BIOINFORMATICS ORIGINAL PAPER

Vol. 27 no. 16 2011, pages 2224–2230
doi:10.1093/bioinformatics/btr387

Structural bioinformatics

Advance Access publication June 29, 2011

Improving the prediction of disulfide bonds in Eukaryotes with machine learning methods and protein subcellular localization

Castrense Savojardo[1,2], Piero Fariselli[1,2,*], Monther Alhamdoosh[1], Pier Luigi Martelli[1], Andrea Pierleoni[3] and Rita Casadio[1]

[1]Biocomputing Group, University of Bologna, CIRI-Life Science and Health Technologies and Department of Biology, Via San Giacomo 9/2, Bologna, [2]Department of Computer Science, Via Mura Anteo Zamboni 7, 40127 Bologna and [3]Externautics s.p.a., Department of Bioinformatics, Via Fiorentina 1, 53100 Siena, Italy

Associate Editor: Burkhard Rost

ABSTRACT

Motivation: Disulfide bonds stabilize protein structures and play relevant roles in their functions. Their formation requires an oxidizing environment and their stability is consequently depending on the redox ambient potential, which may differ according to the subcellular compartment. Several methods are available to predict cysteine-bonding state and connectivity patterns. However, none of them takes into consideration the relevance of protein subcellular localization.

Results: Here we develop DISLOCATE, a two-step method based on machine learning models for predicting both the bonding state and the connectivity patterns of cysteine residues in a protein chain. We find that the inclusion of protein subcellular localization improves the performance of these predictive steps by 3 and 2 percentage points, respectively. When compared with previously developed methods for predicting disulfide bonds from sequence, DISLOCATE improves the overall performance by more than 10 percentage points.

Availability: The method and the dataset are available at the Web page http://www.biocomp.unibo.it/savojard/Dislocate.html. GRHCRF code is available at http://www.biocomp.unibo.it/savojard/biocrf.html.

Contact: piero.fariselli@unibo.it

Received on March 21, 2011; revised on May 26, 2011; accepted on June 21, 2011

1 INTRODUCTION

The formation of disulfide bonds between cysteine residues is essential for folding, stability and maturation of many proteins (Inaba, 2010). Predicting which cysteines in a protein sequence form disulfide bonds plays a relevant role in protein structural and functional annotation (Singh, 2008; Tsai *et al.*, 2007). Several computational methods are available, which can be grouped as: (i) methods that predict the disulfide-bonding state (Chen *et al.*, 2004; Martelli *et al.*, 2002; Mucchielli-Giorgi *et al.*, 2002; Savojardo *et al.*, 2011); (ii) methods that predict the connectivity patterns, assuming that the cysteine-bonding state is known (Fariselli and Casadio, 2001; Ferrè and Clote, 2005; Song *et al.*, 2007; Vullo and Frasconi, 2004); (iii) methods that compute both features (Cheng *et al.*, 2006; Taskar *et al.*, 2005; Vincent *et al.*, 2008).

Proteins that contain disulfide bonds are rarely found in the cytoplasm and are routinely secreted (Kadokura *et al.*, 2003). Disulfide bond formation in Eukaryotes happens in the lumen of the endoplasmic reticulum (Heras *et al.*, 2007; Sevier *et al.*, 2007). These experimental studies show that the localization in the different cell compartments plays a relevant role in disulfide bond generation and stabilization. However, to the best of our knowledge none of the methods developed so far has actually exploited information on the subcellular localization of proteins. Protein datasets with experimentally known subcellular localization are available as well. Several efficient prediction methods were developed to predict subcellular localization (Casadio *et al.*, 2008; Imai and Nakai, 2010). Here, we propose DISLOCATE, a novel two-stage method for disulfide bond prediction in Eukaryotes based on machine learning approaches. We show that the inclusion of protein subcellular localization improves the performance of disulfide bond prediction methods. This improvement is noticeable also when the subcellular localization is predicted with BaCelLo (Pierleoni *et al.*, 2006).

2 MATERIAL AND METHODS

2.1 Datasets

From PDB (release May 2010), we extracted 1797 eukaryotic protein structures with resolution <2.5 Å with at least two cysteine residues and global pairwise sequence similarity <25%. We refer to this dataset as PDBCYS: it includes 7619 free and 3194 bonded cysteines. Since the selected proteins contained some measure of sequence similarity, we clustered the remaining chains using a local sequence similarity score. First, we ran a BLAST sequence search using all the proteins of the set versus themselves. Then, for each pair of proteins we selected the higher bidirectional (say p1 versus p2 or p2 versus p1) sequence identity as reported in the BLAST output. We subsequently treated the proteins as a node of a graph and assigned an edge between two nodes only where local sequence identity between the corresponding protein sequences was >25%. In addition, we computed the connected components of the graph and treated each group of nodes as a protein cluster. Finally, the clusters were grouped in 20 disjoint sets used to train and test the method. For sake of comparison, we also adopted the same procedure on the SPX- set (Cheng *et al.*, 2006).

