PROTEINS: Structure, Function, and Genetics 46:243–249 (2002)
DOI 10.1002/prot.10047

Predicting the Disulfide Bonding State of Cysteines Using Protein Descriptors

M.H. Mucchielli-Giorgi,[1] S. Hazout,[2] and P. Tufféry[2]*
[1]*Equipe Statistique des Séquences Biologiques, UPRESA CNRS, Université d'Evry, Département de Mathématiques, Evry, France*
[2]*Equipe de Bioinformatique Génomique et Moléculaire, INSERM U436, Université Paris, Paris CEDEX, France*

ABSTRACT **Knowledge of the disulfide bonding state of the cysteines of proteins is of major interest in designing numerous molecular biology experiments, or in predicting their three-dimensional structure. Previous methods using the information gained from aligned sets of sequences have reached up to 82% of success in predicting the oxidation state of cysteines. In the present study, we assess the relative efficiency of different descriptors in predicting the cysteine disulfide bonding states. Our results suggest that the information on the residues flanking the cysteines is less informative about the disulfide bonding state than about the amino acid content of the whole protein. Using a combination of logistic functions learned with subsets of proteins homogeneous in terms of their amino acid content, we propose a simple prediction approach, starting from a single sequence, that reaches success rates close to 84%. This score can be improved by avoiding predictions regarding cysteines for which the decision is not well marked. For example, we obtain a score close to 87% correct prediction when we exclude predicting 10% of the cysteines. Proteins 2002;46:243–249.**
© 2002 Wiley-Liss, Inc.

Key words: cysteine redox state prediction; structure prediction; logistic function; protein clustering

INTRODUCTION

Cysteines constitute a special class of amino acids; their oxidation state plays an important role in protein structure and function. The formation of disulfide bridges, involving the oxidized form of cysteines or cystines, is known as a major component of the stabilization of the native conformation of proteins, as it results in forming cross-links between distant regions of the protein sequence.[1–6] In its reduced form, the thiol group of cysteines is particularly reactive. In some enzyme families, such as the cysteine proteinases, the thiol group of cysteines is part of the active site. In the molecular biology field, it can be used in experiments designed to study protein structure and function. Various methodologies, such as spin-labeling techniques and fluorescence labeling,[7–12] use these characteristics.

All these different properties make the prediction of the oxidized state of cysteines a particularly attractive tool concerning both the prediction of the three-dimensional structure of proteins or the engineering of protein stability and functionality. Different experimental, but time-consuming, techniques attempt such determinations (see, e.g., refs. 13 and 14). Different computational studies have focused on this goal, generally based on the analysis of the properties of the regions flanking each cysteine. Early studies by Muskal et al.[15] and Fiser et al.[16] have analyzed these environments and shown that different environments are observed for cysteines and cystines. More recently, Fariselli et al.[17] trained a neural network with the information of the amino acid profiles of the regions flanking the cysteines. These investigators reached a score of correct prediction of 72% using single sequences, increasing up to 81% when using the information of multiple alignments. Finally, Fiser et al.[18] proposed a prediction approach based on a score measuring the conservation of the residues flanking the cysteine in multiple alignments. Interestingly, these workers remarked that, for most cases, the same oxidation state tends to be found for the cysteines of a same protein, which led them to propose a prediction scheme at the protein level: all the cysteines of the same protein are predicted in the same oxidation state, i.e., the state mostly predicted over the cysteines of the protein. Accepting some systematic error for the few cases corresponding to proteins for which different oxidation states occur for the cysteines, Fiser and colleagues also reached scores up to 82% correct prediction. Obviously, such strategy deviates from predicting the oxidation state of cysteines starting from their local sequence environments. One might wonder how informative the amino acids flanking the cysteines are compared with some categorization of the whole amino acid content of the protein.

In the present study, we investigate the relative contributions of the regions flanking the cysteines and of the overall amino acid composition of the protein in predicting the disulfide bonding state of the cysteines. We use different prediction approaches, all based on the training of a logistic function. Despite its simplicity, the use of a single

*Correspondence to: P. Tufféry, Equipe de Bioinformatique Génomique et Moléculaire, INSERM U436; Universite Paris, case 7113, 2 Place Jussieu, 75251 Paris CEDEX 05, France. E-mail: tuffery@urbb.jussieu.fr.

Received 5 April 2001; Accepted 27 September 2001

© 2002 WILEY-LISS, INC.

