Predicting Modifiability of Redox-Sensitive Cysteines with Supervised Machine Learning Hidden Markov Models

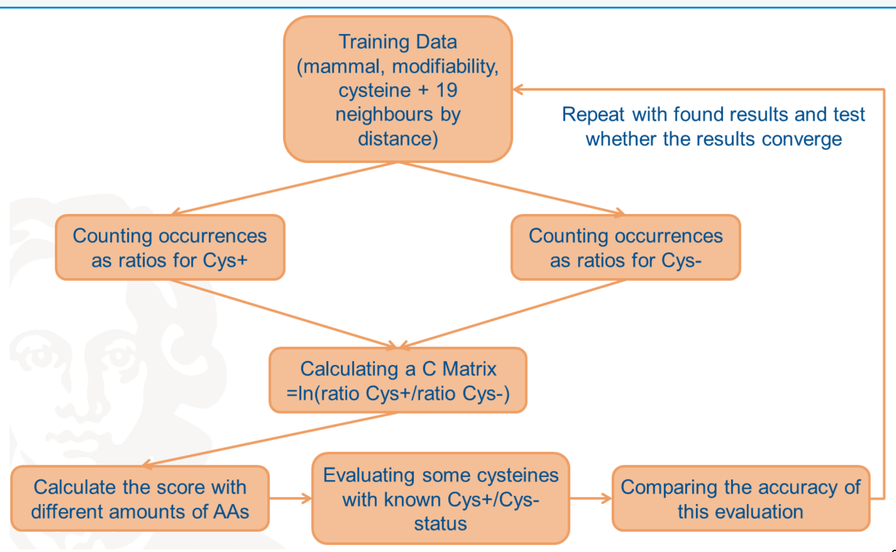
Nils Nover,[1] Marcus Keßler,[1] Jörg Ackermann,[1] and Ina Koch[1]

[1] Molecular Bioinformatics Group, Institute of Computer Science, Robert-Mayer-Str. 11-15, 60325 Frankfurt am Main, Germany

Eukaryotic proteins may undergo reversible post-translational modifications caused by redox reactions of amino acids (AAs). The most relevant redox modification may be the oxidation of cysteine to cystine causing a disulfide bridge to occur and influence the structure and function of a protein [1, 2]. High levels of reactive oxygen species (ROS), such as hydrogen peroxide H_2O_2 and superoxide O^{2-}, promote these redox reactions and consequently are able to promote the redox modification of cysteines [3]. Cysteines can be separated into those which are able to perform these modifications (Cys+) and those which are not (Cys-) [4]. Salsbury et al. (2007) proposed that certain structural properties such as neighboring AAs can be used as predictors for cysteine sensitivity, since these properties show different characteristics for Cys+ compared to Cys-. We propose a method with the ability to predict and evaluate the modifiability of redox-sensitive cysteines using supervised machine learning hidden Markov models (HMM).

Hidden Markov Model

Program Workflow



Cysteines with known modifiability and their ordered neighboring AAs (low to high distance in Å). The C Matrix shows the likelihood of a certain cysteine being either Cys+ or Cys- depending on the position.

Confusion Matrix

a)

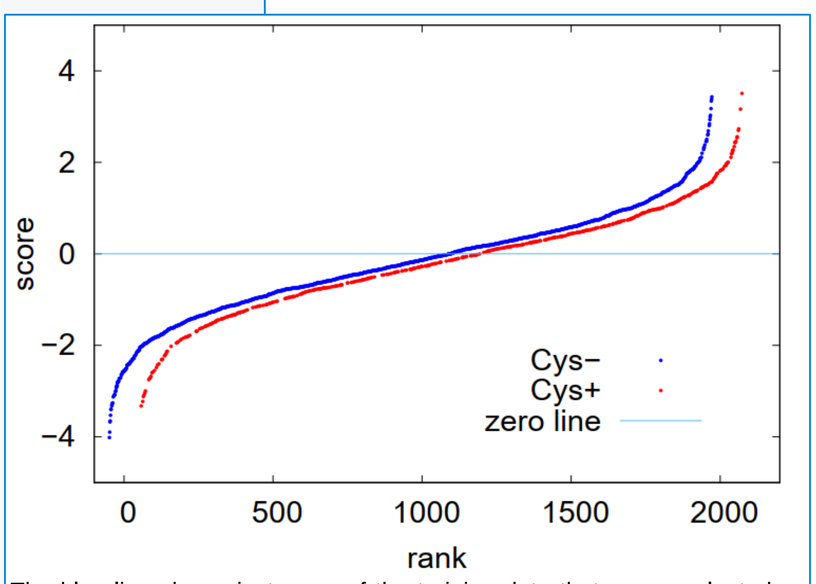
		Prediction by HMM	
		Cys+	Cys-
Input Data	Cys+	259	267
	Cys-	597	900

b)