2.2 PDBCYS subcellular localization

For each protein in PDBCYS, we extracted from the corresponding UniProt file the annotated subcellular localization, considering five different macro compartments: chloroplast, cytoplasm, mitochondrion, nucleus and secreted.

[*] To whom correspondence should be addressed.

Table 1. Subcellular localization of protein chains containing bonded and free cysteines

Localization	Bonded cysteines (%)	Free cysteines (%)	Number of proteins
Chloroplast	11	89	28
Cytoplasm	9	91	472
Mitochondrion	2	98	62
Nucleus	5	95	322
Secreted	79	21	227

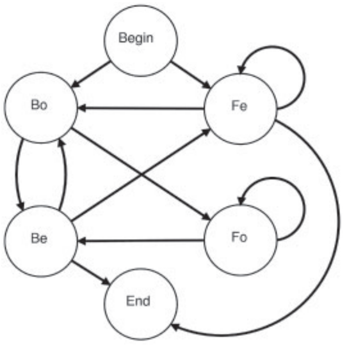

Fig. 1. Automaton adopted to define the cysteine grammar in protein sequences. States Bo and Be define bonding labels while states Fo and Fe indicate free cysteine labels.

The distribution of the different proteins in the various compartments is reported in Table 1. The 62% of the PDBCYS proteins (1121 chains including 4894 free and 1598 bonded cysteines) is endowed with subcellular localization.

To predict subcellular localization, we adopted a cross-validation version of BaCelLo (Pierleoni *et al.*, 2006) also aiming at preventing an overestimation of the predictive contribution.

2.3 Predicting disulfide-bonding state

Disulfide-bonding state of cysteines is predicted with Grammatical-Restrained Hidden Conditional Random Fields (GRHCRFs). GRHRCFs have been recently introduced as a promising framework for solving sequence labeling tasks (Fariselli *et al.*, 2009). Here, for the sake of clarity, we introduce GRHCRFs starting from linear conditional random fields (CRFs). Linear CRFs can also be seen as discriminative versions of Hidden Markov Models (HMMs) and we will describe them applying this approach (Lafferty *et al.*, 2001; Sutton and McCallum, 2007).

2.3.1 From HMMs to CRFs. A HMM is defined by its transition ($a_{s,t}$) and emission ($e_s(c)$) probabilities (Durbin *et al.*, 1998). Given an observed sequence of symbols $X=\{x_1,\dots,x_L\}$ and a sequence of states $Y=\{y_1,\dots,y_L\}$, the joint probability of X and Y can be computed by the HMM as:

$$p(Y,X)=\prod_{j=1}^{L} a_{y_{j-1},y_j} e_{y_j}(x_j) \tag{1}$$

where the variable j runs over the length of the sequence X and an explicit begin state (y_0) is indicated by the index of position 0. Here, we adopt the convention of using uppercase letters for the entire sequences (X and Y) and lowercase letters for the single elements (x_i and y_j). The Equation (1) can be rewritten in an exponential form by introducing new variables: $\tau_{s,t}=\log(a_{s,t})$ and $\mu_s(c)=\log(e_s(c))$:

$$p(Y,X)=\prod_{j=1}^{L}\exp\left[\tau_{y_{j-1},y_j}+\mu_{y_j}(x_j)\right] \tag{2}$$

Taking into consideration Kroneker's deltas ($\delta(a,b)=1$ if $a=b,\ 0$ otherwise), $p(Y,X)$ can be rewritten as follows:

$$p(Y,X)=\prod_{j=1}^{L}\exp\left[\sum_{s,t}\tau_{s,t}f_{s,t}(y_{j-1},y_j)+\sum_{s,c}\mu_s(c)g_{s,c}(y_t,x_j)\right] \tag{3}$$

where $f_{s,t}(y_{j-1},y_j)=\delta(s,y_{j-1})\delta(t,y_j)$ and $g_{s,c}(y_j,x_j)=\delta(s,y_j)\delta(c,x_j)$ are feature functions, respectively, defined over (state, state) and (state, symbol) pairs.

A step forward to the CRF is to '*relax*' the assumption that $\tau_{s,t}$ and $\mu_s(c)$ are log-probabilities, assigning them arbitrary values. However, in order to maintain the meaning of joint probability, a '*global*' normalization factor (Z) is needed, such as:

$$\begin{aligned} p(Y,X) &= \frac{\prod_{j=1}^{L}\exp\left[\sum_{s,t}\tau_{s,t}f_{s,t}(y_{j-1},y_j)+\sum_{s,c}\mu_s(c)g_{s,c}(y_t,x_j)\right]}{Z} \\ &= \frac{1}{Z}\prod_{j=1}^{L}\psi_j(y_{j-1},y_j,x_j) \end{aligned} \tag{4}$$

