Volume 302, number 2, 117–120 FEBS 11024 May 1992

Hypothesis

Different sequence environments of cysteines and half cystines in proteins

Application to predict disulfide forming residues

András Fiser, Miklós Cserzö, Éva Tüdös and István Simon

Institute of Enzymology, Biological Research Center, Hungarian Academy of Sciences, Budapest H-1518, PO Box 7, Hungary

Received 17 March 1992; revised version received 30 March 1992

Protein sequences are often derived by translating genetic information, rather than by classical protein sequencing. At the DNA level cysteines and half cystines are indistinguishable. Here we show that the sequential environments of 'free' cysteine and half cystine are different. A possible origin of this difference is discussed and a simple method to predict cysteines and half cystines from the amino acid sequence is also presented.

Prediction; Free cysteine; Half cystine; Sequential environment

1. INTRODUCTION

Nowadays protein sequences are usually determined via sequencing the corresponding cDNA. However, this efficient technique is not able to recover the complete information content of native proteins, as the post-translational modification and the disulfide bonding pattern are missed. The distinction between the two forms of cysteine is important not only from the viewpoint of the number of reactive sulfhydryl groups, but also because the knowledge of the complete covalent structure seems to be dispensable in the prediction of the 3D structure [1]. Until now just one attempt has been made to predict the covalent status of Cys residues. In that method a neural network approach was applied to the analysis of the 3D structure database (Protein Data Bank). The calculation was rather complicated and did not provide a simple algorithm for prediction [2].

It has been shown that every amino acid has a characteristic sequential environment (within a ± 10-residue distance) in proteins [3,4]. This finding was used to predict protein coding open reading frames in DNA [5], domain boundaries of multidomain proteins [6] and isomorphic residue replacements for protein design [7]. What is more, Frömmel and Preissner pointed out that the sequential environments of Pro residues are different depending on their *cis* or *trans* conformation and this difference can be applied to predict the conformation of Pro in proteins [8].

Correspondence address: I. Simon, Institute of Enzymology, Biological Research Center, Hungarian Academy of Sciences, Budapest H-1518, PO Box 7, Hungary. Fax: (36) (1) 166-5465.

In this paper we show that the sequential environments of 'free' cysteines and half cystines are different. A possible reason for this difference is discussed and a simple method to predict cysteine and half cystine from the amino acid sueqence is also presented here.

2. DATABASES

Protein sequence data were taken from the SWISS-PROT database (release 17) [9]. The database contains sequence data for 20,024 proteins with a total number of 100,482 half cystines and cysteines. The covalent status is documented for about 10% of them. These documented Cys residues were selected from the database with their leading and tailing 10 sequential neighbours. The statistical set was reduced according to a gross homology consideration. Fragments showing 100% homology in their 7 residues with the central heptamer of the any other fragment were removed. We ended up with 411 half cysteine and 1379 free cysteine containing 21-residue-long fragments.

3. RESULTS AND DISCUSSION

For each position from −10 to +10 relative to the central Cys the occurrence of the 20 amino acids were counted. The ratio of corresponding elements is presented in Table I. Values greater than 1.0 indicate that the corresponding residue in the given position appears more often in the vicinity of half cystines than in that of free cysteines.

Some values differ from 1.0 because they reflect pref-

Table I

Ratio of the normalized abundances of various residues in a given position in the vicinity of half cystines and 'free' cysteines

