FEBS Open Bio 4 (2014) 310–314

journal homepage: www.elsevier.com/locate/febsopenbio

A word of caution about biological inference – Revisiting cysteine covalent state predictions

Éva Tüdős [a], Bálint Mészáros [a], András Fiser [b,c], István Simon [a,*]

[a] Institute of Enzymology, Research Centre for Natural Sciences, Hungarian Academy of Sciences, P.O. Box 286, H-1519 Budapest, Hungary
[b] Department of Systems and Computational Biology, Albert Einstein College of Medicine, 1300 Morris Park Ave, Bronx, NY 10461, USA
[c] Department of Biochemistry, Albert Einstein College of Medicine, 1300 Morris Park Ave, Bronx, NY 10461, USA

ARTICLE INFO

Article history:
Received 25 November 2013
Revised 5 March 2014
Accepted 7 March 2014

Keywords:
Protein prediction
Cysteine redox state
Protein structure
Prediction accuracies
Biological inference

ABSTRACT

The success of methods for predicting the redox state of cysteine residues from the sequence environment seemed to validate the basic assumption that this state is mainly determined locally. However, the accuracy of predictions on randomized sequences or of non-cysteine residues remained high, suggesting that these predictions rather capture global features of proteins such as subcellular localization, which depends on composition. This illustrates that even high prediction accuracy is insufficient to validate implicit assumptions about a biological phenomenon. Correctly identifying the relevant underlying biochemical reasons for the success of a method is essential to gain proper biological insights and develop more accurate and novel bioinformatics tools.

1. Introduction

The benefits of protein sequence based computational prediction methods are usually twofold. On one hand, they offer a fast and inexpensive way to obtain new biological information about a protein, which then can be used to design follow-up experiments or to explain existing experimental observations. On the other hand, the success of these prediction algorithms is generally considered to validate the underlying hypothesis about principles governing structure formation or biochemical role of the feature being predicted. For instance, the success of early secondary structure prediction methods in the sixties and seventies, based on α-helix and β-strand forming preferences of individual residues indicated that most of the information about the preferred secondary structure of any segment of the protein is encoded in the local sequence itself [1,2]. Although the development of secondary structure predictions encompassed many years, even the most advanced secondary structure prediction algorithms, reaching as high as 80% accuracy, are able to do so by using local sequence information only [3,4]. In this case the basic assumption (the dominance of local effects in secondary structure formation) is in fact in agreement with both bioinformatics and experimental results. The idea itself predated bioinformatics analysis of three dimensional structures and was later validated by the appropriate calculations, confirming that long range interactions indeed played only a secondary role [5,6]. Experimental investigations also demonstrated that isolated segments have a tendency to prefer conformations similar to the one in folded structures [7].

However, in case of other bioinformatics methods the relationship between a successful prediction and the underlying hypothesis may be less trivial. Highly specific residue-level structural or biochemical features are predicted from local sequence patterns; however it is usually not elaborated why certain features are supposed to be encoded in the immediate sequence environment of residues. Here, as a case study, we focus on the development of a prediction method capturing the covalent state of cysteine (Cys) residues from the amino acid sequence. Despite being a highly specific area of bioinformatics research, it can serve with general conclusions about biological inference and the potential pitfalls of interpreting the success of a prediction method as a verification of underlying assumptions.

As protein sequences are obtained primarily by genome sequencing, the location of post-translational modifications – such as disulfide bonds – is usually unknown for the majority of

Abbreviations: Acc, prediction accuracies; FN, false negatives; FP, false positives; PDB, Protein Data Bank; TN, true negatives; TP, true positives

* Corresponding author. Tel.: +36 1 3826295; fax: +36 1 3826710.
E-mail addresses: tudos.eva@ttk.mta.hu (É. Tüdős), meszaros.balint@ttk.mta.hu (B. Mészáros), andras.fiser@einstein.yu.edu (A. Fiser), simon.istvan@ttk.mta.hu (I. Simon).

http://dx.doi.org/10.1016/j.fob.2014.03.003
2211-5463/

proteins. Protein structure determination would clearly benefit from the knowledge of the oxidation state of various Cys residues, e.g. if a certain Cys does or does not form a disulfide bond [8–11]. Furthermore, disulfide bond connectivity patterns can be used to discriminate between protein folds and can facilitate the accurate superimposition of protein structures [10]. The underlying assumption in these methods is that similar disulfide bond connectivity patterns place similar spatial constraints on proteins, resulting in similar protein structures. Meanwhile, variation in disulfide bridge patterns within the same superfamily may be used to infer variation of protein function. Information on disulfide formation can be incorporated to enhance molecular simulations of folding [12]. Due to all these critical roles that disulfide bonds play in proteins, there is a long standing interest to predict Cys residues that can form disulfide bonds, or more generally, predict which Cys residues are oxidized and bound to another Cys residue (or to other factors) and which Cys are reduced and have a free, highly reactive thiol group.

