Genetics and Molecular Research

A computational method to predict carbonylation sites in yeast proteins

H.Q. Lv[1,2], J. Liu[3], J.Q. Han[1], J.G. Zheng[1] and R.L. Liu[1]

[1]School of Electronic and Information Engineering, Xi'an Jiaotong University, Xi'an, China
[2]School of Life Science and Technology, Xi'an Jiaotong University, Xi'an, China
[3]School of Electrical Engineering, Xi'an Jiaotong University, Xi'an, China

Corresponding author: J. Liu
E-mail: eeliujun@gmail.com

Genet. Mol. Res. 15 (2): gmr.15028006
Received November 5, 2015
Accepted February 11, 2016
Published June 21, 2016
DOI http://dx.doi.org/10.4238/gmr.15028006

ABSTRACT. Several post-translational modifications (PTM) have been discussed in literature. Among a variety of oxidative stress-induced PTM, protein carbonylation is considered a biomarker of oxidative stress. Only certain proteins can be carbonylated because only four amino acid residues, namely lysine (K), arginine (R), threonine (T) and proline (P), are susceptible to carbonylation. The yeast proteome is an excellent model to explore oxidative stress, especially protein carbonylation. Current experimental approaches in identifying carbonylation sites are expensive, time-consuming and limited in their abilities to process proteins. Furthermore, there is no bioinformational method to predict carbonylation sites in yeast proteins. Therefore, we propose a computational method to predict yeast carbonylation sites. This method has total accuracies of 86.32, 85.89, 84.80, and 86.80% in predicting the carbonylation sites of K, R, T, and P, respectively. These results were confirmed by 10-fold cross-validation. The ability to identify carbonylation sites in different kinds of features was analyzed and the position-specific composition of the modification site-

flanking residues was discussed. Additionally, a software tool has been developed to help with the calculations in this method. Datasets and the software are available at https://sourceforge.net/projects/hqlstudio/files/CarSpred.Y/.

Key words: Yeast carbonylation; Carbonylation site prediction; CarSPred.Y

INTRODUCTION

Oxidative stress reflects an imbalance between production and degradation of reactive nitrogen species (RNS) and reactive oxygen species (ROS). (Kim et al., 2010). Oxidative stress arises when an elevated production of ROS has surpassed the detoxification ability of the cell for reactive intermediates (Bollineni et al., 2011). As a result, cellular macromolecules such as proteins, lipids, nucleic acids and carbohydrates can be modified (Dalle-Donne et al., 2003a). Although reversible oxidative modifications are thought to be relevant in physiological processes, the irreversible modifications are known to lead to cellular damage (Møller et al., 2011). Severe oxidative stress can result in reduced cellular signaling capacity, diminished proteasome and lysosome functions, weakened cellular viability, and even cell death (Mullineaux and Baker, 2010; Madian et al., 2011).

Among a variety of oxidative stress-induced PTM, protein carbonylation is an irreversible process, and considered a biomarker of oxidative stress (Dalle-Donne et al., 2003b). Only select proteins can undergo carbonylation that can only occur on four amino acid residues: lysine (K), arginine (R), threonine (T), and proline (P) (Maisonneuve et al., 2009). The yeast proteome has been shown to be an excellent model to study oxidative stress, especially protein carbonylation. As a eukaryotic cell with a short life cycle, yeast can be studied in several different experimental conditions in a short period of time. Furthermore, yeast has a small but well-defined genome. Lastly, yeast expresses numerous proteins that are orthologous to mammalian proteins (MacLean et al., 2001; Longo and Fabrizio, 2002; Mirzaei and Regnier, 2008). However, the yeast genome encodes for approximately 6000 proteins (Mirzaei and Regnier, 2006), and current experimental approaches used to identify carbonylation sites are expensive, time-consuming, and limited in protein processing abilities. Moreover, there is no bioinformatics method to predict carbonylation sites in yeast proteins. The only tools currently employed, CSPD (Maisonneuve et al., 2009) and CarSPred 1.0 (Lv et al., 2014), are limited to *Escherichia coli* and human proteomes, respectively. Therefore, it is necessary to develop a computational method for prediction of carbonylation sites in yeast proteins.