The notation is simplified by introducing the so-called potential functions ψ_j (Lafferty *et al.*, 2001). In spite of the additional flexibility of $\tau_{s,t}$ and $\mu_s(c)$, it can be shown that $p(Y,X)$ describes exactly the HMM class (Sutton and MacCallum, 2007). Generative models as these model both the sequence of states Y and the observed sequence of symbols X. The last step toward linear CRFs is to write the conditional distribution $p(Y|X)$ using the previous definition of $p(Y,X)$ as:

$$\begin{aligned} p(Y|X) &= \frac{p(Y,X)}{\sum_{Y'}p(Y',X)} \\ &= \frac{\prod_{j=1}^{L}\psi_j(y_{j-1},y_j,x_j)}{\sum_{Y'}\prod_{j=1}^{L}\psi_j(y_{j-1},y_j,x_j)} \\ &= \frac{1}{Z(X)}\prod_{j=1}^{L}\psi_j(y_{j-1},y_j,x_j) \end{aligned} \tag{5}$$

where the normalization factor over all possible sequences of labels Y' is usually referred to as a partition function $Z(X)$. The discriminative nature of CRFs (as the conditional probability $p(Y|X)$ is directly modeled) offers several advantages over generative approaches such as HMMs, including the relaxation of the strong independence assumptions implied in HMMs (Fariselli *et al.*, 2009; Lafferty *et al.*, 2001). Finally, linear CRFs can further relax the definition of the functions making them dependent on the sequence $X:\psi_j(y_{j-1},y_j,X)$. With this notation, for instance, we can have feature functions that take into consideration a window around the j-th position or global sequence descriptors such as the subcellular localization.

2.3.2 From CRFs to GRHCRFs. One of the problems with linear CRFs is the fact that the set of observed labels $\{y_j\}$ coincides with the set of states $\{s_j\}$. However, in order to identify biological meaningful predictions, the observed sequence of label Y can be considered as generated by an automaton with several different states, sharing the same type of label. For instance, the automaton presented in Figure 1 defines the simplest disulfide grammar with four states, but the observed sequences of labels contain only two symbols: free (F) and bonded (B). This makes the model cumbersome to treat because, in order to map the automaton into the observed sequence, a new artificial (and unambiguous) relabeling of the observed sequence has to be created (for instance, the sequence of observed labels Y = FBFB must be converted in Fe, Bo, Fo, Be). Any time that a new model is tested, a new artificial sequence of labels must be generated. Furthermore, grammatical rules such as forbidden transitions must be learned from the examples. This may lead to erroneous predictions if the rules are not sufficiently represented into the training examples. Alternatively, the rules can be hard-coded in the source code (and have to be consequently adapted when the grammar changes) by setting the corresponding transitions to $-\infty$.

To overcome these limitations, we introduced the GRHCRFs (Fariselli *et al.*, 2009). GRHCRFs decouples the observed sequence of labels from the set of states introducing a hidden set of variables. Like HMMs, GRHCRFs can be represented through a finite state machine (FSM) with some missing transitions between states. The structure of the FSM is determined by the specific grammar used for the problem at hand. In order to better generalize, GRHCRFs (as well as HMMs) define a one-to-many mapping between labels and FSM states and, at the same time, restrict the accepted predictions to only those that correspond to an allowed path in the FSM. A function $\Lambda(s)=y$, is defined to map each state s to a given observed label y. The potential functions ϕ_j for each sequence position j are defined as:

$$\phi_j(s_{j-1},s_j,y_j,X)=\psi_j(y_{j-1},y_j,X)\Gamma\left(s_{j-1},s_j\right)\,\Omega(s_j,y_j) \tag{6}$$

The ϕ_j are defined similarly to the CRF potential functions with the added constraints:

$$\Gamma(s,t)=\begin{cases}1 & \text{if } (s,t) \text{is a valid transition}\\ 0 & \text{otherwise}\end{cases} \tag{7}$$

$$\Omega(s,y)=\begin{cases}1 & \text{if } \Lambda(s)=y\\ 0 & \text{otherwise}\end{cases} \tag{8}$$

that ensure that the only valid path of the FSM is considered. The probability of a sequence of labels Y given an observation sequence X is obtained as:

$$p(Y|X)=\frac{Z(Y,X)}{Z(X)} \tag{9}$$

where $Z(Y,X)$ and $Z(X)$ are normalization factors defined as:

$$Z(Y,X)=\sum_{s}\prod_{j=1}^{L}\phi_j(s_{j-1},s_j,y_j,X) \tag{10}$$

$$Z(X)=\sum_{Y}Z(Y,X) \tag{11}$$

that can be computed using the forward–backward procedure (Fariselli *et al.*, 2009).

The model parameters $\Theta=\{\theta_k\}=\{\tau_{s,t},\mu_{s,\sigma}(c)\}$ associated to each feature are learned by maximizing log-likelihood over training data $D=\{(X^{(i)},Y^{(i)})|i=1,\dots,N\}$:

$$\begin{aligned}\ell(\Theta,D)&=\log\prod_{i=1}^{N}p(Y^{(i)}|X^{(i)};\,\Theta)-\sum_{k}\frac{\theta_k^2}{2\sigma^2}\\ &=\sum_{i=1}^{N}\log Z(Y^{(i)},X^{(i)})-\log Z(X^{(i)})-\sum_{k}\frac{\theta_k^2}{2\sigma^2}\end{aligned} \tag{12}$$

where the last term is a Gaussian prior regularizer (Lafferty *et al.*, 2001). The maximization is carried out using the Limited memory Broyden–Fletcher–Goldfarb–Shanno (L-BFGS) quasi-Newton optimization algorithm (Byrd *et al.*, 1995).