During the 1980's the sequence databases had become large enough to enable the recognition of the non-random pairing of residues within their sequential vicinity [13,14]. Early prediction methods were based on the observed different sequence environments of cysteines and half-cystines of the disulfide bonds [15]. The success of a neural network based approach considering the sequence environments of cysteines and half-cystines [16] also suggested that most of the information on the preferred oxidation state of Cys residues must be encoded in the surrounding sequence segments. This hypothesis formed the basis of a prediction approach, where a disulfide bond forming statistical potential was calculated for Cys residues from the relative frequency of residues in the surrounding decapeptides [15]. This method – performing with 71% accuracy – was developed nearly two decades ago and according to citation data of the paper it is still in use; although since then more accurate, albeit more complicated methods have been developed that use a wide range of other input data in addition to the protein sequence [11] (reaching 82% accuracy, e.g. [17]). Furthermore, the incorporation of mass spectrometry data into S–S bond determination methods (see [18–20]) has also significantly improved their accuracy, even surpassing the efficiency of sequence based methods, albeit on smaller datasets [21]. The primary source of information for this study was the Protein Data Bank (PDB [22]), which is approximately 40 times larger now than it was at the time. Hence, it seems plausible that the re-parametrization of the method would significantly increase the accuracy.

By now it is clear that due to the high glutathione concentration in the cytoplasm (due to the constituently active glutathione reductase enzyme), and the oxidative environment in the extracellular matrix, almost all Cys residues in the cytoplasm (intracellular Cys residues) are in the free, thiol form, while almost all Cys residues outside the cytoplasm (in the extracellular matrix or in the various compartments of the cell, extracellular Cys residues) are either in disulfide bonds or in a liganded form [23]. However, it has also become known that the amino acid compositions of intracellular and extracellular proteins are significantly different [17,24–26]. Therefore one can present an alternative hypothesis, according to which the 20 residues in the flanking decapeptides of Cys residues might only identify the extracellular or intracellular nature of the entire protein and the oxidation state of the Cys residues is simply the consequence of the subcellular localization. This idea is supported by the fact that the inclusion of subcellular localization in the prediction algorithms improve their accuracy only by a few points [27] indicating that this information is already largely encoded in other input data (e.g. the sequence).

It is rather common that prediction methods are used for a different purpose than what they were developed for. For example, the above mentioned Cys prediction method [15] was successfully used to design a site directed mutagenesis experiment, to predict where a Cys residue needs to be inserted into a sequence to have a larger probability of forming a disulfide-bond [28] or to find out which of the designed Cys residues form a disulfide bond with a sequentially distant Cys residue [29]. However, in theory, for this type of predictions the method can only be meaningful if there is a significant direct influence of the local sequence environment on the disulfide forming potential.

In this study we explore the origin of the potential of a Cys to exist in a disulfide bond/liganded (bound) form and in thiol (free) form, as calculated from the signal of characteristic local sequence patterns in the neighborhood of Cys residues. Based on these results, we aim to assess the correctness of considering the success of the developed method as a proof of the idea that Cys oxidation state is directly encoded in the local sequence. In order to do this, the original prediction method was re-implemented in this study using the current, 40 fold larger PDB database. We also revisited earlier reports that claim that certain relative sequential positions around Cys residues play specific roles in determining its oxidation state. Subsequently, the method implementation was repeated using shuffled protein sequences to remove all local sequence information. Finally, the specificity of the prediction algorithm was further challenged by implementing it for all the 20 residue types, not only for Cys. The results can either reinforce the view that Cys redox state is mainly determined by local effects or can show the non-causal relationship between this underlying view and prediction accuracy, serving as a general warning about biological inferences concerning bioinformatics methods.

2. Datasets and methods

2.1. Real protein sequence dataset

Entries in Protein Data Bank (PDB [22]) were filtered at 25% sequence identity level using BlastClust (http://www.ncbi.nlm.nih.gov/blast/Blast/) to remove redundant sequences. The resulting dataset contains 3881 polypeptide chains with 19202 Cys residues. Out of these sequences 569, 2039 and 132 contain half-cystines, free cysteines and both half-cystines and free cysteins, respectively. Using a more general classification we also discriminate between oxidized Cys that are bound to either another Cys residue in a disulfide bond or to ligands and free Cys that are not bound. In this classification there are 693 proteins containing bound Cys only, 1804 of them contain only free cysteins and 243 contain both. Our dataset also contains 1117 polypeptide chains that do not contain any Cys residues at all and 24 chains were discarded as the oxidation state of Cys residues could not be determined. According to the UniProt annotations [30] (http://www.uniprot.org/) the 3881 polypeptide chains can be split into three groups: 1783 intracellular, 1024 extracellular and 910 transmembrane proteins. There are 9515 half-cystines and 9687 free cysteines or – using the alternative definition – there are 10996 Cys residues in some sort of bound state (oxidized) and 8206 in free thiol (reduced) state.

2.2. Random protein sequence dataset

Randomly mixed sequences were obtained by shuffling the order of residues within each protein from the *Real protein sequence dataset* individually.

2.3. Disulfide bond forming potentials

First position specific residue preference matrices were calculated. For each position in a ±10 residue window centered on Cys residues the occurring amino acids were counted. Next,