Patterns in the sequence context of protein disulfide bonds

by

Aron Charles Eklund

B.S. Physics
University of California, San Diego, 1996

Submitted to the Department of Biology in partial
fulfillment of the requirements for the degree of

Master of Science in Biology
at the
Massachusetts Institute of Technology

February 2002

Signature of Author _____
Department of Biology
January 2002

Certified by _____
Chris A. Kaiser
Professor of Biology
Thesis Supervisor

Accepted by _____
Alan D. Grossman
Co-Chair, Department Committee on Graduate Studies

Patterns in the sequence context of protein disulfide bonds

by

Aron Charles Eklund

Submitted to the Department of Biology in January 2002 in partial
fulfillment of the requirements for the degree of Master of Science in Biology

ABSTRACT

Disulfide bonds play an important role in the structural stability of the proteins that contain them. Yet, little is known about the specificity with which they are formed. To address this, a representative set of disulfide bonds from nonhomologous eukaryotic polypeptides was created. The amino acid sequences flanking these disulfide bonds were searched for conserved patterns that may reflect recognition sites by the disulfide bond forming enzyme protein disulfide isomerase (PDI). Several methods of classifying disulfide bonds were explored, and each class was analyzed for conserved sequence patterns. To maximize the chances of finding a conserved recognition site, a simulated annealing algorithm was implemented to divide a set of disulfide-bonded cysteines into two sets of cysteines with an average sequence environment that is as far from randomly-distributed as possible. No significant conserved patterns were found in the set of disulfide bonds or within any of the classification schemes introduced. Additionally, several methods for predicting disulfide bond connectivity were explored. The most successful methods predicted connectivity based on the sequential distance between cysteines.

TABLE OF CONTENTS

INTRODUCTION

Disulfide Bonds

A protein disulfide bond is the covalent bond formed between the sulfur atoms of two cysteine residues. The correct formation of disulfide bonds is necessary for the proper function of many proteins (Raina and Missiakas, 1997). Disulfide bonds stabilize proteins entropically by crosslinking the linear polypeptide (Johnson *et al*., 1978).

The formation of a structural disulfide typically occurs in extracytoplasmic environments; the cytoplasm contains reducing enzymes that disfavor disulfide bonds. In eukaryotic cells, disulfide bonds are formed in the endoplasmic reticulum (ER) (Braakman *et al*., 1991). In prokaryotes, disulfide bonds are formed in the periplasm (Debarbieux and Beckwith, 1999). In both eukaryotes and prokaryotes, a hydrophobic N-terminal signal peptide causes the nascent polypeptide to enter the ER or periplasm. For example, an estimated 10-20% of yeast proteins contain a signal sequence and are therefore delivered to the ER (Kaiser *et al*., 1997). After translocation into the ER, the protein attains its native state, aided by numerous chaperones (Wei and Hendershot, 1996). Typically, a protein does not form its final tertiary structure until all disulfide bonds have been formed (Wedemeyer *et al*., 2000).

A protein undergoing oxidative folding in the ER may contain several cysteines, and a disulfide could possibly form between any two of these cysteines. The number of potential disulfide connectivities (mappings of connected cysteines) N is given by the formula

$$N(c,d) = \frac{c!}{2^d\, d!(c-2d)!}$$

where c is the number of cysteines in the protein, and d is the number of disulfide bonds. For a medium-sized protein with three disulfides formed among seven cysteines, there are $N(7,3) = 105$ possible connectivities. For a more complicated protein with eight disulfide bonds, there are over two million distinct ways of connecting cysteines. Despite this large number of possible connectivities, disulfide bonds are found primarily in a single conformation within a healthy cell.

One model for the remarkable accuracy of oxidative folding is that disulfide bonds are formed randomly and then reshuffled until they reach a topological arrangement that allows a stable tertiary structure, as is thought to occur *in vitro* (Creighton, 1997). Another model is that

there is some degree of specificity in the formation of disulfide bonds, such that the majority of disulfide bonds are initially formed between the proper cysteines. Both models are dependent on the activity of protein disulfide isomerase (PDI), an enzyme that can both form and reshuffle disulfide bonds.