As far as expressiveness is concerned, linear CRFs and GRHCRFs are in many instances theoretically equivalent. However, the decoupling between states and observed labels allows the GRHCRFs to model ambiguous conditions: the sequence of observed labels can be associated with several different paths on the FSA. In these cases, GRHCRFs can exploit the ambiguity summing over all possible solutions and obtaining better performance (Fariselli *et al.*, 2009). However, in the case of the FSA of Figure 1, both models are theoretically similar, and GRHCRFs collapses to '*constrained CRFs*'. Nonetheless, GRHCRFs are simpler to deal with when grammar rules are introduced.

2.3.3 Bonding state prediction with GRHCRFs. For the bonding state prediction, we adopted the automaton described in Figure 1. The arrows represent the allowed transitions, while the B and F circles, respectively, represent the bonding and non-bonding cysteine states. The labels 'e' (even) and 'o' (odd) indicate the number of cysteines in the bonding state so far processed. The path can end only from an e-label state. This guarantees that only correct even predictions are assigned when considering intra-chain disulfide bonds.

To assign the bonding state, we encoded each cysteine with a 'local vector' representing the sequence nearest neighborhood. The vector is computed starting from the Position Specific Scoring Matrix (PSSM) as internally computed by PSI-BLAST using BLOSUM62 (Altschul *et al.*, 1997). The input vector represents each cysteine in the protein sequence and its neighborhoods, by defining a window of size $w=2k+1$ centered on each cysteine. The encoding vector consists in $20\times w$ components, where 20 is the number of residue types. We supplemented the local encoding (PSSM) with the piece of information provided by the subcellular localization (PSSM + SL) as obtained by the BaCelLo predictor in cross-validation.

On a more practical level, for each cysteine in position j and for each state s, the GRHCRFs defines the following state features:

- $g_s(x_{j+k}[a])$ for k in $\{-w/2,\dots,w/2\}$ and a in $\{$Residue Alphabet$\}$ and
- $g_s(o)$ for o in $\{$Global Features$\}$

where w is the width of the window around the cysteine of position j, and the set of global features can be 1 or 0 depending on the different types of subcellular localizations. The functions $g_s(x_{j+k}[a])$ are weighted by the corresponding position $(j+k)$ and residue type (a) values extracted from the PSSM.

2.3.4 Details on the employment of BaCelLo. The predictions provided by the BaCelLo have been integrated into the input vector of our method. As training dataset for BaCelLo, we employed its original training set described in Pierleoni *et al.* (2006). In order to make this procedure as fair as possible, we proceeded as follows: each protein in our dataset has been aligned using BLAST against the BaCelLo training set and the relevant hits with an $e<$ 1e-3 for that protein have been identified. Then, the hits found have been removed from the BaCelLo training set and the predictor has been retrained on this reduced set. Finally, the protein subcellular localization has been predicted using the re-trained BaCelLo predictor. This guarantees that each protein has been processed with a training set that does not contain homologous proteins.

2.4 Predicting connectivity patterns

Once the cysteine-bonding state is assigned, we predict the connectivity pattern of the subsets of proteins that contain at least a pair of cysteines in the bonding state. The connectivity pattern is assigned by applying a support vector regression (SVR) approach (similarly to Song *et al.*, 2007). The SVR predictions of each possible pair of cysteines is used as edge weight and the Edmond–Gabow algorithm is adopted to predict the most probable disulfide pattern (Fariselli and Casadio, 2001).

In order to evaluate SVR, we use the same 20-fold cross-validation procedure described above, considering only proteins with at least two disulfide bridges. SVRs were trained using an input encoding based on global and local information. The global information (that does not depend on each particular cysteine pair) is defined by the Normalized Protein Length (one real value), the Protein Molecular Weight (one real value) and the protein amino acid composition (20 real values). The local pairwise encoding (that depends on each particular cysteine pair) consists of the following descriptors:

- two PSSM-based windows centered into the cysteines forming the pairs. We used a window of length 13, the one that performed better among the several different-size windows we tested. With this choice, we ended up with a vector of $13\times 20\times 2=520$ components;
- the relative order of the cysteines. This feature is encoded with 2 real values that represent the normalized relative order of a cysteines pair.

Given a protein with n cysteines $(C_1, C_2, \ldots, C_n)$, the corresponding normalized ordered list of cysteines is given by $(1/n,\ 2/n,\ \ldots,\ n/n)$. For each pair of cysteines, the corresponding values are then taken from the list (e.g. the pair (C_1, C_4) is encoded as $(1/n, 4/n)$);

- the cysteine separation distance. This feature is encoded with 1 real value that represents the log-cysteine sequence separation computed as $\mathrm{SEP}(C_i, C_j) = \log(|j-i|)$ where i and j are sequence positions of cysteines Ci and Cj, respectively.

Finally, we provided to the SVR an input vector of 545 components based on all features described above.

