **16**, 321–357 (2002)