In this paper, a computational method for predicting carbonylation sites on K, R, T, and P in yeast proteins was proposed. Datasets were gathered from the latest proteomic studies on yeast carbonylation. Information regarding amino acid composition (AAC), position-specific amino acid propensity (PSAAP), as well as physicochemical and biochemical properties, were extracted from sample sequences. Student's *t*-test evaluation criterion and incremental feature selection (IFS) were combined to determine the final optimization feature sets. Weighted support vector machine (WSVM) was applied for the classification of unbalanced training samples. In addition, the ability to identify carbonylation sites using different types of available

information was analyzed. Position-specific composition of flanking residues at modification sites was also discussed. Finally, a software tool was developed as a computational aid for the proposed method under the win32 environment.

MATERIAL AND METHODS

Datasets

Collected datasets consisted of carbonylated protein sequences and K, R, T and P carbonylation sites in yeast. Since carbonylation data cannot be found in any of the public databases, relevant data were extracted from literature. A dataset for carbonylation sites of yeast proteins was established that comprised 224 carbonylated protein sequences as well as 135 K, 92 R, 62 T, and 90 P carbonylation sites from six yeast proteomic studies. The statistics and corresponding references of all sources are shown in **S1 Table**.

Positive and negative carbonylation sites

Protein sequences were excluded if there was any ambiguity regarding the carbonylation sites. The accession numbers, corresponding references, reasons for elimination, and other relevant information about these protein sequences are shown in **S2 Table**. The remaining sequences were used to prepare positive and negative carbonylation sites. Carbonylation sites in the sequences were regarded as positive carbonylation sites, unless certain criteria were fulfilled to be considered as a negative site. If a residue has the same amino acid type with an experimentally verified carbonylation site, and this residue has not been reported to be a positive site, it can be considered as a negative site. In addition, a negative site has to be located within a protein sequence which contains positive sites. Lastly, a negative site should be extracted from a dataset which contained the same type of positive sites. Residues that satisfied all three criteria were denoted as negative carbonylation sites.

Sample preparation

The ± n (N = 6, 7, …, 13) flanking residues of positive and negative carbonylation sites were used as positive and negative candidate sample sequences, respectively. Since the central residue is always the same, it was excluded from each candidate sample sequence. The CD-HIT program (Huang et al., 2010) was employed to retrieve non-redundant sample sequences with a cut-off threshold of 65%.

Sample imbalance correction

The order of magnitude for the numbers of positive and negative sample sequences was different. It is known that carbonylation sites are vastly dominated by the same type of residues in a carbonylated protein. However, highly imbalanced samples may induce inaccuracy of some classifiers (Japkowicz and Stephen, 2002), and may result in artificial evaluation of these methods. In view of this, negative samples were chosen to match the

positive samples. The number of negative samples was approximately six times that of positive samples. This approach was applied to all the four types of residues (K, R, T and P). The final samples contained 86 K, 56 R, 44 T, and 59 P positive training sample sequences, as well as 536 K, 363 R, 271 T, and 358 P negative training sample sequences. The dataset is summarized in Table 1.

Table 1. Carbonylation datasets of the yeast presented in this study.

Group	No. of carbonylated proteins	No. of carbonylation sites			
		K	R	T	P
Original sequences	224	135	92	62	90
Positive samples	216	86	56	44	59
Negative samples	216	536	363	271	358

No. of carbonylated proteins corresponding to the original sequences is smaller than that corresponding to the samples, since eight proteins have been filtered out. For details, please refer to **S2 Table**.

Feature extraction

AAC, PSAAP, and HQI, were included in the feature extraction procedure. A total of 20 native amino acids and one dummy amino acid, X, was included in the feature extraction approach.

AAC features

It has been demonstrated that a large number of carbonylation sites are found in RKPT-enriched regions (Maisonneuve et al., 2009; Rao and Møller, 2011). In addition, the sequences flanking the RKPT-enriched regions are rich in various residues including iron-binding sites and hydrophobic amino acids (Maisonneuve et al., 2009). Based on the above, AAC was employed to extract amino acid composition information from residues flanking the carbonylation sites.