	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	Cys	1	2	3	4	5	6	7	8	9	10
A	1.50	0.73	1.44	0.72	0.59	0.99	0.99	1.18	1.17	1.12		1.74	1.26	0.95	1.15	0.65	0.88	0.70	0.63	0.74	1.54
C	0.77	1.64	0.62	1.65	1.25	1.44	1.44	0.50	0.44	0.69		0.81	0.41	0.61	1.33	1.54	1.30	1.79	0.78	1.39	0.87
D	0.94	0.81	1.16	1.60	1.94	0.62	0.82	1.42	1.08	1.04		0.90	0.73	1.17	1.60	1.35	1.06	1.09	1.16	1.31	1.28
E	0.66	1.27	0.51	1.88	1.27	1.06	1.03	1.25	0.81	0.97		0.94	0.81	0.71	0.44	0.74	0.53	0.57	0.59	0.49	0.46
F	0.90	0.17	0.41	0.43	0.88	0.49	0.52	0.79	0.65	0.80		0.97	0.49	0.99	0.41	1.21	0.59	1.11	0.75	0.96	1.08
G	1.55	2.07	1.35	1.08	1.18	1.94	1.90	1.25	1.24	1.29		1.95	1.75	1.38	1.53	1.04	1.46	1.54	1.25	0.92	1.53
H	0.78	0.72	0.68	0.43	0.43	0.67	0.22	1.00	1.08	1.31		0.22	0.23	0.78	2.25	0.76	0.69	0.35	0.54	0.51	0.44
I	0.70	0.98	0.58	0.96	1.16	1.06	0.47	0.65	1.30	1.02		0.42	1.31	0.94	0.76	1.30	1.31	0.90	1.36	1.13	0.78
K	0.67	0.77	0.87	1.08	0.81	1.26	0.67	0.90	1.11	0.72		1.00	0.94	1.15	0.79	1.16	0.64	1.12	1.02	1.01	0.87
L	0.60	0.57	0.71	0.54	0.44	0.56	0.61	0.21	0.78	0.50		0.48	0.59	0.35	0.74	0.45	0.46	0.94	0.81	0.80	0.61
M	0.66	0.80	0.34	0.53	0.76	0.62	0.33	0.46	1.99	0.44		0.54	0.75	0.48	0.53	0.37	0.59	0.54	0.36	0.57	0.44
N	1.60	1.44	1.55	0.87	1.06	0.88	1.91	2.08	2.03	1.25		1.80	1.58	1.65	1.47	1.07	2.35	1.89	1.52	1.17	1.03
P	0.93	0.72	1.18	1.03	0.87	0.59	0.99	0.89	0.33	0.81		0.70	0.61	0.70	0.74	0.84	1.14	1.38	1.05	0.81	1.57
Q	0.79	1.18	0.82	1.13	1.17	0.92	1.03	1.07	0.40	0.72		1.05	1.11	0.91	0.84	1.24	0.65	0.73	0.88	1.34	0.53
R	0.95	1.12	0.99	1.09	0.94	0.50	0.53	0.72	1.10	0.73		0.77	0.88	0.61	0.47	0.49	0.77	0.90	0.68	0.99	0.41
S	1.51	0.94	1.55	0.98	1.06	1.09	1.00	1.53	0.83	0.98		1.69	1.40	1.52	1.16	1.29	1.23	0.92	1.38	0.96	1.12
T	0.80	1.03	1.21	1.12	1.03	2.04	1.23	1.43	1.10	1.86		0.73	1.40	1.53	0.87	0.98	1.36	0.56	1.32	1.24	1.27
V	1.05	1.26	0.71	0.69	1.29	1.05	1.12	0.81	0.73	1.08		0.62	1.25	0.95	0.57	1.02	0.63	0.56	0.61	1.14	1.27
W	0.81	0.17	2.57	1.63	1.25	1.15	1.66	0.14	0.51	1.67		0.78	0.36	1.17	1.28	1.77	1.24	1.42	1.00	2.15	1.34
Y	1.63	1.28	2.05	1.55	1.11	1.03	1.37	1.25	2.73	1.81		2.21	0.89	1.81	1.60	2.22	1.63	1.36	2.50	1.59	1.03

erence for half cystine or cysteine, some others may be due to statistical uncertainty.