For the SVR implementation, we used the libsvm package (http://www.csie.ntu.edu.tw/~cjlin/libsvm) with a RBF kernel.

2.5 Measuring scoring efficiency

Here Tp, Tn, Fp and Fn are, respectively, true positives, true negatives, false positives and false negatives with respect to the disulfide-bonding state class. The disulfide-bonding state predictions are evaluated using the following indices:

- Q_2 or accuracy that evaluates the number of correctly predicted cysteines divided by the total number of cysteines:

$$Q_2 = \frac{\mathrm{Tp} + \mathrm{Tn}}{\mathrm{Tp} + \mathrm{Tn} + \mathrm{Fp} + \mathrm{Fn}} \tag{13}$$

- Precision (Pr) of the disulfide-bonding state class that is the number of correctly predicted cysteines divided by the total number of predicted cysteines in the positive class:

$$\mathrm{Pr} = \frac{\mathrm{Tp}}{\mathrm{Tp} + \mathrm{Fp}} \tag{14}$$

- Recall (Rc) of the disulfide-bonding state class is the number of correctly predicted cysteines divided by the total number of observed bonded cysteines:

$$\mathrm{Rc} = \frac{\mathrm{Tp}}{\mathrm{Tp} + \mathrm{Fn}} \tag{15}$$

- F_1, defined as the harmonic mean of Pr and Rc:

$$F_1 = \frac{2 \times \mathrm{Pr} \times \mathrm{Rc}}{\mathrm{Pr} + \mathrm{Rc}} \tag{16}$$

- Matthews Correlation Coefficient (CC) defined as follows:

$$\mathrm{CC} = \frac{(\mathrm{Tp} \times \mathrm{Tn} - \mathrm{Fp} \times \mathrm{Fn})}{\sqrt{(\mathrm{Tp} + \mathrm{Fp}) \times (\mathrm{Tp} + \mathrm{Fn}) \times (\mathrm{Tn} + \mathrm{Fp}) \times (\mathrm{Tn} + \mathrm{Fn})}} \tag{17}$$

- Q_{prot} is the number of correctly predicted proteins N_{cp} divided by the total number of proteins N_p:

$$Q_{\mathrm{prot}} = \frac{N_{cp}}{N_p} \tag{18}$$

When we score the connectivity pattern prediction, we also compute the following indices:

- P_b is the number of correctly predicted bonds N_c divided by the total number of predicted bridges N_p:

$$P_b = \frac{N_c}{N_p} \tag{19}$$

- R_b is the number of correctly predicted bonds N_c divided by the number of observed bonds N_b:

$$R_b = \frac{N_c}{N_b} \tag{20}$$

- Q_p is the number of correctly predicted disulfide patterns N_{pat} divided over the total number of proteins N_p:

$$Q_p = \frac{N_{\mathrm{pat}}}{N_p} \tag{21}$$

Table 2. GRHCRF and linear CRF performance as a function of different inputs

Model	Input	Q_{prot} (%)	CC (%)	Q_2 (%)	Pr (%)	Rc (%)	F_1 (%)
CRF	PSSM	79	70	88	84	73	78
GRHCRF	PSSM	83	80	91	91	83	87
GRHCRF	PSSM + OSL	87	85	93	92	87	90
GRHCRF	PSSM + PSL	86	83	93	91	86	87

OSL and PSL, respectively, represent the observed and predicted subcellular information. Relative errors associated to each index are all below 1%. For index definition, see Section 2.

3 RESULTS AND DISCUSSION

3.1 Model selection procedure

Both CRF and SVR depend on hyperparameters that need to be adjusted (σ^2 in the regularization term of CRF, γ and C in SVR). Furthermore, both in the bonding state prediction and in the disulfide connectivity prediction, part of the input is based on a window of flanking residues centered on the cysteines. The size of these windows needs to be set as well.

All these parameters have been chosen by performing a cross-validation procedure. The dataset was first divided into a number of balanced sets as described in Section 2.1. Using this data split, we selected the parameters by averaging the best values obtained for each training set (in many cases, they are the same). With this procedure, we ended up with the following training-based parameters: $\sigma^2 = 0.05$ and $w = 15$ for the CRF and $\gamma = 0.0625$, $C = 11.31$ and $w = 13$ for the SVR.

3.2 Predicting the disulfide-bonding state: the role of subcellular localization

We applied the GRHCRF method for the prediction of the disulfide-bonding states of cysteines with a 20-fold cross-validation procedure. The best results (with input window $w = 15$) are reported in Table 2. For sake of comparison (and using the same input encoding), we also trained CRF models based on the same FSA of Figure 1, using artificial sequences of labels derived from a FSA parsing. However, in order to highlight the effect of the grammatical constraints, CRF models in Table 2 were not provided with hard-coded forbidden transitions (τ_{st}) set to $-\infty$. The results show that the grammatical rules are not easy to be acquired from the examples, since the GRHCRF models outperform the CRF ones in Table 2. As mentioned above, in the simple case of the FSM of Figure 1, GRHCRFs collapse to linear CRFs when both label and grammatical constraints are taken into account.

