Analysis of residue interaction networks to improve prediction of protein succination

Analysis of residue interaction networks to improve prediction of protein succination

Arianna Carolina Rosa[1], Gianluca Miglio[1,2]

[1]Department of Drug Science and Technology, and [2]Scientific Computing Competence Centre (C³S), University of Turin, Italy

E-mail: gianluca.miglio@unito.it

Introduction

Protein succination (Figure 1) results from a specific adduction of fumarate to certain cysteine (Cys) residues in proteins. In the last years, the interest toward this type of post-translational Cys modification is markedly increased following the observation of an aberrant level of modified proteins in tumours found in patients affected by hereditary leiomyomatosis and renal cell cancer (Yang et al., *Metabolites* 2014; 4: 640-54), in tissues of obese and diabetic mice (Thomas et al., *Obesity* 2012, 20: 263-9), as well as in cells exposed to dimethyl fumarate (Piroli et al., *Biochem J* 2016; 462: 231-45), which is a drug used to treat patients with multiple sclerosis or psoriasis. Despite the compelling findings, however, further efforts remain to be accomplished to obtain a comprehensive examination and characterization of this type of Cys modification. The aim of this study was to elucidate the role of microenvironment-related factors in governing the specificity of protein succination.

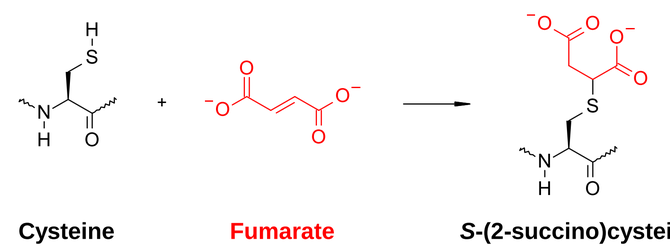

Figure 1. Reaction between fumarate and Cys sites. Nucleophilic adduction of Cys to fumarate *via* a Michael reaction yields S-(2-succino)cysteine (2SC) sites.

Methods

The analytical procedure is shown in Figure 2. A dataset of fumarate-sensitive proteins and sites was built by collecting data as previously described (Miglio et al., *Biochim Biophys Acta* 2016; 1864: 211-8). Data were generated/collected from two workflows leading to the: (*a*) generation of topological data by the analysis of the residue interaction networks (RINs), using together the UCSF Chimera (1.11.2) software (http://www.cgl.ucsf.edu/chimera/) and RINalyzer (http://www.rinalyzer.de), a Cytoscape-plugin for protein structure network assessment; (*b*) collection of biochemical and biological data from web sources/tools [PropKa (http://nbcr-222.ucsd.edu/pdb2pqr_2.0.0/), and DSSP web tool (http://swift.cmbi.ru.nl/gv/dssp/)]. Finally, the collected data were analysed and visualized using the **R** software (The R Project for Statistical Computing; https://www.r-project.org/).

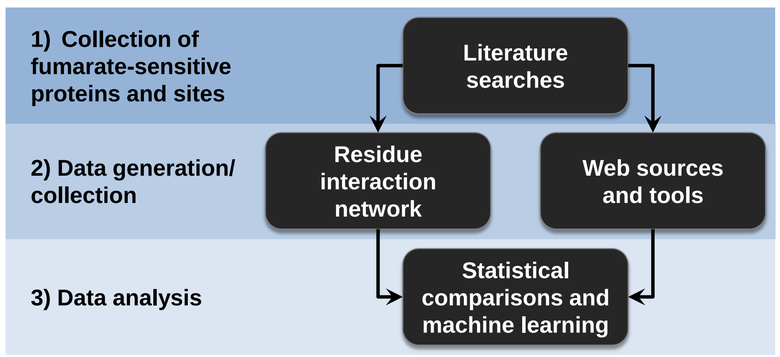

Figure 2. Overview of the analytical procedure.