TABLE I. Assessment of the Prediction Using Different Information*

Data type	Q_2 (%)	E_{SS} (%)	E_S (%)
bE	67.3 ± 2.3 (68.5)	38.8 (43.1)	27.9 (22.9)
fE	70.1 ± 2.1 (71.9)	46.2 (43.0)	17.1 (17.1)
Pr	77.7 ± 4.3 (83.1)	29.3 (26.3)	16.6 (10.1)
fP	79.3 ± 3.3 (82.6)	26.3 (25.6)	16.3 (11.4)

bE, results based on the binary encoding of the nature of residues flanking the cysteines; *fE*, results based on the frequency encoding of the nature of residues flanking the cysteines; *Pr*, results based on descriptors typing each protein; *fP*, a combination of *fE* and *Pr*.
*The uncertainty values correspond to 1 SD. Values within parentheses correspond to the results obtained with a database including proteins with one cysteine, for comparability with other studies.

logistic function is well suited for deciphering some of the determinants that undergo a prediction process, as it is possible to analyze the weights associated with each input, and thus to extract information about the parameters that do condition the prediction. We apply a simple predictor to a series of data extracted from single sequences; assessment of the accuracy of the prediction is achieved through a cross-validation process. First, we perform a learning–prediction cycle on a whole set of nonredundant proteins using, as input of the logistic function, information about either the residues flanking the cysteines or the amino acid content of the whole protein. This leads us to iterate this process on groups of proteins extracted by clustering the initial set of proteins on the basis of their relative content in amino acids. Finally, a global prediction strategy is assessed.

MATERIALS AND METHODS

Database of Cysteine-Containing Fragments

For training and validation, we have used a collection of nonredundant protein chains from the culled Protein Data Bank (PDB)[19] (not models, percentage of sequence identity less than 25%, and a resolution of >2.5 Å). Two sets of proteins were defined.

A first set (set 1) was used to assess the prediction scores obtained using different descriptors and to define groups of proteins. Starting from a collection of 869 proteins, only for 683 does at least one cysteine occur in the sequence, and only for 559 do at least two cysteines occur in the sequence. As we are interested in predicting the disulfide bonding state of cysteines, the studies were made from this set of 559 proteins, except when noted in Table I. This corresponds to a total of 2,882 cysteines. The disulfide bonding state of the cysteines was determined by analyzing the interatomic distance between the SG sulfur atoms of the cysteines of each protein, with disulfide bonds corresponding to a distance of < 2.15 Å. Among the 559 proteins for which disulfide bond could exist, none is observed for 348. For 156 chains, all the cysteines are involved in disulfide bonds, and for the remaining 55 proteins, different oxidation states occur depending on each particular cysteine. Overall, this resulted in a collection of 1,270 cysteines involved in disulfide bonds (denoted SS), and 1,612 not involved (denoted S).

A second set of proteins (set 2), unrelated to the previous set, was defined among the newcomers of the culled PDB. This set consists of 375 proteins, for which at least two cysteines occur. For 75 proteins, all the cysteines are involved in disulfide bonds, for 274 proteins, no disulfide bond occur, and for 26 proteins, different states occur. This corresponds to 1,838 cysteines, with 464 involved in disulfide bonds. This set was used to assess the global prediction score.

Descriptors of the Cysteine Disulfide Bonding States

Cysteine environment descriptors

The environment of the cysteine was defined by a window of size $2n + 1$, centered on the cysteine of interest, where n is the number of amino acids flanking the cysteine each side. Only the $2n$ positions of such a window are of interest, the central position always corresponding to a cysteine. Two different ways of encoding the environment of a cysteine were considered. For each residue in the window, the first method consists of defining a vector of dimension 20 (for the 20 different amino acid types), with all elements set to 0, but one set to 1, according to the type of the particular amino acid encoded. We denote it as *bE* for "binary" description. A second approach, denoted *fE* for "frequency" description, was to use for each residue of the window a single number, $\log(f^l_k/f^l)$, where f^l_k corresponds, for the protein k, to the occurrence frequency of the type of amino acid l observed at a given position of the window, and f^l the occurrence frequency of the same type of amino acid in the whole database. This type of descriptor can reflect the under- or overrepresentation of some amino acid types at the different locations in the window. This set of descriptors was supplemented with a measure of the position of the cysteine within the sequence, obtained as

$$(p/N_k)*[(N_k - p)/N_k]$$

where p is the position of the amino acid in the sequence, and N_k the number of amino acids of the protein k. This descriptor measures the distance of the cysteine to the extremities of the sequence.