Oxidative Folding

Classic experiments performed by Anfinsen in the 1950s showed that reduced, denatured ribonuclease A (RNase A) is capable of uncatalyzed folding into its native structure in the presence of an oxidant. Native RNase A has eight cysteines, giving 105 possible combinations of four disulfide bonds. This unexpected observation of proper folding in the absence of other large molecules led to the "thermodynamic hypothesis", which states that all necessary information for proper folding is contained in the amino acid sequence, and that the native conformation is that in which the free energy is lowest (Anfinsen, 1973).

However, the observed rate of RNase A refolding was much slower *in vitro* than the expected physiological rate. A search for an *in vivo* catalyst led to the discovery of PDI (Goldberger *et al*., 1963), which increases the rate of *in vitro* refolding 15-fold (Lambert and Freedman, 1983).

Protein Disulfide Isomerase

The structure of PDI has not been solved, but proteolysis experiments and NMR data indicate an **a**-**b**-**b'**-**a'**-**c** domain structure (Figure 1). The **a** and **a'** domains are homologous to thioredoxin, a multifunctional 12-kDa protein that is a major cytoplasmic disulfide reductant. The **a** domain, when expressed alone, folds into the thioredoxin fold (Kemmink *et al*., 1996). The **b** and **b'** domains have no significant sequence similarity to thioredoxin, but surprisingly the **b** domain is also found in a thioredoxin-like structure (based on solution NMR) when expressed alone (Kemmink *et al*., 1997). The **a** and **a'** domains, but not the **b** and **b'** domains, contain the CXXC motif, which is known to be the redox-active site in thioredoxin and in several bacterial oxidoreductases. The **c** domain is a putative high-capacity Ca^{2+} binding site, and the extreme C-terminus contains an ER retention signal (Ferrari and Soling, 1999).

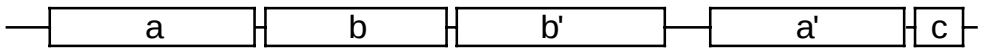


Figure 1. The putative domain structure of human PDI.

Multiple PDI homologs are known, including five in yeast (Pdi1p, Eug1p, Eps1p, Mpd1p, Mpd2p) (Frand *et al*., 2000) and seven in mammals (PDI, P5, ERp72, ERp57, PDIp, PDIR, PDI-Dβ). All have a signal sequence for import into the ER as well as C-terminal ER retention signals (Ferrari and Soling, 1999).

The roles of the multiple PDI homologs remain unclear, but it seems that their functions partially overlap. Several combinations of genomic deletions were recently performed in yeast (Norgaard *et al*., 2001). The genes encoding Eug1p, Eps1p, Mpd1p, and Mpd2p can be deleted without causing a noticeable phenotype; thus, among this homology group, Pdi1p appears to be sufficient for normal growth. The deletion of PDI1 is lethal but can be rescued by overexpression of any of the other four homologs, although growth rate is retarded by a factor of two. Furthermore, overexpressed Mpd1p alone (with the other homologs deleted) is sufficient for growth. Overexpressed Mpd2p alone is not sufficient, but overexpressed Mpd2p in combination with wildtype-level Mpd1p is sufficient.

Eug1p is interesting because it is the only PDI homolog without the full CXXC motif. Instead, Eug1p has two CXXS motifs, allowing it to act as an isomerase but not as an oxidase. Overexpressed Eug1p is sufficient for growth when Mpd1p and Mpd2p are present at wildtype levels. However, when the serines in the active sites are replaced with cysteines, overexpressed Eug1p alone is sufficient for normal growth. It appears that the multiple PDI homologs have overlapping functions but are not completely redundant in their activity. One possible explanation is that each PDI homolog has evolved to efficiently catalyze the oxidative folding of certain substrates, while the overexpression of certain PDIs can compensate for the lack of others.

PDI as a Protein Oxidase

The study of PDI has focused largely on its disulfide isomerase activity, in part because the isomerase activity is most easily assayed. However, recent work in yeast has suggested that PDI may have a different function *in vivo*. Two key techniques have been employed to accurately measure the *in vivo* oxidation state of PDI. First, by disrupting cells in a solution of

10% trichloroacetic acid, disulfide exchange is stopped quickly and completely ("acid trapping"), because all thiols become protonated (Weissman and Kim, 1991). Second, the reagent 4-acetamido-4'-maleimidylstilbene-2,2'-disulfonic acid (AMS) will conjugate with free thiols but not with disulfides, allowing the relative oxidation state of a protein to be assayed by a shift in SDS-PAGE mobility (Kobayashi *et al*., 1997). Using this "acid trap" followed by AMS treatment, it was demonstrated that the vast majority of PDI is fully oxidized *in vivo* (Frand and Kaiser, 1999). Since the fully oxidized form of PDI cannot act as an isomerase, it is likely that the primary cellular role of PDI is as an oxidase.