Results

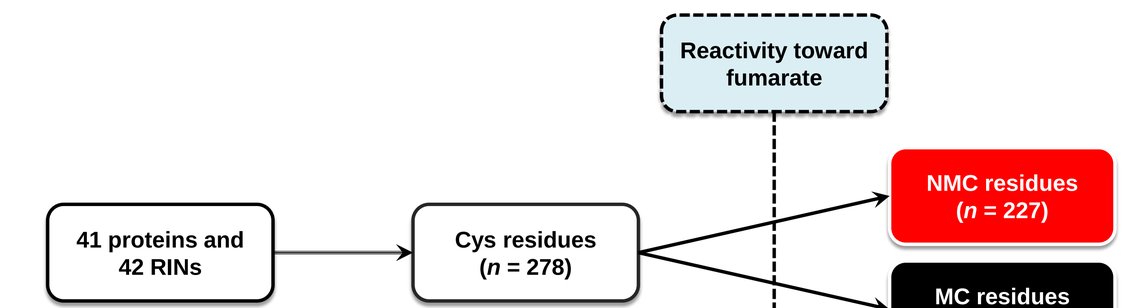

Figure 3. Overview of the included proteins and sites. A total of 278 Cys residues were found in 42 RINs generated from the 41 proteins included in this study. According to their reactivity toward fumarate, 51 and 227 sites were judged as modifiable cysteine (MC) and non-modifiable Cys (NMC) residues, respectively.

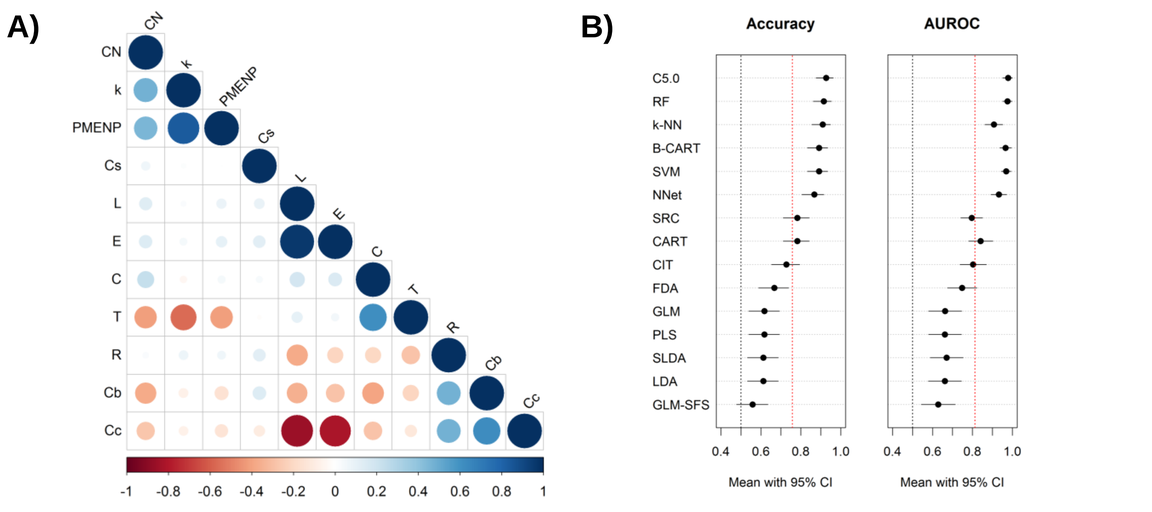

Figure 5. Prediction of Cys reactivity toward fumarate. (A) Network-based features (betweenness centrality, Cb; closeness centrality, Cc; clustering coefficient, C; degree, k; eccentricity, E; neighbourhood connectivity, NC; partner of multi-edged node pairs, PMENP; radiality, R; shortest path length, L; stress centrality, Cs; topological coefficient, T) were initially considered as potential predictors for cysteine reactivity toward fumarate. To extract potential predictors for Cys reactivity, colinearities were initially evaluated upon these features, and 3 (Cc, E and PMENP) were excluded to limit bias toward the classification resulting from using correlated predictors. (B) Network-based measures were analysed using a library of 15 models. Accuracy and Area under the ROC curve (AUROC) were determined to quantify the model performance. The average values were shown as red vertical lines. PLS: Partial Least Square; SLDA: Stabilized Linear Discriminant Analysis; LDA: Linear Discriminant Analysis; GLM-SFS: Generalized Linear Model with Stepwise Feature Selection; GLM: Generalized Linear Model; FDA: Flexible Discriminant Analysis; CART: Classification and Regression Tree; CIT: Conditional Inference Tree; SRC: Single Rule Classification; NNet: Neuronal Network; k-NN: k-Nearest Neighbors; B-CART: Bagged ClAssification and Regression Tree; SVM: Support Vector Machine; C5.0: C5.0; RF: Random Forrest.