Protein descriptors

For each protein, we define a vector of dimension 23. The 20 first elements are $\log(f^l_k/f^l)$, where f^l_k corresponds to the occurrence frequency of amino acid type l in the protein k, and f^l the occurrence frequency of the amino acid type l in the whole database. The 21th corresponds to a normalized size of the protein, as $\log(N_k/N_{\text{mean}})$ where N_k is the size of the protein k, and N_{mean} the average size of the proteins of the database. The two last elements are an explicit measure of the number of cysteines occurring in the protein (number of occurrences and frequency of occurrences in the protein k). This set of descriptors is denoted *Pr* for protein descriptors.

The Logistic Function: A Simple Model for Prediction

The method for predicting the disulfide bonding state of a cysteine is based on the use of a single logistic function,

equivalent to a perceptron, i.e., a neural network without a hidden layer. In our case, the basic input of the logistic function is a vector $X[X_i]$ of descriptors describing either the cysteine environment (*bE* or *fE*) or the protein (*Pr*), or a combination of both. The probability of observing a cysteine in a disulfide bond is evaluated as

$$P(X) = 1/[1 + \exp(-w_0 - \sum_i w_i X_i)]$$

where w_i is the weight of the descriptor at position i of the input vector. The vector of weights $W[w_i]$ is estimated by a maximum likelihood method on a learning set, using the general linear model (GLM) package of the *R* software.[20]

Decision Rule and Assessment of the Prediction Accuracy

The rate of error R associated with the prediction was defined as

$$R = P(Pred = SS/Obs = S) \cdot P(Obs = S) + P(Pred = S/Obs = SS) \cdot P(Obs = SS)$$

where *Pred* and *Obs* denote, respectively, the prediction and the reality, *SS* (resp. *S*), a cysteine (resp. not) involved in a disulfide bond. It corresponds to the probability of either predicting a *SS* state when a *S* state is observed, or a *S* state when a *SS* state is observed. The rate of correct prediction (or accuracy of the prediction) Q_2 is equal to $1 - R$. We also define E_S (resp. E_{SS}) as the fraction of cysteines in the state *S* (resp. *SS*) predicted in the *SS* state (resp. *S*), i.e.

$$E_S = P(Pred = SS/Obs = S) \quad \text{and} \quad E_{SS} = P(Pred = S/Obs = SS)$$

Given a logistic function and its associated vector of weights, the decision rule consists simply of testing whether the output $P(X)$ of the logistic function is more or less than a given threshold *Ps*. If $P(X) > Ps$, the prediction is *SS*; otherwise, it is *S*. After estimating the weights of the logistic model, we determined the threshold by minimizing the rate of error *R*. This was achieved simply by slicing the [0 . . . 1] interval by steps of 0.01, computing the value of *R* for each step, and keeping the value of the slice for which the minimum of *R* was obtained. This decision threshold was applied during the validation to assess Q_2.

Cross-Validation of the Results for Set 1

To permit cross-validation of the prediction procedure, the collection of the cysteines of set 1 was split into five randomly selected subsets. For each of the five subsets, the four other subsets were used as the learning set, while this particular subset was used for validation. Each cysteine observed in the database is present either in the learning or the validation set.

Clustering of Proteins

Classification of the proteins was carried out, to describe each protein, using a vector $V[V_k^l]$ of dimension 20, where $V_k^l = f_k^l/\sqrt{f^l}$, f_k^l the frequency of the amino acid l in the protein k, and f^l the frequency of the amino acid in the whole database. Hence, a Euclidian distance between two proteins k and k' is expressed as

$$d(k, k') = \sum_{l=1}^{20} (V_k^l - V_{k'}^l)^2 = \sum_{l=1}^{20} (f_k^l - f_{k'}^l)^2/f^l$$

This is equivalent to using a χ^2 metrics. Hierarchical classification was performed using Ward's minimum variance method.[21] After visual inspection of the tree to determine a reasonable number of groups, the clusters were determined by a *K*-means approach,[22] using as input the means of the clusters extracted by hierarchical clustering. The stability of the groups was assessed by verifying that the means of the clusters move slightly. Finally, a Sammon map[23] was used to display the repartition of the proteins in the clusters. We used the *R* software[20] implementation of these methods.

