Analysis of Topological Features for the Prioritization of Protein Succination

Analysis of Topological Features for the Prioritization of Protein Succination

Benedetta Ferrara[1], Gianluca Miglio[1,2]

[1]Department of Drug Science and Technology, and [2]Scientific Computing Competence Centre (C[3]S), University of Turin, Italy

E-mail: gianluca.miglio@unito.it

An aberrant adduction of fumarate to certain cysteine (Cys) residues in proteins (succination; Figure 1) has been implicated in the pathogenesis of different disorders, such as those caused by mutations in the gene encoding fumarate hydratase on chromosome 1q42 (e.g., hereditary leiomyomatosis and renal cell cancer; Yang et al., *Oncogene* 2014; 33: 2547-2556), metabolic diseases (e.g., type-2 diabetes mellitus; Frizzell et al., *Free Radic. Res.* 2011; 45: 101-109), and possibly other conditions (Piroli et al., *Mol. Cell Proteomics*. 2016; 15: 445-361). Moreover, by perturbing the intra/inter-protein signal transmission, succination could contribute to the variability in the response to drugs. In the last years, over 200 modified sites across more than 180 eukaryotic proteins have been identified experimentally. However, the biological role of protein succination still remains elusive. In addition, the need of methods that predict the functional impact of this type of post-translational modification on protein behaviour has not yet been met (Miglio et al., *Biochim. Biophys. Acta* 2016; 1864: 211-218). In this study, the predictive and diagnostic value of new computational approaches, combining concepts of network theory and machine learning, has been evaluated for the prioritization of protein succination.

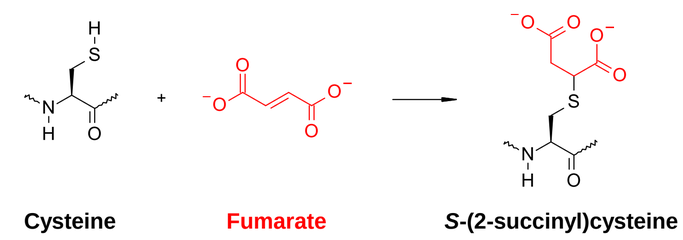
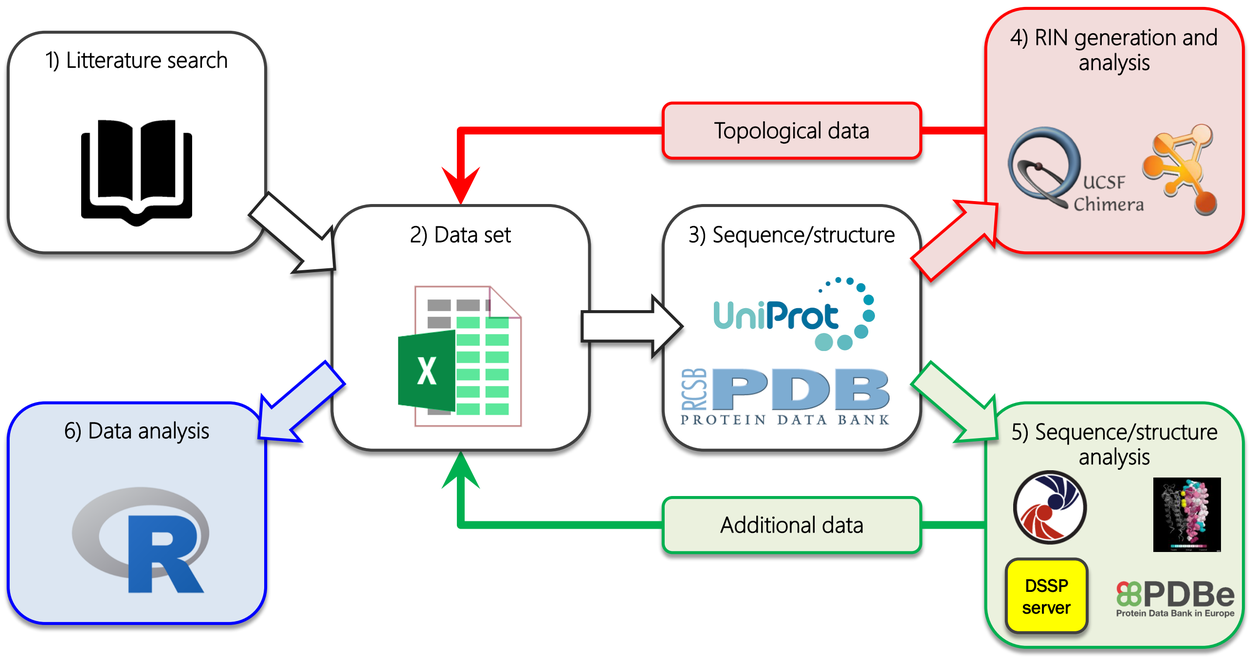