The composition of different types of amino acids in each sample sequence was considered. The frequencies corresponding to 20 native amino acids were calculated, and the dummy amino acid X was neglected. Therefore, a 20-dimensional feature vector was extracted from each sample sequence, with a sum of 1. The dimension of AAC vector was independent of residue length of sample sequences, as AAC describes the composition of amino acids, and is not affected by residue positions.

PSAAP features

PSAAP has been successfully used in various applications as well as in PTM prediction of phosphorylation and S-nitrosylation sites (Tang et al., 2007; Xu et al., 2013). In the PSAAP encoding scheme, the absolute frequencies of different types of residues in sample sequences were computed to construct a position-specific amino acid propensity matrix. The feature vector of a query sample sequence can be generated by looking up the corresponding elements in this matrix. The position-specific amino acid propensity matrix was given by:

$$\begin{cases} P(\mathrm{i},\mathrm{j}) = 0, & \delta_{NEG(j)} = 0 \\ P(i,j) = \dfrac{F_{POS}(i,j) - F_{NEG}(i,j)}{\delta_{NEG(j)}}, & \delta_{NEG(j)} \neq 0 \end{cases} \quad \text{(Equation 1)}$$

where the dimension of $P(i,j)$ was 21 x 2n; 21 is the number of amino acid types, and 2n denotes the residue length of sample sequences. $F_{POS}(i,j)$ represents the absolute frequency of amino acid type i appearing at position j in positive training samples; $F_{NEG}(i,j)$ is similarly represented. $\delta_{NEG(j)}$ denotes the standard deviation of column j of the absolute frequency matrix F_{NEG}.

We generated a 2n-dimensional feature vector for each given sample sequence using the PSAAP encoding scheme. The vector described the position-specific probability of amino acids in the residue fragments flanking a possible carbonylation site.

HQI features

AAIndex was used to determine the biochemical and physicochemical properties of amino acids. It is a database widely used in PTM site predictions (Kawashima et al., 2008; Chen et al., 2013). One problem was that overfitting and computational tractability would arise if a large number of properties are involved in a classification problem (Trost and Kusalik, 2013). Therefore, high-quality indices (HQI) was introduced to deal with the electric properties, hydrophobicity, alpha and turn propensities, physicochemical properties, residue propensity, composition, beta propensity, as well as intrinsic propensities of amino acids using a sophisticated method called consensus fuzzy clustering (Saha et al., 2012).

The normalized HQI8 was used to generate a 2n x 8-dimensional feature vector for each sample sequence. The corresponding value of residue X was set to 0. This vector described the biochemical and physicochemical properties of residues flanking a potential carbonylation site.

Student *t*-test and the IFS curve

The Student *t*-test is a univariate feature selection criterion, which ranks all features according to a statistical significance score. Surprisingly, the simple Student's *t*-test can outperform more complex wrappers or embedded feature selection methods in the study of molecular signatures (Haury et al., 2011). In addition, it has been used to analyze the solvent accessible property of PTM sites (Xu et al., 2010). In this paper, three kinds of features were extracted from each sample sequence, and the total number of these features was 21 + 2n + 2n x 8. For instance, when the window size parameter n is equal to 6, the total number of feature will be 129. These features were then ranked according to the Student *t*-test evaluation criterion, and IFS curves were employed to determine the dimensions of the final optimization feature sets.

WSVM classifier

Support vector machine (SVM) is a popular classifier used in many bioinformatics

approaches (Liu et al., 2015). It can be applied to determine the decision surface from two distinct classes of positive and negative samples in a feature space. WSVM is based on the standard SVM, and has additional abilities to compensate for bias due to imbalanced dataset by assigning each subset a different penalty coefficient. In this paper, the WSVM in libsvm (Chang and Lin, 2011) was used to solve the classification problem of unbalanced samples, which is available at http://www.csie.ntu.edu.tw/~cjlin/libsvm/.

Performance assessment

In this paper, the 10-fold cross-validation was chosen for preparation of validation datasets. Matthew's correlation coefficient (MCC) and total accuracy (TA) were used to quantitatively evaluate the reliability and capability of the proposed method. The TA and MCC were given by:

$$\begin{cases} TA = \dfrac{TP + TN}{TP + TN + FP + FN} \\ MCC = \dfrac{TPxTN - FPxFN}{\sqrt{(TP + FP)(\mathrm{TP}+ \mathrm{FN})(\mathrm{TN}+ \mathrm{FP})(\mathrm{TN}+ \mathrm{FN})}} \end{cases} \quad \text{(Equation 2)}$$

where TP and FN are the number of positive data that are predicted to be positive and negative, respectively; TN and FP denote the number of negative data that is predicted to be negative and positive.