Global Decision Rule Using a Jury Approach

Given the definition of clusters of proteins, it is possible to estimate the probability that one new sequence belongs to one cluster by assuming that the distances intra-clusters are distributed according to Gaussian laws. Given the means and standard deviations of the clusters, one can estimate the probability that a given sequence *s* belongs to a group *g* as

$$ps_g = f(ds_g) \Big/ \sum_i f(ds_i)$$

with

$$f(ds_i) = \exp[-ds_i^2/(2\sigma_i^2)]/(\sqrt{2\pi}\sigma_i)$$

where $f(ds_i)$ is the probability density that sequence *s* belongs to group *i*, ds_i is the distance between sequence *s*, and the center of the group *i*, σ_i is the standard deviation of the distances of group *i*.

Given a set of predictors (logistic functions) associated with each group of proteins, it is possible to combine them into a global jury score defined as

$$J(s) = \sum_g V(X)_g \times ps_g$$

where ps_g is the probability that sequence *s* belongs to group *g* and $V(X)_g$ the "vote" obtained using the predictor of group *g*, expressed as 0 when the prediction is that the cysteine is not involved in disulfide bond, 1 else. Hence, the nearest groups condition mostly the jury score. The decision rule simply consists of testing whether the output $J(s)$ is more or less than a given threshold J_{ss}. If $J(s) > J_{ss}$, the prediction is *SS*, else *S*. Here, we have taken a value of 0.5 for J_{ss}.

RESULTS AND DISCUSSION

Assessing the Relative Contribution of the Predictors

Table I presents the cross-validation results obtained using different predictors on the whole set 1. The *bE* (resp

TABLE II. Properties of the Groups of Proteins According to Their Amino Acid Composition

Group	N	All_{SS}	All_S	Het_{SS}	Mean_{sze}	f_{CYS} (%)	f_{SS} (%)
g_s	68	42	20	6	92	10.1	74.1
g_1	125	16	103	6	276	1.7	17.0
g_2	137	29	95	13	230	1.8	37.1
g_3	128	13	108	7	329	1.4	13.6
g_4	101	56	22	23	276	1.9	78.2

N, Number of proteins of the group; All_{SS}, number of proteins for which all cysteines form disulfide bonds; All_S, number of proteins for which all cysteines form no disulfide bonds; Het_{SS}, number of proteins for which only some cysteines form disulfide bonds; Mean_{sze}, mean size of proteins of the group; f_{CYS}, frequency of the cysteines in the group; f_{SS}, frequency of cysteines forming disulfide bonds.

fE) line corresponds to the results obtained with the binary (resp. frequency) encoding of the residues flanking the cysteines (for a window of seven residues each side, i.e., the optimum observed by Fariselli et al.[17]). The data within parentheses correspond to results obtained from a database including the proteins for which only one cysteine exists, for comparability with related studies. One observes rates of prediction on the order of 70%, i.e., slightly less than the rate of 72% obtained by Fariselli and coworkers when training a neural network, and carrying out a prediction from single sequences. The use of the *fE* encoding leads to better results. A larger portion of the errors comes from a poorer attribution to the *SS* state, on the order of two times worse than that obtained for the *S* state.

The *Pr* line corresponds to the results obtained when using only the set characterizing the protein composition. Better scores are obtained, comparable to the score of 82% obtained by Fiser et al.,[18] using a conservation score within aligned proteins. Compared with the *bE* and *fE* results, most of the improvement comes from a better discrimination of the disulfide bonding cysteines, as shown by the E_{SS} column.

Looking at the weights associated with the logistic function, it is possible to extract some of the factors that determine the prediction. For *bE*, we obtain very few significant contributions, with poor reproducibility among the different subsets of the cross-validation: the negative contributions of Met at location −3 and Ile at location +2, with the cysteine located in position 0. For *fE*, which measures the under- or overrepresentation of the amino acids observed at certain locations, whatever their type, the only significant contributions detected for some cross-validation sets are associated with locations −3 and +3, and they are negative. For *Pr*, we obtain a much larger number of significant contributions, with good reproducibility through the cross-validation. In this study, we only report the contributions of the amino acids that are extremely significant ($P < 10^{-6}$). Asp, Phe, Asn, Pro, Ser, Thr, and Tyr contribute positively to the *SS* state, while His, Met, Val, and a large normalized size of protein are negative parameters. Interestingly, this is in partial agreement with the observations by Abkevich and Shakhnovich,[24] who reported, using lattice model simulations, a positive correlation of disulfide content with Gln, Ser, Thr, and Asn, and a negative correlation with Glu, Asp, Lys, and Arg.