Figure 1. Reaction between fumarate and cysteine (succination). Nucleophilic adduction of cysteine to fumarate via a Michael reaction yields S-(2-succinyl)cysteine (2SC) sites.

Methods



The protocol adopted in this study starts with the selection of proteins and sites to be investigated (Steps 1-3). A dataset of fumarate-sensitive proteins was built taking advantage of large sets of eukaryotic proteins and 2SC sites experimentally verified reported in previous studies (Ternette et al. *Cell Rep.* 2013; 3: 689-700; Yang et al. *Metabolites* 2014; 4: 640-654.; Merkley et al. *Mass Spectrom. Rev.* 2014; 33: 98-109). These datasets have been combined and updated by including additional proteins and sites described elsewhere: tubulin α-1B chain, tubulin β chain (Piroli et al., *Biochem. J.* 2014; 462: 231-245), and kelch-like ECH-associated protein 1 (Adam et al., *Cancer Cell.* 2011; 20: 524-537). Then, all these proteins and sites were assessed for eligibility, which was established by adopting two criteria: 1) proteins with established crystal or NMR structure and proteins that can be modelled; 2) overall sequence homology of at least 80% (100% around the Cys residues).

Sequence and structure of the included proteins were retrieved from the UniProt (http://www.uniprot.org/) and the Protein Data Bank (PDB) repositories (http://www.rcsb.org/pdb/home/home.do). Data were generated from two workflows leading to the: generation of topological data by the analysis of the residue interaction networks using the UCSF Chimera (1.11.2) software (http://www.cgl.ucsf.edu/chimera/), converted and analysed using RINalyzer (http://www.rinalyzer.de), a Cytoscape-plugin for protein structure network assessment (Step 4); collection of additional data extracted from several web sources [PropKa (http://nbcr-222.ucsd.edu/pdb2pqr_2.0.0/); DSSP web tool (http://swift.cmbi.ru.nl/gv/dssp/); ConSurf server (http://consurf.tau.ac.il/2016/); PDBsum (http://www.ebi.ac.uk/thornton-srv/databases/cgi-bin/pdbsum/GetPage.pl?pdbcode=index.html); Steps 5]. Finally (Step 6), data were prepared, analysed and visualized using the R software (The R Project for Statistical Computing; https://www.r-project.org/).

Results

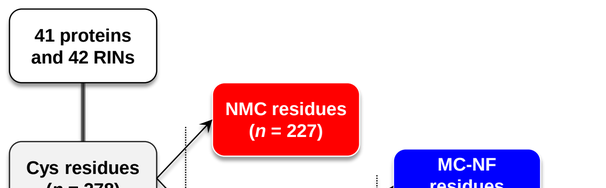


Figure 1. A total of 278 Cys residues were found in 42 RINs generated from the 41 proteins included in this study. According to their reactivity toward fumarate, 51 and 227 sites were judged as modifiable cysteine (MC) and non-modifiable Cys (NMC) residues. According to their functional roles or proximity to a functional site in the folded proteins, 22 MC residues were suggested to be functional MC residues (MC-F). The remaining MC residues were postulate to be non-functional MC (MC-NF) residues.