		Prediction by HMM	
		Cys+	Cys-
Input Data	Cys+	12.80%	13.20%
	Cys-	29.51%	44.49%

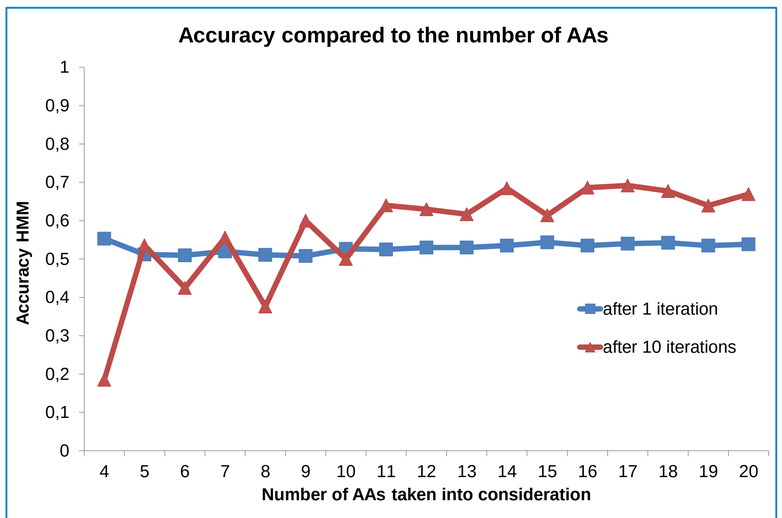
The confusion matrix a) shows the amount of Cys+ and Cys- predicted by the HMM tool. b) shows the predictions in percent for a total of 2023 cysteines. The accuracy of this model is 0.57. This means, 57% of the predictions are right. The matrix shows which predictions were made.

Data Visualization: scores



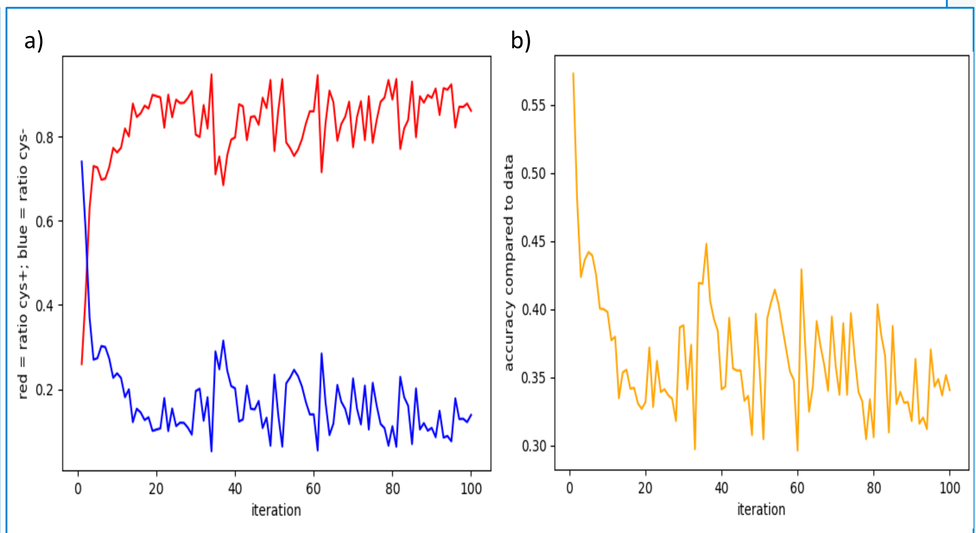
The blue line shows instances of the training data that were evaluated as Cys- and the red dots as Cys+. Positive scores indicate a Cys+ occurrence while low scores tend to be Cys-.