Finally, we have considered combining both types of descriptors. We report only results obtained for *fP*, a combination of *fE* and *Pr*. For *fP*, we obtain only a small gain (1.6%) compared with *Pr*, and no gain at all when considering proteins for which at least two cysteines occur. This suggests that most of the information lies within *Pr* and that comparatively little complementary information lies within *fE*, i.e., the local sequence environment. We have checked this explicitly by proposing as input of a new logistic function the outputs of *Pr* and *fE*, to analyze the relative weights resulting for the two types of input. We obtain weights on a magnitude on the order of 10 times larger for *Pr* compared with that of *fE*, which confirms a major contribution of the amino acid composition of the protein compared with that of the local sequence environment.

Overall, these results suggest that the *bE* or *fE* information is, in fact, much less informative than that of *Pr*. Other combinations of descriptors for various window sizes of either *bE* or *fE* (not reported) always lead to the same conclusion.

Prediction Within Clusters of Unrelated Proteins of Similar Amino Acid Content

Given the observation that protein characterization predictors lead to the best prediction rates, we have investigated whether it is possible to categorize subsets of our collection of proteins of set 1 in terms of the set of protein descriptors. Clustering of the proteins of set 1 (see Materials and Methods) first resulted in isolating a small homogeneous group of 68 proteins, distant from all others, denoted g_s. As shown in Table II, the major characteristics of this group, compared with others, are the smaller size of the proteins and the richer cysteine content (on the order of six times larger than for other proteins). Also, the fraction of cysteines involved in disulfide bonds is large. Such a group has already been noted.[25] Using hierarchical classification, clustering of the remaining 491 proteins leads to the identification of four stable groups that were satisfactorily homogeneous. Figure 1(a) shows a Sammon map illustrating how these four groups are differentiated. As shown Table II, these groups show different combinations of their protein mean size, their number of All_{SS} or