RESULTS

Final optimization feature set

The final optimization feature sets were determined by Student *t*-test feature evaluation criterion and IFS curves. Average MCC values with incremental Student's *t*-test features and different window sizes (N = 7-13) were computed using 10-fold cross-validation, and the corresponding IFS curves were plotted in Figure 1. It can be seen that the IFS curves for the four types of carbonylation sites (K, R, T and P) peaked when n was equal to 11, 10, 8, and 13; the top 14, 12, 7, and 11 Student's t-test features were selected. Therefore, the four optimization feature sets were eventually chosen to serve for K, R, T, and P carbonylation site predictions respectively. Additionally, only IFS curves for N = 8-13 are shown in Figure 1 for simplification.

Method performance

The WSVM classifiers were trained and tested through 10-fold cross-validation based on the K, R, T, and P carbonylation datasets in yeast. The probability results of the 10 iterations were spliced into one to serve the average TA and MCC computation. The proposed method achieved total accuracies of 86.32, 85.89, 84.80, and 86.80% for the four types of

carbonylation site (K, R, T, and P) predictions, as evaluated by 10-fold cross-validation. The corresponding MCC values were 0.2422, 0.2487, 0.1530 and, 0.3284.

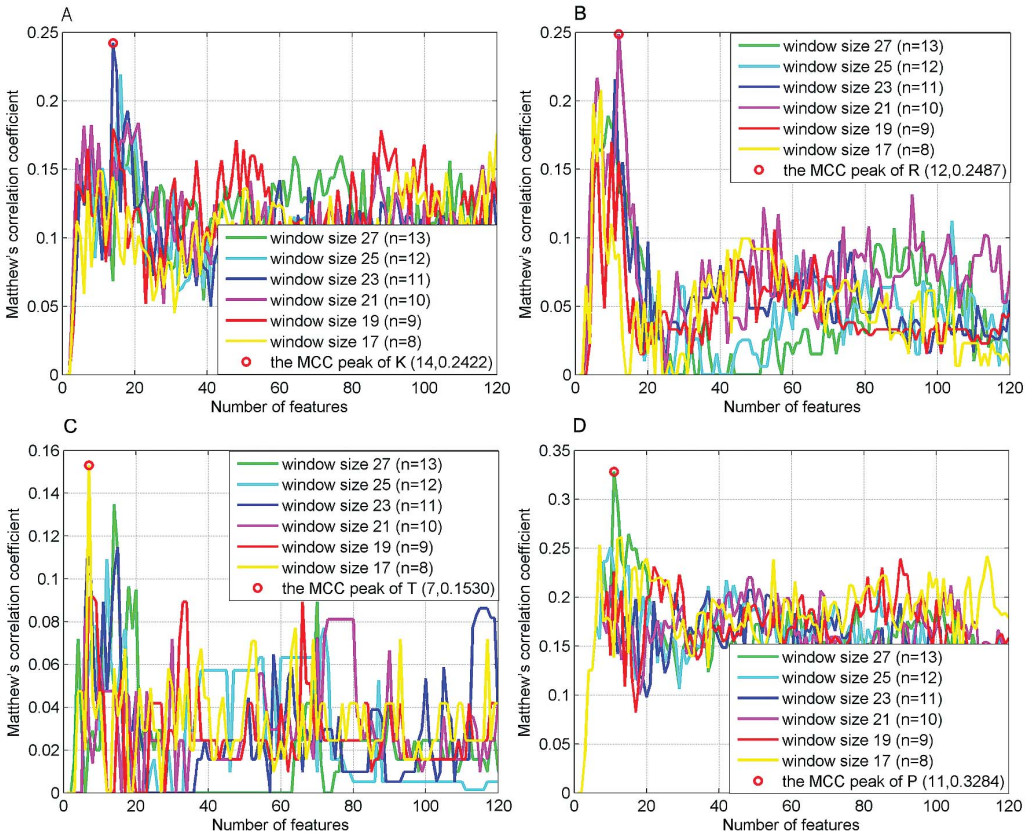

Figure 1. Change of average MCC values versus the number of Student *t*-test features and different window sizes using 10-fold cross-validation.The MCC values show peaks (red circle) when N = 11, 10, 8, and 13. The top 14, 12, 7, and 11 features were selected separately. **A.** K carbonylation site prediction. **B.** R carbonylation site prediction. **C.** T carbonylation site prediction. **D.** P carbonylation site prediction.