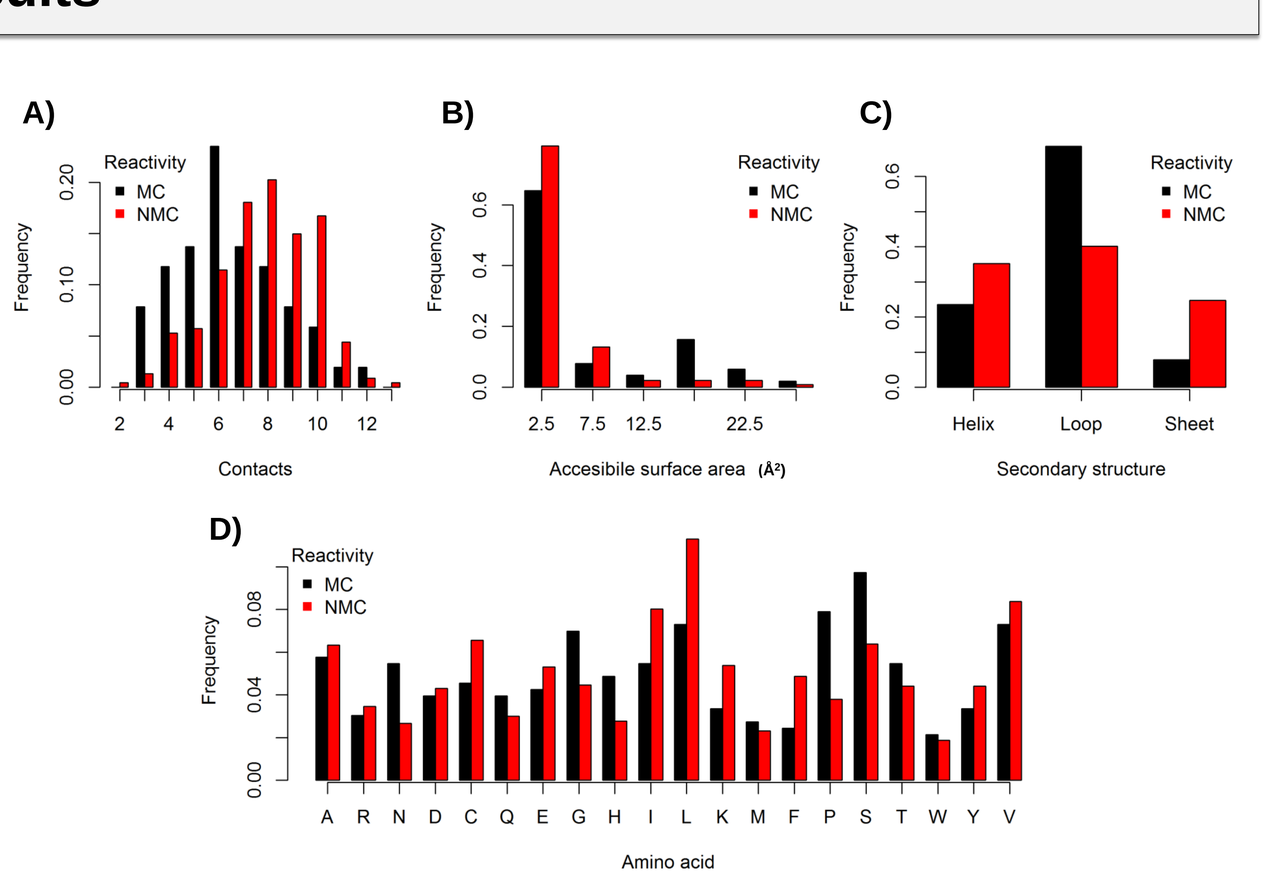

Figure 4. Comparisons between MC and NMC sites. (A) Number of Cys-interacting amino acids (contacts; $p = 5.76 \times 10^{-3}$, Pearson's χ^2 test). (B) Accessible surface area of the sulphur atom ($p = 8.84 \times 10^{-4}$, Wilcoxon rank sum test). (C) Secondary structure of the Cys-bearing peptide ($p = 6.51 \times 10^{-4}$, Pearson's χ^2 test). (D) Cysteine-interacting amino acids ($p = 2.72 \times 10^{-4}$, Pearson's χ^2 test).

Conclusions

- Significant differences between MC and NMC sites were determined when their microenvironments were compared.
- The reactivity of a Cys site toward fumarate was accurately predicted when the data for 8 RIN-based topological features were analysed.
- The adoption of concepts of network theory and machine learning could provide helpful strategies to profile a Cys site and quantify its likelihood to be modified by fumarate.

Acknowledgements

Dr Francesca Damiano and Dr Carlotta Corgiat Loia (Department of Drug Science and Technology, University of Turin) are gratefully acknowledged for their support in the method implementation. This work was supported by funding of the «Università degli Studi di Torino, Ricerca Locale Ex-60%, 2015 e 2016-2017» to GM.