From Table 2, it is also evident that subcellular localization plays a significant role in predicting the cysteine-bonding state. In particular, Q_{prot} (accuracy per protein) and CC (Matthews correlation coefficient) scores indicate that information derived from the observed subcellular localization (OSL) increases up by five percentage points the method performance. The improvement is slightly lower when the predicted subcellular localization (PSL) is included in the input vector, indicating that BaCelLo is endowed with a high prediction score. It is worth mentioning that in PDBCYS the trivial cases are not present (chains that contain a single cysteine).

Table 3. Scoring the prediction of disulfide connectivity when the cysteine-bonding state is known

Index	Number of bonds				
	2	3	4	5	All
$R_b = P_b$	75 (1)	60 (2)	57 (2)	46 (3)	60 (2)
Q_p	75 (1)	48 (2)	44 (3)	19 (5)	54 (2)

Relative errors for each index in terms of percentage are reported within parenthesis. SVR description and index definition are provided in Section 2.

Table 4. Prediction of protein disulfide bonds with DISLOCATE as a function of the number of disulfide bonds

#B	PSSM			PSSM + PSL		
	R_b	P_b	Q_p	R_b	P_b	Q_p
1	80 (1)	38 (3)	72 (1)	83 (1)	46 (3)	76 (1)
2	64 (2)	50 (2)	60 (2)	67 (2)	52 (2)	61 (2)
3	46 (3)	41 (3)	34 (3)	47 (3)	41 (3)	35 (3)
4	52 (2)	35 (3)	33 (3)	52 (2)	37 (3)	35 (3)
5	39 (3)	39 (3)	15 (6)	39 (3)	39 (3)	15 (6)
All	52 (2)	41 (3)	34 (3)	52 (2)	42 (3)	36 (3)

#B, number of bonds. PSSM, input with PSSM. PSSM + PSL, input that add the predicted subcellular localization. Relative errors for each index in terms of percentage are reported within parenthesis. For index definition see Section 2.

3.3 Predicting connectivity patterns

To train and test the predictor of connectivity patterns based on SVR, we adopted the same 20-fold partition of the dataset after removing chains that contained less than two disulfide bonds per structure (Table 3). Here, we assume a perfect knowledge of the disulfide-bonding state of cysteines (Table 3). The SVR aims to predict the connectivity pattern that gets increasingly complex as the number of disulfide bonds increases (Fariselli and Casadio, 2002). The procedure does not restrict the prediction to the connectivity patterns that are present in the dataset and allows prediction of never-seen-before patterns (the restricted procedure can be implemented as well, improving the method performance, Singh 2008; Tsai *et al.*, 2007; Vincent *et al.*, 2008).

3.4 DISLOCATE: the integrated predictor of cysteine bonds in proteins considering subcellular localization

The prediction of subcellular localization, of cysteine-bonding states and of their topology, is then integrated into DISLOCATE, that takes a protein sequence as input. To evaluate DISLOCATE, both wrong disulfide state predictions and wrong connectivity assignments are taken into account when scoring the performance. Subcellular localization is considered as an input added feature. Values reported in Table 4 are obtained with a cross-validation procedure and as a function of the number of known disulfide bonds in the protein chain. It is patent that information on subcellular localization, albeit predicted, increases DISLOCATE performance for proteins with up to four disulfide bridges.

Table 5. Prediction of protein disulfide bonds with DISLOCATE as a function of the number of cysteines

#Cys	PSSM			PSSM + PSL		
	R_b	P_b	Q_p	R_b	P_b	Q_p
2	48 (2)	64 (2)	92 (1)	75 (1)	44 (3)	96 (1)
3	19 (5)	33 (3)	89 (1)	41 (3)	62 (2)	93 (1)
4	58 (2)	68 (2)	85 (1)	58 (2)	68 (2)	85 (1)
5	43 (3)	67 (2)	86 (1)	43 (3)	67 (2)	86 (1)
6	47 (3)	45 (3)	76 (1)	48 (2)	46 (3)	76 (1)
7	49 (2)	58 (2)	86 (1)	45 (3)	55 (2)	85 (1)
8	48 (2)	46 (3)	73 (1)	47 (3)	44 (3)	74 (1)
9	38 (3)	46 (3)	89 (1)	56 (2)	56 (2)	90 (1)
10	46 (3)	46 (3)	60 (2)	47 (3)	46 (3)	60 (2)
All	47 (3)	51 (2)	83 (1)	49 (2)	48 (2)	86 (1)

#Cys, number of cysteines; PSSM, input with PSSM; PSSM + PSL, input that add the predicted subcellular localization. Relative errors for each index in terms of percentage are reported within parenthesis. For index definition, see Section 2.

In Table 5, we report its accuracy as a function of the number of cysteines in the proteins (up to 10 cysteines), independently of the observed or predicted bonding state. Data show that when the predicted subcellular localization is added (PSL), the performance of the method increases. It is worth mentioning that when accuracy is scored as a function of the number of cysteines (Table 5), the vast majority of the protein sequences contain only free cysteines, resulting in higher Q_p values if compared to the case that consider only proteins with disulfide bonds (Table 4).