The oxidative role for PDI was probed with carboxypeptidase Y (CPY), a lumenal protein with five disulfide bonds that is oxidized and folded in the ER. PDI was shown to interact directly with CPY by the acid-trapping of a PDI-CPY heterodimer ("mixed disulfide"). Finally, the depletion of Pdi1p resulted in the accumulation of reduced CPY in the ER, suggesting that Pdi1p is necessary for protein oxidation in the ER (Frand and Kaiser, 1999).

PDI Substrate Specificity

Several studies have addressed the peptide binding specificity of PDI. In one experiment, the reduction of insulin by reduced glutathione (GSH) was used to assay the activity of purified bovine PDI. Several peptides of varying lengths and composition were tested, and all were found to competitively inhibit insulin reduction. The level of inhibition correlated well with the length of the peptide (tripeptide: K_i > 10mM, 29-mer: K_i = 200μM). Additionally, peptides containing cysteines were found to have 8-10 fold lower K_i than a peptide of similar length without a cysteine (Morjana and Gilbert, 1991). While these results are informative, the number of peptides tested was not sufficient to detect sequence-specific binding preferences.

Another study (Westphal *et al*., 1998) probed the binding specificity of human PDI when it acted as a disulfide reductase. A library of 30,000 partially-random peptides was synthesized on plastic beads, such that each bead contained a single species of peptide. The fluorescent group o-aminobenzoyl (Abz) was attached to the N-terminus. A peptide containing the quenching group 3-nitrotyrosine was attached by a disulfide bond, such that the fluorescence of Abz would increase when the disulfide bond was reduced (Figure 2).

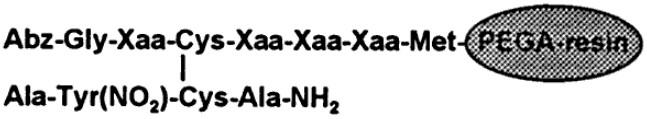


Figure 2. Fluorogenic substrate for probing the specificity of PDI as a reductase.

The library was first incubated in 1mM GSH, and the beads that became fluorescent were considered unstable and were removed from the library. The remainder of the library was incubated with PDI, and thirteen beads that became fluorescent within a short time (< 1 hour) were selected as candidates for a good substrate for PDI. Sequencing the peptides on these beads revealed a preference for the pattern (small/helix breaker)-Cys-Xaa-(hydrophobic/basic)-hydrophobic. This result suggests that the amino acid residues surrounding a substrate cysteine can influence catalytic reduction by PDI.

Statistical Analysis of Disulfide Bonded Sequences

PDI's substrate specificity might also be deduced from the structures of its natural substrates. A starting point for this analysis would be to look for features shared by disulfide-bonded cysteines. The features could be a specific sequence of amino acids, a certain secondary structure, or even a tertiary structure. A significant limitation of this approach is that it is not known which cysteines or disulfide bonds actually interact with PDI *in vivo*. However, there are hundreds of proteins of known structure that contain disulfide bonds and are therefore possible substrates. If a common feature could be found among the cysteines involved in disulfide bonds that was absent in cysteines with free thiols, this pattern would be a candidate for a PDI recognition signal. Alternatively, a common feature might be indicative of an environment where a disulfide bond is energetically favorable.

There have been a few attempts to correlate the native redox state of a cysteine with the amino acid sequence adjacent to the cysteine residue (its "sequence environment"). One group evaluated 10,000 cysteines of known redox state in the SWISS-PROT database (Fiser *et al*., 1992). The amino acid sequence from positions -10 to +10 relative to the cysteines were examined. About 1800 sequences were used in the actual analysis after identical sequences were removed. The distribution of residues at each position was calculated for oxidized cysteines and for reduced cysteines. A statistically significant bias was found: the presence of certain residues at certain positions frequently corresponded to either an oxidized cysteine or a reduced cysteine.

Given the sequence environment of a novel cysteine, the biases could be used to predict the cysteine oxidation state with 71% accuracy.