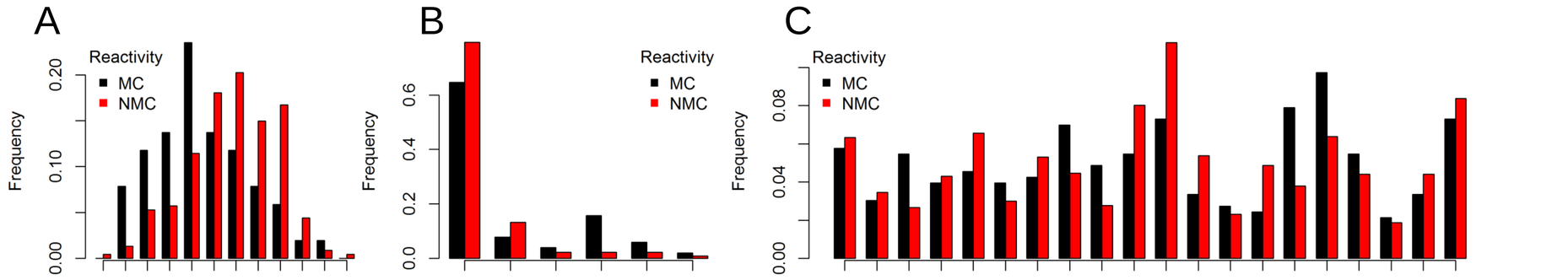


Figure 2. Characterization of the microenvironment surrounding the cysteine residues. In order to better understand the role of microenvironmental factors in determining specificity of succination in vivo, the structure and RIN of the included proteins were analysed. Data collected for three attributes were analysed and results were displayed as frequency distributions. (A) Number of cysteine-interacting amino acids (contacts; $P = 5.757 \times 10^{-3}$, Pearson' s $\chi 2$ test for the difference MC vs. NMC). (B) Accessible surface area of the sulphur atom ($P = 8.844 \times 10^{-4}$, Wilcoxon rank sum test, for the difference MC vs. NMC). (C) Cysteine-interacting amino acids ($P = 2.717 \times 10^{-4}$, Pearson' s $\chi 2$ test for the differences between MC and NMC). Collectively, these analysis demonstrate that the NMC are typically core residues, surrounded by hydrophobic amino acids. In contrast, the MC can be either buried or exposed residues often surrounded by polar or charged amino acids.

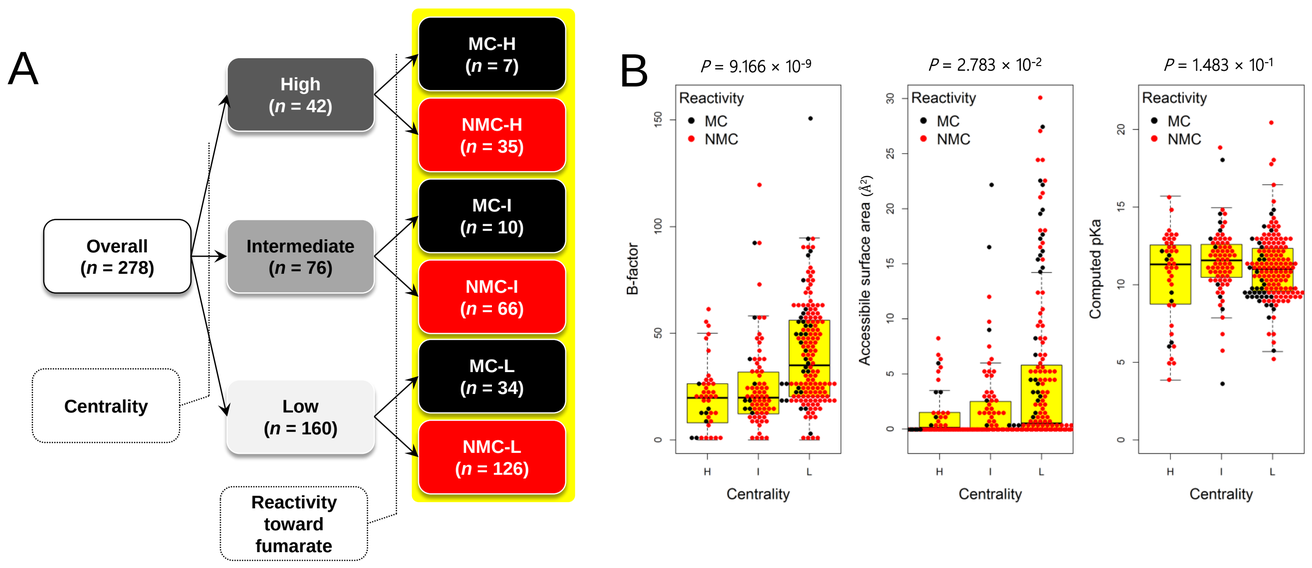


Figure 3. Centrality in the residue networks, structural/biochemical attributes and reactivity of cysteine residues. (A) According to their centrality in the protein networks (high: H vs. intermediate: I vs. low: L) and reactivity toward fumarate (modifiable cysteine: MC vs. non-modifiable cysteine: NMC), cysteine sites were divided into six subsets. (B) Data for the experimental B-factor, the accessible surface area of the sulphur atom and the acid dissociation constant (pKa) of the sulfhydryl group were analysed to evaluate possible associations (Kruskal-Wallis rank sum test).