The predictive power of this method is still not very high. It is restricted by the following factors in addition to the method itself. Some of the carbonylatable residues were assumed to be negative samples, but may be revealed as carbonylation sites under different experimental conditions. Therefore, assignment of negative carbonylation sites can only be tentative. Additionally, there is a limitation to the training sample size, which will result in insufficient training of the classifier to some degree.

CarSPred software

A software tool named CarSPred.Y for win32 was developed to facilitate the application of this method. In this software, the four types of carbonylation sites (K, R, T

and P) are separately predicted in individual modules. The input format of these modules can be FASTA format sequences or a file, and the output can also be chosen from a list outputs or bed printed into a file. For the latter output, the locations and probabilities of putative carbonylation sites are clearly indicated with new annotations. More detailed instructions can be found on the software manual.

DISCUSSION

Feature analysis

The AAC, PSAAP and HQI features discussed in the paper vary in their abilities to identify carbonylation sites in yeast proteins. However, relying on counting the total number of features in the final optimization feature sets is unreliable as the total dimensions of the three types of features are different. Therefore, the average Student *t*-test scores of the three kinds of features based on the positive and negative sample sequences were considered instead. The scores corresponding to K, R, T, and P final optimization feature sets at the optimal window sizes were computed and shown in Figure 2. It was hypothesized that the AAC and PSAAP features play a more important role in the identification of yeast carbonylation sites.

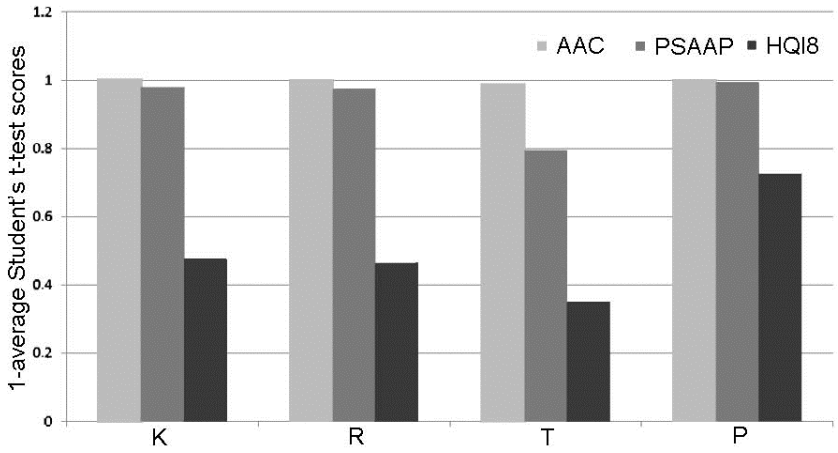

Figure 2. Average Student *t*-test scores of the three kinds of features in the K, R, T, and P feature sets at the optimal window size.

Position-specific composition analysis

The AAC and PSAAP are important features in the Student *t*-test feature list. Therefore, a web-based analysis application named Two Sample Logo (TSL) (Vacic et al., 2006) was used to analyze the position-specific composition of residues flanking carbonylation and non-carbonylation sites. Using the TSL tool with the default parameter options, statistical significance was calculated for each flanking residue at the modification site, which was graphically represented.

The position-specific composition differences between positive and negative

sample sequences of yeast carbonylation sites were shown in Figure 3. There were obvious differences among the K, R, T, and P carbonylation sites. However, for residues sharing the same composition as the carbonylation site, the degree of enrichment downstream of the site was lower than that found of upstream of the site. This was consistent with our previous study on carbonylation sites of human proteins (Lv et al., 2014). Therefore, this may be an important and general rule for K, R, T, and P carbonylation sites.