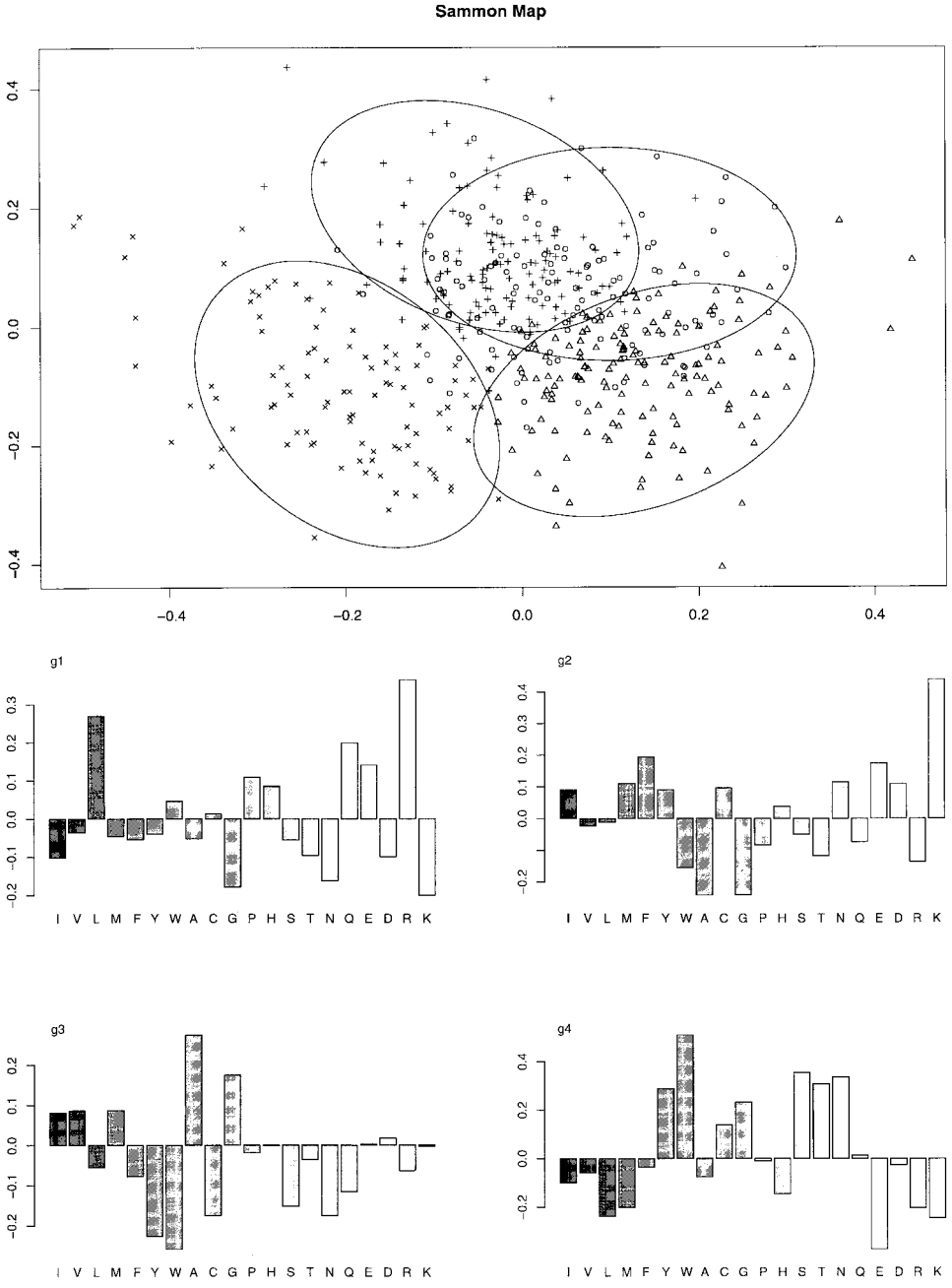


Fig. 1. **a:** Sammon map of the four groups of proteins, after exclusion of the small cysteine-rich proteins. Symbols correspond to the different groups. Ellipses roughly indicate the location of the highest densities for each group. **b:** Relative propensities of the amino acids in the four groups of proteins.

All_S proteins, and their cysteine content that is involved in disulfide bonds. Figure 1(b) presents the amino acid profiles of these proteins. As expected, the four groups show very distinct profiles. Group 1 is associated with a lack of Gly, Asn, and Lys and with an excess of Leu and Arg. Group 2 is characterized by an excess of Lys and with a lack of Ala and Gly; group 3 by an excess of Ala and Gly and a lack of Tyr and Trp; and group 4 by an excess of Tyr, Trp, Ser, Thr, and Asn, and a lack of Glu.

The proteins corresponding to each group were isolated from the database, and the learning/validation process was iterated on each group separately. Table III presents the results obtained from these subsets. In light of the results obtained from the whole database, the results are reported for only three combinations of descriptors. Again, the protein predictors attain much higher scores than do the cysteine environment predictors. For *Pr*, the rate of correct predictions in all cases is higher than 86% and can be as high as 95%, with a mean of 89.8%, which is much larger than the results obtained for the whole database. For the local cysteine environment predictors, the rates of correct prediction, within the range of 75–85%, are also

TABLE III. Assessment of the Prediction Using Different Information for Homogeneous Groups of Proteins*

	Data type								
	Pr			*fE*			*fP*		
Group	Q_2 (%)	E_{SS} (%)	E_S (%)	Q_2 (%)	E_{SS} (%)	E_S (%)	Q_2 (%)	E_{SS} (%)	E_S (%)
g_s	95.3 ± 1.6	3.9	8.4	76.1 ± 4.7	5.1	76.3	93.7 ± 1.8	4.2	12.4
g_1	86.9 ± 2.5	70.2	1.4	82.3 ± 4.6	97.1	1.5	85.2 ± 3.4	56.1	5.7
g_2	86.4 ± 2.1	27.3	5.6	78.2 ± 4.1	40.9	10.6	83.9 ± 2.1	28.7	8.7
g_3	93.0 ± 2.3	23.1	4.5	84.9 ± 2.7	96.4	2.4	90.6 ± 1.9	33.1	5.7
g_4	86.7 ± 3.9	4.8	41.7	77.8 ± 3.6	1.0	98.2	83.7 ± 3.1	6.2	50.8

Pr, results based on descriptors typing each protein; *fE*, results based on the frequency encoding of the nature of residues flanking the cysteines; *fP*, a combination of *fE* and *Pr*.
*The uncertainty values correspond to 1 SD.

better than those obtained from the whole database, but the combination of both sets of descriptors (*fP*) leads to worse scores than for *Pr*. The errors appear to be distributed differently, according to the different subsets of proteins. For g_1, most of the errors come from a poorer prediction of the cysteines involved in disulfide bonds, while for g_4, most errors correspond to cysteines that are not involved in disulfide bonds. These tendencies appear to be related to the frequencies of cysteines forming disulfide bonds in the various groups, and show the limits of the discrimination of the logistic function, choosing to favor errors in the smaller set. For *Pr*, this tendency is much smaller, however, than when using *fE*. Interestingly, this behavior is not observed for g_s, where the errors appear to be well balanced between the *SS* and the *S* cysteines.