Calculating with the right amount of AAs



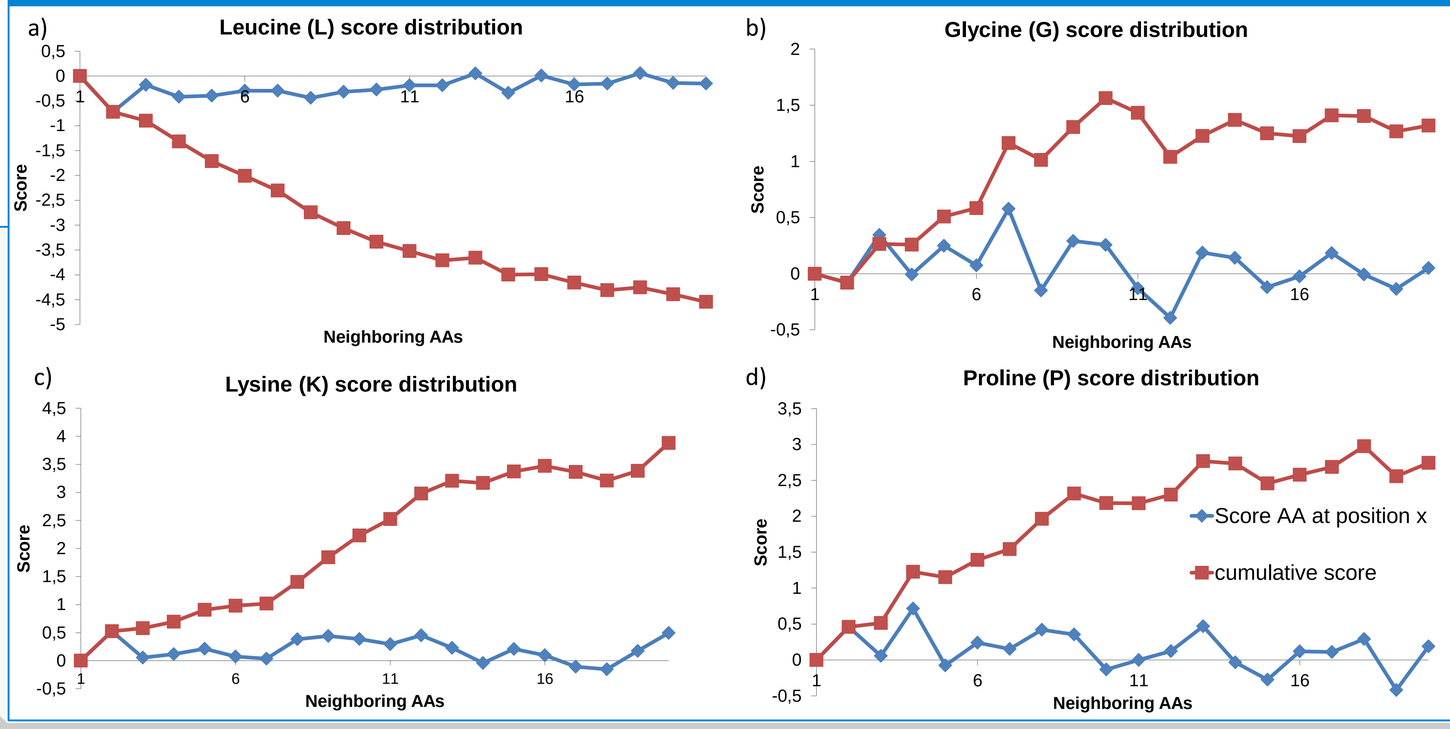
The accuracy of the HMM compared to the number of AAs used for the calculation. The accuracy can not be improved by taking more AAs into consideration for the HMM model. For further calculation we will continue to use all 19 neighboring AAs.

Using the output data to repeat the HMM



The HMM can be performed iteratively when the predictions are used to overwrite the existing data. a) shows the distribution of Cys+ to Cys- during 100 iterations. b) shows how the accuracy compared to the input data evolves over the iterations. This approach is not pursued any further as the accuracy is very low compared to the input data.

Discussion



- Prior research has shown a high relative frequency of leucine (a) in the neighborhood of Cys- and a high relative frequency of lysine (b), glycine (c), proline (d) and other AAs in the neighborhood of Cys+
- (Presentation “Prediction and Analysis of Redox-Sensitive Cysteines Using Machine Learning and Statistical Methods” by Marcus Keßler is recommended)

Conclusion:

- Low accuracy to the input data
- The findings conform that the frequency of certain AAs in the neighborhood of cysteines indicate the modifiability of them.
- For a higher accuracy in the HMM, it is needed to have a relatively high frequency in certain neighborhood positions. The findings however imply that the frequency are almost evenly distributed over all neighboring positions.

References

[1] Paul D. Ray, Bo-Wen Huang, and Yoshiaki Tsuji. Reactive oxygen species (ROS) homeostasis and redox regulation in cellular signaling. Cellular Signalling, 24, 2012.

[2] Matsumura M, Matthews BW. Stabilization of functional proteins by introduction of multiple disulfide bonds. Methods Enzymol. 1991;202:336-356. doi:10.1016/0076-6879(91)02018-5

[3] Cross CE, Halliwell B, Borish ET, et al. Oxygen radicals and human disease. Ann Intern Med. 1987;107(4):526-545. doi:10.7326/0003-4819-107-4-526

[4] Salsbury FR Jr, Knutson ST, Poole LB, Fetrow JS. Functional site profiling and electrostatic analysis of cysteines modifiable to cysteine sulfenic acid. Protein Sci. 2008;17(2):299-312. doi:10.1110/ps.073096508

[5] Keßler et al. Prediction and Analysis of Redox-Sensitive Cysteines Using Machine Learning and Statistical Methods, to appear, 2020