Table II is given to show the significant differences in deviation units. Elements of Table II were calculated in the following way: the difference of normalized abundances of a given residue in the vicinity of half cystines and free cyteines was divided by the sum of the standard deviations of the two normalized abundances. Since the abundance data are binomially distributed, the latter were calculated as:

$$D^2(x) = n \times p \times (1-p)$$

where p is the probability of the given element, and n is the number of all occurrences. If the average occurrence of the various residues were the same in the neigh-

Table II

Disulfide forming preferences in standard deviation units, i.e. difference between the normalized abundances of various residues in the vicinity of half cystines and 'free' cysteines divided by the sum of their standard deviation. The values above (+2) and under (−2) are underlined with a broken or solid lines, respectively

	−10	−9	−8	−7	−6	−5	−4	−3	−2	−1	Cys	1	2	3	4	5	6	7	8	9	10
A	1.50	−0.97	1.24	−1.09	−1.71	−0.04	−0.06	0.66	0.58	0.39		2.03	0.92	−0.20	0.51	−1.51	−0.43	−1.22	−1.48	−1.04	1.49
C	−0.71	1.30	−1.25	1.58	0.66	1.06	1.01	−1.39	−1.98	−1.15		−0.61	−2.22	−0.90	0.82	1.24	0.71	1.86	−0.54	0.81	−0.35
D	−0.15	−0.56	0.45	1.36	2.23	−1.36	−0.61	1.10	0.24	0.11		−0.31	−0.97	0.50	1.36	0.89	0.16	0.23	0.46	0.87	0.75
E	−1.17	0.73	−1.85	1.98	0.77	0.16	0.09	0.74	−0.62	−0.08		−0.17	−0.63	−0.96	−2.44	−0.98	−1.82	−1.49	−1.41	−1.99	−2.22
F	−0.29	−3.80	−2.11	−1.83	−0.32	−1.80	−1.51	−0.57	−1.00	−0.60		−0.07	−1.61	−0.03	−1.98	0.44	−1.39	0.26	−0.77	−0.12	0.20
G	1.83	2.94	1.25	0.31	0.65	2.68	2.58	0.90	0.86	0.94		2.78	2.41	1.49	1.96	0.14	1.68	1.80	0.86	−0.30	1.74
H	−0.39	−0.60	−0.77	−1.28	−1.50	−0.68	−2.40	0.01	0.15	0.60		−2.89	−2.75	−0.57	2.11	−0.56	−0.79	−1.95	−1.11	−1.36	−1.60
I	−0.99	−0.05	−1.50	−0.13	0.44	0.16	−1.96	−1.20	0.80	0.07		−2.00	0.68	−0.15	−0.71	0.74	0.59	−0.26	0.81	0.36	−0.76
K	−1.16	−0.70	−0.39	0.24	−0.64	0.71	−1.05	−0.29	0.30	−0.94		−0.01	−0.19	0.43	−0.62	0.45	−1.08	0.30	0.06	0.03	−0.42
L	−2.10	−2.13	−1.41	−2.39	−3.15	−2.31	−1.92	−5.16	−1.06	−2.76		−2.92	−1.91	−3.70	−1.19	−2.63	−2.68	−0.25	−0.79	−0.82	−1.65
M	−0.69	−0.45	−1.80	−1.17	−0.50	−0.82	−1.49	−1.33	1.31	−1.23		−1.12	−0.45	−1.21	−0.89	−1.29	−0.82	−0.98	−1.36	−0.78	−1.25
N	1.25	0.97	1.11	−0.39	0.16	−0.32	1.83	2.10	1.96	0.60		1.56	1.17	1.44	1.17	0.19	2.33	1.64	1.19	0.41	0.08
P	−0.22	−1.05	0.50	0.09	−0.44	−1.53	−0.03	−0.35	−3.02	−0.63		−1.26	−1.58	−1.20	−1.00	−0.55	0.42	1.27	0.17	−0.70	1.42
Q	−0.66	0.48	−0.46	0.32	0.42	−0.21	0.07	0.16	−2.02	−0.83		0.16	0.25	−0.26	−0.40	0.61	−1.03	−0.76	−0.33	0.77	−1.55
R	−0.15	0.34	−0.02	0.22	−0.15	−1.81	−1.45	−0.96	0.23	−0.88		−0.68	−0.36	−1.37	−1.83	−1.99	−0.74	−0.32	−1.12	−0.04	−2.24
S	1.56	−0.25	1.64	−0.06	0.21	0.35	−0.01	1.71	−0.73	−0.08		2.13	1.44	1.59	0.54	0.99	0.91	−0.36	1.24	−0.16	0.44
T	−0.66	0.09	0.68	0.35	0.10	2.35	0.69	1.38	0.29	2.39		−0.99	1.12	1.46	−0.42	−0.08	1.08	−1.79	0.90	0.71	0.77
V	0.16	0.77	−1.15	−1.24	0.86	0.14	0.42	−0.74	−1.12	0.24		−1.44	0.84	−0.16	−1.72	0.05	−1.39	−1.90	−1.70	0.42	0.77
W	−0.32	−2.14	1.47	0.87	0.31	0.22	0.93	−2.57	−0.97	0.76		0.39	−1.32	0.26	0.42	0.99	0.34	0.50	0.00	1.19	0.46
Y	1.08	0.61	1.72	0.92	0.24	0.08	0.78	0.48	3.06	1.74		1.74	−0.24	1.30	1.20	1.95	1.19	0.72	2.46	1.07	0.08