We report only the contributions of the amino acids that are extremely significant ($P < 10^{-6}$). For g_s, the occurrence of disulfide bonds appears to be favored by Asn and Thr, while for g_3, it is Asp and Thr. For g_1, it appears not to be favored by Ile, Met, and Ala. For g_4, it is unfavored by Ile, Val, Met, and Asp and by large frequencies of occurrence of cysteines in the protein. For g_2, only this last descriptor is extremely significant. These amino acids and the sign of their contributions are rather consistent with those obtained from the whole set of proteins. Also, residues that contribute most to the prediction do not correspond to the residues with the large propensity values shown in Figure 1(b).

Predicting the Oxidation State of Cysteines Without Prior Knowledge of Cluster Assignment

Finally, such an increase in the scores is not useful per se, as one does not know the group to which new proteins belongs. Therefore, we have to consider a strategy in which group assignment and prediction of the oxidation state of the cysteines are embedded. We face the problem that proteins might be incorrectly assigned to their "true" group. Furthermore, newcomers might correspond to proteins that are rather distant from the proteins used in set 1. Taking these possible problems into account, we have considered a full prediction scheme based on combining the predictors obtained for the five groups of proteins previously defined (see Materials and Methods). A first assessment of the strategy was performed using a cross-validation of set 1. We obtained a correct prediction score of 84.2% ± 2.3% (1 SD), with associated E_{SS} and E_S values of 23.7% ± 4% and 9.3% ± 0.9%. Compared with the average score close to 89% obtained on the different well-characterized clusters, a loss close to 5% appears to come from the cluster assignment. It is possible to improve this strategy by considering the validity of the decision for proteins for which the jury score is close to the limit. We have investigated a strategy in which we accept carrying out a prediction only if the decision is marked (i.e., no prediction is performed for values that are too close to the decision threshold). By providing no conclusion for values of scores within ±0.1 of the threshold (i.e., 0.4–0.6), we obtain a prediction for 90.5% of the cysteines. Among these cysteines, the correct prediction score is 87.4% ± 1.8%, associated with E_{SS} and E_S values of 18.3% ± 2.7% and 5.8% ± 0.7%. Corresponding values for a window of ±0.2 around the decision threshold are a prediction for 71.5% of the cysteines, with a correct prediction score of 91%, and E_{SS} and E_S values of 10.4% and 3.2%. Hence, well-marked decisions appear to be associated with more reliable predictions.

Finally, we measured the efficiency of the procedure on the newcomers set (set 2). The predictors are learned from the groups defined from protein set 1, and we assess the prediction scores using the unrelated protein set 2. We obtained a correct prediction score of 84.2%, with associated E_{SS} and E_S values of 41.8% and 6.5%. The scores obtained when providing no conclusion for values of votes within ±0.1 are of 85.2% with associated E_{SS} and E_S values of 33.1% and 6.5%, without providing a conclusion for 10.3% of the cysteines. For a window of ±0.2, the corresponding scores are 87.7%, 31%, and 2.5%, providing no conclusion for 20.6% of the cysteines. Such results are consistent with those obtained during the cross-validation.

CONCLUSIONS AND PERSPECTIVES

The present results suggest that the local sequence environment of the cysteines contains much less information than that of the amino acid composition of the protein. Indeed, predicting the disulfide bonding state of cysteines from a set of protein descriptors appears to be more efficient than focusing on the specificity associated with the regions flanking the cysteines. The accuracy of prediction increases when the proteins are partitioned into homogeneous subsets. This finding is fully consistent with

the previous observation, by Fiser and Simon,[18] that the oxidation states of the cysteines of a same polypeptide chain tend to be similar. Overall, we obtain better scores using single sequences than obtained from previous studies using alignment information. Alignment information, however, is not so independent of the global amino acid content of the protein.

Concerning the design of molecular biology experiments, based on chemical reactions using the thiol group of cysteines, it is somewhat interesting to note that, for four of five groups we obtain better scores of success in predicting "no disulfide bond" proteins. As for predicting the three-dimensional structure of proteins, the issue that arises becomes one of attempting to predict the pairs of cysteines that form disulfide bonds for "all disulfide bond" proteins.