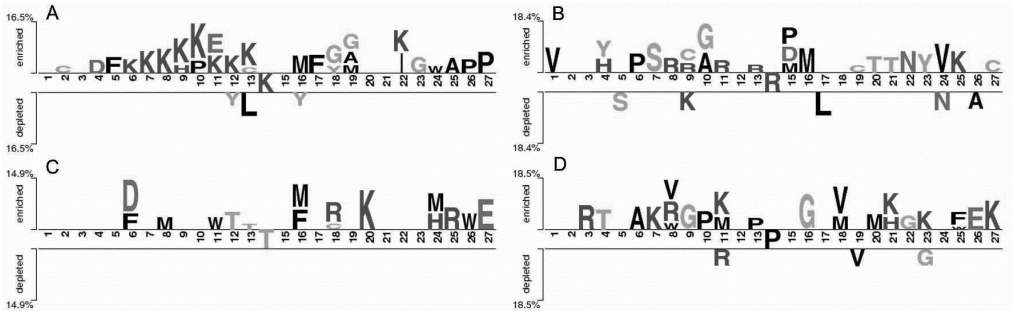

Figure 3. Two-sample-logos of the position-specific composition of residues flanking the positive and negative carbonylation sites in yeast proteins. Graphical residues in the positive samples are separated in two groups, enriched, and depleted. **A.** K carbonylation sites, **B.** R carbonylation sites, **C.** T carbonylation sites, **D.** P carbonylation sites.

Conflicts of interest

The authors declare no conflicts of interest.

ACKNOWLEDGMENTS

Research supported by grants from China Postdoctoral Science Foundation (#2015M580851).

REFERENCES

Bollineni RCh, Hoffmann R and Fedorova M (2011). Identification of protein carbonylation sites by two-dimensional liquid chromatography in combination with MALDI- and ESI-MS. *J. Proteomics* 74: 2338-2350. http://dx.doi.org/10.1016/j.jprot.2011.07.002

Chang CC and Lin CJ (2011). LIBSVM: A library for support vector machines. *ACM Trans. Intell. Syst. Technol.* 2: 27:1-27:7.

Chen X, Qiu JD, Shi SP, Suo SB, et al. (2013). Incorporating key position and amino acid residue features to identify general and species-specific Ubiquitin conjugation sites. *Bioinformatics* 29: 1614-1622. http://dx.doi.org/10.1093/bioinformatics/btt196

Dalle-Donne I, Giustarini D, Colombo R, Rossi R, et al. (2003a). Protein carbonylation in human diseases. *Trends Mol. Med.* 9: 169-176. http://dx.doi.org/10.1016/S1471-4914(03)00031-5

Dalle-Donne I, Rossi R, Giustarini D, Milzani A, et al. (2003b). Protein carbonyl groups as biomarkers of oxidative stress. *Clin. Chim. Acta* 329: 23-38. http://dx.doi.org/10.1016/S0009-8981(03)00003-2

Haury AC, Gestraud P and Vert JP (2011). The influence of feature selection methods on accuracy, stability and interpretability of molecular signatures. *PLoS One* 6: e28210. http://dx.doi.org/10.1371/journal.pone.0028210

Huang Y, Niu B, Gao Y, Fu L, et al. (2010). CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 26: 680-682. http://dx.doi.org/10.1093/bioinformatics/btq003