3.5 Comparison with other methods

Our method exploits protein subcellular localization in Eukaryotes according to the computations resulting by a cross-validated version of BaCelLo (Pierleoni *et al.*, 2006). However, for sake of comparison, we benchmarked DISLOCATE with other methods that were tested on SPX- (Cheng *et al.*, 2006). This dataset comprises 51% of proteins from Prokaryotes (out of 2547 protein structures). Unfortunately, it is not possible to compare the performance of the two separate steps (disulfide-bonding state and connectivity pattern predictions) with the available methods on a single dataset, since in the literature this information is not reported. In particular, for the prediction of the connectivity patterns most of the approaches adopt the highly redundant SP39 dataset (Fariselli and Casadio, 2001), where the sequence homology (if not properly handled) can mask the real performance (Supplementary Table S1). With this aim, here we retrain DISLOCATE (with the same parameter selected for PDBCYS) using a 10-fold cross-validation procedure. These 10 subsets were selected in order to prevent sequences with local similarity >25% from being extracted from two different sets (the cross-validation folds are available on the DISLOCATE Web page). BaCelLo predictions are obtained using the cross-validation procedure described above (Section 2.3.4).

DISLOCATE results are shown in Table 6 without and with the added subcellular localization feature (first and second groups of columns, respectively). Our data are compared with the scoring indices of state-of-the-art predictors as derived from the literature (Cheng *et al.*, 2006; Vincent *et al.*, 2008). The higher DISLOCATE overall accuracy is due to the fact that the vast

Table 6. Comparison with other approaches on SPX- dataset

#B	DISLOCATE PSSM			DISLOCATE PSSM + PSL			APTK + 1-NN[a]			DISULFIND +1-NN[a]			DIpro[a]		
	R_b	P_b	Q_p	R_b	P_b	Q_p	R_b	P_b	Q_p	R_b	P_b	Q_p	R_b	P_b	Q_p
1	90	69	88	92	74	90	30	30	27	30	30	30	–	–	–
2	70	54	69	71	60	70	51	54	47	51	51	49	–	–	–
3	60	54	53	63	54	55	63	65	58	66	67	61	–	–	–
4	46	36	30	48	37	32	50	51	40	48	49	37	–	–	–
All	60	51	50	62	52	53	43	44	37	43	44	39	32	48	–

#B, number of bonds. PSSM, input with PSSM; PSSM + PSL, input that add the subcellular localization. All the values are percentage (%).

[a] Values taken from Vincent *et al.* (2008). For the indices, see Section 2.

majority of the proteins in SPX- contains 1 or 2 bridges and in this case, DISLOCATE outperforms other methods. DISLOCATE, unlike other predictors, does not restrict the spectrum of possible connectivity patterns considered by the nearest neighbor approach (1-NN). This is reflected in the accuracy indices for 3 and 4 bonds where the DISLOCATE performance is lower than those including 1-NN approach that filters out connectivity patterns not present in SPX- (Vincent *et al.*, 2008). The probability of randomly predicting a correct connectivity pattern decreases exponentially as the number of disulfide bonds increase (Fariselli and Casadio, 2001). Indeed, the number of possible patterns is 1, 3, 15 and 105 when the number of disulfide bonds is 1, 2, 3 and 4, respectively. Accordingly, a high number of disulfide bonds determines a decrease in DISLOCATE performances. DISLOCATE and the other methods predict all the possible patterns at 1 and 2 number of bonds. However, at 3 and 4 bonds the 1-NN restricts the selection to 13 (out of 15) and 18 (out of 105) patterns, respectively. Therefore, DISLOCATE outperforms its random predictor 7.6 ($Q_p \times 15$) and 30 folds ($Q_p \times 105$), when 3 and 4 bonds in the disulfide patterns are considered. In turn, for the same number of bonds in the pattern, the best 1-NN-filtered approach scores 7.6 ($Q_p \times 13$) and 7.3 ($Q_p \times 18$) higher than random.

Furthermore, adding the subcellular localization feature improves DISLOCATE performance for each number of bonds in the disulfide pattern. In this benchmarking, the role of subcellular localization in improving the prediction is blurred by the fact that only 49% of the sequences are from Eukaryotes.

4 DISLOCATE SERVER

GRHCRF code is available under the GPL license and can be downloaded at the page http://www.biocomp.unibo.it/savojard/biocrf.html. The complete implementation of DISLOCATE is also freely accessible as Web server at the Web page http://www.biocomp.unibo.it/savojard/Dislocate.html. The server interface is extremely user friendly, and requires only to paste or upload a protein sequence from Eukaryotes. The only choice left to the user is the section of the organism kingdom type: animals (default), fungi and plants. This is necessary to activate the BaCelLo subcellular localization prediction (Pierleoni *et al.*, 2006). The Web server takes only one protein sequence at a time and it is not intended for intensive wide-genome scanning.

5 CONCLUSIONS

In this article, we present a new two-step predictor of disulfide bonds based on a newly developed machine learning model (Fariselli *et al.*, 2009) and taking protein sequence as input. We show that the inclusion of protein subcellular localization improves its performance, indicating that this piece of biological information is relevant for the classification of the bonding state of cysteine residues. We also show that the method matches up to with the available state-of-the-art predictors.