REFERENCES

1. Zhou NE, Kay CM, Hodges RS. Disulfide bond contribution to protein stability: positional effects of substitution in the hydrophobic core of the two-stranded alpha-helical coiled-coil. Biochemistry 1993;32:178–187.
2. Clarke J, Fersht AR. Engineered disulfide bonds as probes of the folding pathway of barnase: increased the stability of proteins against the rate of denaturation. Biochemistry 1993;32:4322–4329.
3. Betz SF. Disulfide bonds and the stability of globular proteins. Protein Sci 1993;2:1551–1558.
4. Creighton T. Disulfide-coupled protein folding pathways. Philos Trans R Soc Lond B 1995;348:5–10.
5. Roesler KR, Rao AG. A single disulfide bond restores thermodynamic and proteolytic stability to an extensively mutated protein. Protein Sci 2000;9:1642–1650.
6. Clarke J, Hounslow AM, Bond CJ, Fersht AR, Daggett V. The effect of disulfide bonds on the denatured state of barnase. Protein Sci 2000;9:2394–2404.
7. Creighton T. Proteins: structures and molecular properties. 2nd ed. New York: Freeman; 1993.
8. Guixe V. Chemical modification of SH groups of *E. coli* phosphofructokinase-2 induces subunit dissociation: monomers are inactive but preserve ligand binding properties. Arch Biochem Biophys 2000;376:313–319.
9. Kim J, McNamee MG. Topological disposition of Cys 222 in the alpha-subunit of nicotinic acetylcholine receptor analyzed by fluorescence-quenching and electron paramagnetic resonance measurements. Biochemistry 1998;37:4680–4686.
10. Imamoto Y, Kataoka M, Tokunaga F, Palczewski K. Light-induced conformational changes of rhodopsin probed by fluorescent alexa594 immobilized on the cytoplasmic surface. Biochemistry 2000;39:15225–15233.
11. Sano A, Nakamura H. Fluorescence quenching properties and chemiluminescence responses of alpha-ketothiols derivatized with o-phthalaldehyde and primary amino compounds. Luminescence 2001;16:25–28.
12. Schindel C, Zitzer A, Schulte B, Gerhards A, Stanley P, Hughes C, Koronakis V, Bhakdi S, Palmer M. Interaction of *Escherichia coli* hemolysin with biological membranes. A study using cysteine scanning mutagenesis. Eur J Biochem 2001;268:800–808.
13. Morris HR, Pucci P. A new method for rapid assignment of S-S bridges in proteins. Biochem Biophys Res Commun 1985;126:1122–1128.
14. Xue J, Kalafatis M, Silveira JR, Kung C, Mann KG. Determination of the disulfide bridges in factor Va heavy chain. Biochemistry 1994;33:13109–13116.
15. Muskal S, Holbrook S, Kim S. Prediction of the disulfide-bonding state of cysteine in proteins. Protein Eng 1990;3:667–672.
16. Fiser A, Cserzo M, Tudos E, Simon I. Different sequence environments of cysteins and half cysteines in proteins: application to predict disulfide forming residues. FEBS Lett 1992;302:117–120.
17. Fariselli P, Riccobelli P, Casadio R. Role of evolutionary information in predicting the disulfide-bonding state of cysteine in proteins. Proteins 1999;36:340–346.
18. Fiser A, Simon I. Predicting the oxydation state of cysteines by multiple sequence alignment. Bioinformatics 2000;16:251–256.
19. Hobohm U, Scharf M, Schneider R. Selection of representative protein data sets. Protein Sci 1993;1:409–417.
20. Ihaka GR. A language for data analysis and graphics. J Comp Graph Stat 1996;5:299–314.
21. Ward JH. Hierarchical grouping to optimize an objective function. J Am Stat Assoc 1963;58:236–244.
22. Blashfield RK, Aldenderfer MS, Morey LC. Cluster analysis software. In: Krisnaiah PR, Kanal LN, editors. Handbook of statistics. Vol 2, Amsterdam: North-Holland; 1982. p 245–266.
23. Sammon JW. A non linear mapping for data structure analysis. IEEE Trans Comp C 1969;18:401–409.
24. Abkevich VI, Shakhnovich EI. What can disulfide bonds tell us about protein energetics, function and folding: simulations and bioinformatics analysis. J Mol Biol 2000;4:975–985.
25. Harrison PM, Sternberg MJ. The disulfide beta-cross: from cystine geometry and clustering to classification of small disulfide-rich protein folds. J Mol Biol 1996;264:603–623.