Table 1. Distribution of the cysteine sites according to their centrality in the protein structure networks and conservation.

Set	Sites with: Intermediate-to-high centrality — Highly conserved (%)	Intermediate-to-high centrality — Poorly conserved (%)	Low Centrality — Highly conserved (%)	Low Centrality — Poorly conserved (%)	Any level of centrality — Moderately conserved (%)	*P*-value[a] vs MC-F
MC-F (*n* = 22)	10 (45.5)	0 (0.0)	3 (13.6)	7 (31.8)	2 (9.4)	NA
MC-NF (*n* = 29)	1 (3.4)	3 (10.3)	3 (10.3)	15 (51.7)	7 (24.1)	3.976×10^{-3}
NMC (*n* = 227)	44 (19.4)	32 (14.1)	20 (8.8)	60 (26.4)	71 (31.3)	9.166×10^{-3}

Highly conserved positions: conservation ≥ 75%, Poorly conserved positions: conservation ≤ 25%.

[a]Pearson' s $\chi 2$ test. NA: non-applicable.

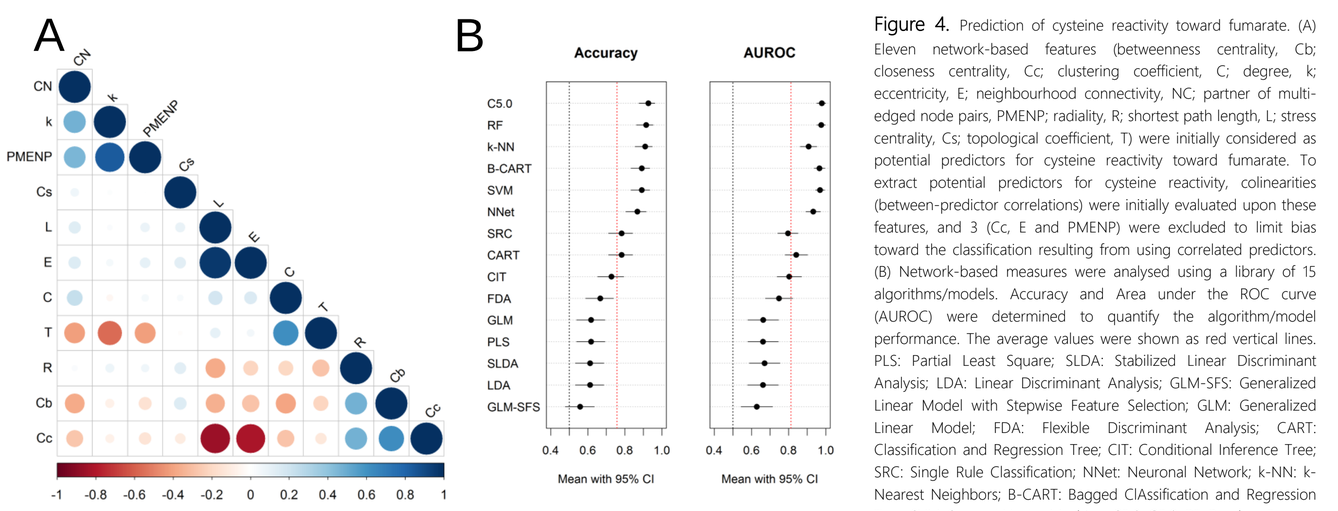


Figure 4. Prediction of cysteine reactivity toward fumarate. (A) Eleven network-based features (betweenness centrality, Cb; closeness centrality, Cc; clustering coefficient, C; degree, k; eccentricity, E; neighbourhood connectivity, NC; partner of multi-edged node pairs, PMENP; radiality, R; shortest 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 cysteine reactivity, colinearities (between-predictor correlations) 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 algorithms/models. Accuracy and Area under the ROC curve (AUROC) were determined to quantify the algorithm/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 CIAssification and Regression Tree; SVM: Support Vector Machine; C5.0: C5.0; RF: Random Forrest.

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 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.