ACKNOWLEDGEMENTS

We thank Dr Orsola Brizio from the ‘L. Heilmann’ Interfaculty Centre for Applied and Theoretical Linguistics of the University of Bologna for proofreading.

Funding: MIUR-FIRB grant number RBLA039M7M for the LIBI project delivered to R.C.

Conflict of Interest: none declared.

REFERENCES

Altschul,S.F. *et al.* (1997), Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.

Byrd,R.H. *et al.* (1995) A limited memory algorithm for bound constrained optimization. *SIAM J. Sci. Stat. Comput.*, **16**, 1190–1208.

Casadio,R. *et al.* (2008) The prediction of protein subcellular localization from sequence: a shortcut to functional genome annotation. *Brief. Funct. Genomic Proteomic*, **7**, 63–73.

Chen,Y.C. *et al.* (2004) Prediction of the bonding states of cysteines using the support vector machines based on multiple feature vectors and cysteine state sequences. *Proteins*, **55**, 1036–1042.

Cheng,J. *et al.* (2006) Large-scale prediction of disulphide bridges using kernel methods, two-dimensional recursive neural networks, and weighted graph matching. *Proteins*, **62**, 617–629.

Durbin,R *et al.* (1998) *Biological Sequence Analysis: Probabilistic Models of Proteins and Nucleic Acids*. Cambridge University Press, Cambridge, UK.

Fariselli,P. and Casadio,R. (2001) Prediction of disulfide connectivity in proteins. *Bioinformatics*, **17**, 957–964.

Fariselli,P. *et al.* (2009) Grammatical-Restrained Hidden Conditional Random Fields for Bioinformatics applications. *Algorithms Mol. Biol.*, **4**, 1–10.

Ferrè,F. and Clote,P. (2005) Disulfide connectivity prediction using secondary structure information and diresidue frequencies. *Bioinformatics*, **21**, 2336–2346.

Heras,B. *et al.* (2007) The name’s bond. . . disulfide bond. *Curr. Opin. Struct. Biol.*, **17**, 691–698.

Imai,K. and Nakai,K. (2010) Prediction of subcellular locations of proteins: where to proceed? *Proteomics*, **10**, 3970–3983.

Inaba,K. (2010) Structural basis of protein disulfide bond generation in the cell. *Genes Cells,* **15**, 935–943.

Kadokura,H. *et al.* (2004) Protein disulfide bond formation in prokaryotes. *Annu. Rev. Biochem.*, **72**, 111–135.

Lafferty,J.D. *et al.* (2001) Conditional random fields: probabilistic models for segmenting and labeling sequence data. In *Proceedings of the Eighteenth International Conference on Machine Learning (ICML’01)*, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, pp. 282–289.

Martelli,P.L. *et al.* (2002) Prediction of the disulfide bonding state of cysteines in proteins with hidden neural networks. *Protein Eng. Des. Sel.*, **15**, 951–953.

Mucchielli-Giorgi,M.H. *et al.* (2002) Predicting the disulfide bonding state of cysteines using protein descriptors. *Proteins*, **46**, 243–249.

Pierleoni,A. *et al.* (2006) BaCelLo: a balanced subcellular localization predictor. *Bioinformatics,* **22**, e408–e416.

Savojardo,C. *et al.* (2011) Prediction of the bonding state of cysteine residues in proteins with machine-learning methods. In Rizzo,R. and Lisboa,P.J.G. (eds) *CIBB 2010, LNBI*, Springer, Berlin, Heidelberg, pp. 98–111.

Sevier,C.S. *et al.* (2007) Modulation of cellular disulfide-bond formation and the ER redox environment by feedback regulation of Ero1. *Cell*, **129**, 333–344.

Singh,R. (2008) A review of algorithmic techniques for disulfide-bond determination. *Brief. Funct. Genomic Proteomic,* **7**, 157–172.

Song,J. *et al.* (2007) Predicting disulfide connectivity from protein sequence using multiple sequence feature vectors and secondary structure. *Bioinformatics,* **23**, 3147–3154.

Sutton,C. and McCallum,A. (2007) An introduction to conditional random fields for relational learning. In Getoor,L. and Taskar,B. (eds) *Introduction to Statistical Relational Learning*, MIT Press, Cambridge, MA, pp. 103–127.

Taskar,B. *et al*. (2005) Learning structured prediction models: a large margin approach. In *Proceedings of the Twenty Second International Conference on Machine Learning (ICML05)*, ACM New York, NY, USA, p. 102.

Tsai,C.H. *et al.* (2007) Bioinformatics approaches for disulfide connectivity prediction. *Curr. Protein Pept. Sci.*, **8**, 243–260.

Vincent,M. *et al.* (2008) A simplified approach to disulfide connectivity prediction from protein sequences. *BMC Bioinformatics*, **9**, 20.

Vullo,A. and Frasconi,P. (2004) Disulfide connectivity prediction using recursive neural networks and evolutionary information. *Bioinformatics*, **20**, 653–659.