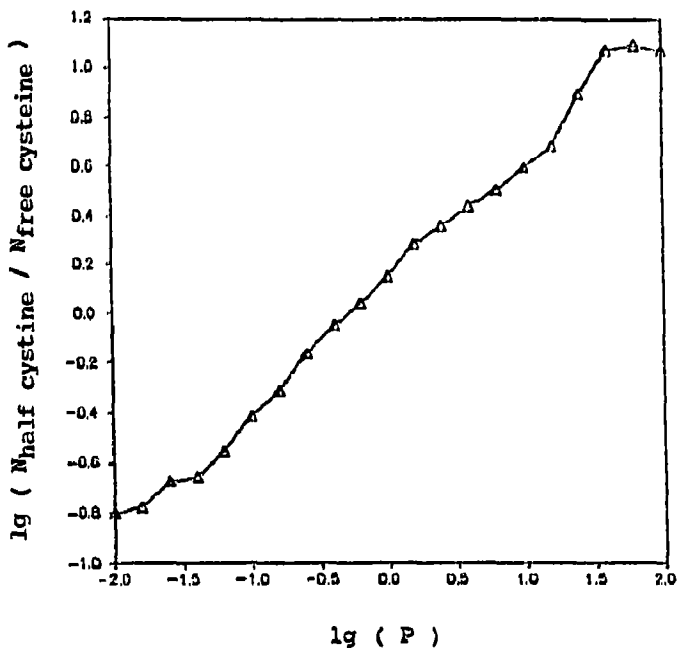

Fig. 1. The logarithm of the ratio of cysteine and 'free' cysteine abundances versus the logarithm of the disulfide forming potential (P).

bourhood of half cystines and cysteines then (in accordance with Gaussian-distribution) less then 18 out of the 400 values would be expected outside the (−2, +2) interval and only one outside the (−3,+3) interval. In fact, 39 and 6 fall outside the (−2,+2) and (−3,+3) intervals, respectively (Table II).

It can be seen in Table II that the polar, neutral, H-bond forming residues appear more often in the neighbourhood of half cysteines, while the sequential environments of 'free' cysteines are rather hydrophobic. The charged residues occur also preferentially in the vicinity of free cysteines (see underlined values in Table II).

Gly is the most abundant residue in the vicinity of half cystines. Since it has no side chain it endows the backbone with large flexibility which may be necessary for the exact coordination of the covalently bonded half cystines. In addition, the very good H-bond forming peptide bond is not shadowed by any side chain.

Among the favouring residues, their H-bond former character seems to be the definitive property in the half cystine preference (e.g. Asn, Tyr, Ser, Thr).

In contrast, among the half cystine negating (i.e. free cysteine favouring) residues the dominant quality is the bulkiness of the apolar side chain (e.g. Leu, Phe, Ile, Val, Ala).

For a given residue the balance of these properties may determine the half cystine preferring quality. The importance of H-bond forming property can be seen in the example of Tyr and Phe which otherwise have similar bulkiness, but the first one distinctly favours the half cystine while the second one prefers free cysteine. Likewise the importance of bulkiness can be seen for the examples of Asn and Gln which have similar H-bond forming character but the bulkier Gln prefers half cystine much less than the smaller Asn.