Japkowicz N and Stephen S (2002). The class imbalance problem: a systematic study. *Intell. Data Anal. J.* 6: 429-449.
Kawashima S, Pokarowski P, Pokarowska M, Kolinski A, et al. (2008). AAindex: amino acid index database, progress report 2008. *Nucleic Acids Res.* 36: D202-D205. http://dx.doi.org/10.1093/nar/gkm998
Kim JH, Sedlak M, Gao Q, Riley CP, et al. (2010). Dynamics of protein damage in yeast frataxin mutant exposed to oxidative stress. *OMICS* 14: 689-699. http://dx.doi.org/10.1089/omi.2010.0051
Liu J, Han J and Lv H (2015). ADPRtool: A novel predicting model for identification of ASP-ADP-Ribosylation sites of human proteins. *J. Bioinform. Comput. Biol.* 13: 1550015. http://dx.doi.org/10.1142/S0219720015500158
Longo VD and Fabrizio P (2002). Regulation of longevity and stress resistance: a molecular strategy conserved from yeast to humans? *Cell. Mol. Life Sci.* 59: 903-908. http://dx.doi.org/10.1007/s00018-002-8477-8
Lv H, Han J, Liu J, Zheng J, et al. (2014). CarSPred: a computational tool for predicting carbonylation sites of human proteins. *PLoS One* 9: e111478. http://dx.doi.org/10.1371/journal.pone.0111478
MacLean M, Harris N and Piper PW (2001). Chronological lifespan of stationary phase yeast cells; a model for investigating the factors that might influence the ageing of postmitotic tissues in higher organisms. *Yeast* 18: 499-509. http://dx.doi.org/10.1002/yea.701
Madian AG, Myracle AD, Diaz-Maldonado N, Rochelle NS, et al. (2011). Differential carbonylation of proteins as a function of in vivo oxidative stress. *J. Proteome Res.* 10: 3959-3972. http://dx.doi.org/10.1021/pr200140x
Maisonneuve E, Ducret A, Khoueiry P, Lignon S, et al. (2009). Rules governing selective protein carbonylation. *PLoS One* 4: e7269. http://dx.doi.org/10.1371/journal.pone.0007269
Mirzaei H and Regnier F (2006). Identification and quantification of protein carbonylation using light and heavy isotope labeled Girard's P reagent. *J. Chromatogr. A* 1134: 122-133. http://dx.doi.org/10.1016/j.chroma.2006.08.096
Mirzaei H and Regnier F (2008). Protein:protein aggregation induced by protein oxidation. *J. Chromatogr. B Analyt. Technol. Biomed. Life Sci.* 873: 8-14. http://dx.doi.org/10.1016/j.jchromb.2008.04.025
Møller IM, Rogowska-Wrzesinska A and Rao RS (2011). Protein carbonylation and metal-catalyzed protein oxidation in a cellular perspective. *J. Proteomics* 74: 2228-2242. http://dx.doi.org/10.1016/j.jprot.2011.05.004
Mullineaux PM and Baker NR (2010). Oxidative stress: antagonistic signaling for acclimation or cell death? *Plant Physiol.* 154: 521-525. http://dx.doi.org/10.1104/pp.110.161406
Rao RS and Møller IM (2011). Pattern of occurrence and occupancy of carbonylation sites in proteins. *Proteomics* 11: 4166-4173. http://dx.doi.org/10.1002/pmic.201100223
Saha I, Maulik U, Bandyopadhyay S and Plewczynski D (2012). Fuzzy clustering of physicochemical and biochemical properties of amino acids. *Amino Acids* 43: 583-594. http://dx.doi.org/10.1007/s00726-011-1106-9
Tang YR, Chen YZ, Sheng ZY and Zhang Z (2007). Predicting protein phosphorylation sites with neuralgenetic network algorithm, Bioinformatics and Biomedical Engineering, 2007. ICBBE 2007. The 1st International Conference on: 111-114.
Trost B and Kusalik A (2013). Computational phosphorylation site prediction in plants using random forests and organism-specific instance weights. *Bioinformatics* 29: 686-694. http://dx.doi.org/10.1093/bioinformatics/btt031
Vacic V, Iakoucheva LM and Radivojac P (2006). Two Sample Logo: a graphical representation of the differences between two sets of sequence alignments. *Bioinformatics* 22: 1536-1537. http://dx.doi.org/10.1093/bioinformatics/btl151
Xu G, Paige JS and Jaffrey SR (2010). Global analysis of lysine ubiquitination by ubiquitin remnant immunoaffinity profiling. *Nat. Biotechnol.* 28: 868-873. http://dx.doi.org/10.1038/nbt.1654
Xu Y, Ding J, Wu LY and Chou KC (2013). iSNO-PseAAC: predict cysteine S-nitrosylation sites in proteins by incorporating position specific amino acid propensity into pseudo amino acid composition. *PLoS One* 8: e55844. http://dx.doi.org/10.1371/journal.pone.0055844

Supplementary material

S1 Table. Carbonylation sites of yeast proteins collected in proteomic studies.

S2 Table. Carbonylation sites and their corresponding proteins identified in proteomic studies but excluded in this paper.