Our findings are in fair agreement with those of [2] which also shows that hydrophobic residues accumulate in the vicinity of free cysteines while polar residues and Gly are more abundant in the neighbourhood of half cystines. However, there are some differences referring to the ranking within these two main groups, especially for some rare amino acids. The most striking differences can be found for Met, His, Cys and Tyr. Their relative abundances are as low as: 2.27%, 2.36%, 2.14% and 3.74%, respectively [10]. A part of the differences between our results and that of [2] may reflect the very small data set of [2], i.e. less then 9,000 residues around Cys were studied compared to our more than 37,000 residues. Furthermore in [2] only a special subset of all proteins, namely the crystallizable ones were analysed, but any enlargement of the database necessarily requires the revision of all prediction methods based on statistical analysis [10].

The observed differences between the sequential environments of half cystines and 'free' cysteines (Table I) may be applied to predict disulfide forming cysteines from amino acid sequence.

A simple prediction method may work in the following way. Let us introduce the disulfide forming potential for a Cys containing segment as the product of the corresponding values of Table I according to the sequence of 21 membered segment. If the 21-residue-long segment considered overlaps the end of the amino acid sequence, or in the case undefined amino acids, one must obviously use a value of 1.0. The greater the product (the disulfide bond forming potential), the greater is the probability that the Cys in question is a half cystine. Likewise, disulfide forming potentials close to zero indicate greater probability of 'free' Cys.

One of the best ways to test a prediction power was suggested by Rooman and Wodak [11]. According to this 'jack-knife' procedure one protein was removed from the database and from the remainder of the database a new table similar to Table I was created, and the prediction power was tested on the removed protein. Then this procedure was repeated for every single protein and the average of the observed power was considered as the prediction power of the method. In this case the efficiency of the prediction was about 71% (choosing 1.0 for the limit). We should like to note that using linearly decreasing weights moving away from the Cys or taking shorter segments into account does not improve the prediction power.

For extremely high or low disulfide forming potentials the method is reliable. Fig. 1. shows the prediction power as a function of the applied threshold limit. For example choosing 2.0 for the prediction limit, then the yield for half cystines is more than 87% and nearly 50%

of all the examined half cystines have higher predicted values; on the other hand, if the prediction limit is 0.5 then the predictions yield about 75% for 'free' cysteines, and nearly 70% of all of the 'free' cysteines have smaller predicted values.

We can conclude that the sequential environments of half cystine and cysteine are different enough to make the prediction of the disulfide forming probabilities of various Cys residues in protein possible.

Acknowledgements: This work was supported by a research grant from the Hungarian Academy of Sciences (OTKA 1361). The authors are grateful to Dr. A. Aszódi for his advice and help in the database management and to Dr. F. Vonderviszt in the preparing this manuscript.

REFERENCES

[1] Simon, I., Glasser, L. and Scheraga, H.A. (1991) Proc. Natl. Acad. Sci. USA 88, 3661–3665.
[2] Muskal, S.M., Holbrook, S.R. and Kim, S.-H. (1990) Protein Engineering 3, 667–672.
[3] Vonderviszt, F., Mátrai, Gy. and Simon, I. (1986) Int. J. Peptide Protein Res. 27, 483–492.
[4] Cserző, M. and Simon, I. (1989) Int. J. Peptide Protein Res. 34, 184–195.
[5] Mitra, C.K., Cserző, M. and Simon, I. (1991) J. Theor. Biol. (submitted).
[6] Vonderviszt, F. and Simon, I. (1986) Biochem. Biophys. Res. Commun. 139, 11–17.
[7] Tüdös, É., Cserző, M. and Simon, I. (1990) Int. J. Peptide Protein Res. 36, 236–239.
[8] Frömmel, C. and Preissner, R. (1990) FEBS Lett. 277, 159–163.
[9] Bairoch, A. Dept. de Biochimie Medicale, CMU, 1 Rü Michel Servet, 1211 Geneve 4., Switzerland (in: Kahn, P. and Cameron, G. (1990) Methods in Enzymology, 183, 23–31).
[10] Simon, I. and Cserző, M. (1990) Trends Biochem. Sci. 15, 135–136.
[11] Rooman, M.J. and Wodak, S.J. (1988) Nature 335, 45–49.