A second study employed an iterative learning algorithm to predict the oxidation state of a given cysteine (Fariselli *et al*., 1999). First, the amino acid sequence environments of 2500 cysteines from nonhomologous proteins of known structure were used to "train" the algorithm. For each cysteine, a window of 13 residues was used as the input; that is, the program was given the identity of the six residues towards the N-terminus and the six residues towards the C-terminus. When asked to predict the oxidation state of a cysteine not used in the training stage, the program was 72% accurate.

The prediction accuracy could be futher improved by using homology information taken from the HSSP (homology-derived secondary structure of proteins) database (Sander and Schneider, 1991). For every structure record in the Research Collaboratory for Structural Bioinformatics (RCSB) Protein Data Bank (PDB), a corresponding entry in the HSSP database contains multiple protein sequence alignments for all known homologous proteins. Instead of using the sequence information from a single protein, the average of the aligned sequences was used to train the program. The availability of homology information increased the prediction accuracy to 78%.

For each amino acid in a PDB entry, the HSSP database generates a conservation weight based on the homologous sequence alignments. The program was then trained with the conservation weight in addition to the averaged sequences for each of the 2500 cysteines. With this additional evolutionary information, the program could predict the oxidation state of a cysteine with 82% accuracy.

Although their prediction methods are far from perfect, these two computational analyses demonstrate that amino acid sequences contain information relevant to the oxidation state of a cysteine.

Statistical Analysis of Disulfide Bond Connectivity

A few researchers have performed comparisons of the properties of several disulfide-bonded proteins in hopes of finding common features that might characterize the process of oxidative folding.

In an early study (Thornton, 1981), 28 disulfide-containing proteins of known structure and 26 disulfide-containing proteins with known connectivity were analyzed. In this data set, the majority (83%) of proteins are extracellular. It was noted that very few of the proteins (6%) contain more than a single free thiol.

The 128 independent disulfide bonds present in these proteins was compared. The observed separation between linked residues was compared to the separation expected by chance, and a bias was found towards connections between nearby residues. Half of the disulfides were separated by fewer than 24 residues, while disulfides separated by more than 150 residues were rarely observed. This distribution may reflect the domain structure of proteins, and it could be attributed to folding kinetics, where nearby residues come into contact more frequently than distant ones.

A more recent study found similar trends among a larger data set (Petersen *et al*., 1999). 351 disulfide bonds from 131 nonhomologous single-chain proteins of known structure were analyzed. Again, a tendency towards disulfides separated by a small number of residues was observed, but no statistical analysis was performed. It was noticed that among proteins containing at least one disulfide bond, there was four times as many proteins with an even number of cysteines than an odd number. This probably reflects the reactivity of proteins containing free thiols; the oxidizing machinery of the ER may tend to form as many disulfide bonds as possible, and a protein may evolve to contain as few free thiols as possible to avoid improper intermolecular disulfide bonds.

Also, smaller proteins were observed to have a larger fraction of their residues involved in disulfide bonds. Since the tertiary structure of smaller proteins tends to be less stable, they benefit more from the structural stabilization of disulfide bonds.

Topology of Disulfide-Bonded Proteins

The entire topology of disulfide-bonded proteins has been studied (Benham and Jafri, 1993). A pattern of disulfide bonds on a protein can be efficiently represened by labeling each disulfide alphabetically and writing the order in which disulfide-bonded cysteines appear in a protein sequence. For example, a protein with one disulfide bond, which can only have one disulfide pattern, is represented as *aa*. A protein with two disulfide bonds has one of three

topological patterns, depending on whether the disulfide bonds overlap in the primary amino acid sequence. Hence, it may be represented as *aabb*, *abba*, or *abab*.

Two topological properties were analyzed: symmetry and reducibility. Here, a disulfide pattern is considered symmetric if its topological mirror image is equivalent. For example, *aabbcc* and *abccab* are symmetric, but *aabcbc* and *abcacb* are not. All patterns with one or two disulfides are symmetric. A disulfide pattern is considered reducible if the peptide backbond can be cut such that both both pieces contain disulfide bonds but no disulfide bonds connect the two pieces. For example, *aabb* is reducible to *aa* and *bb*, but *abba* and *abab* are irreducible.

To analyze naturally occurring proteins, a data set was derived from 62 independent, complete, disulfide-containing proteins in the PDB and from 186 independent, disulfide-containing proteins of known connectivity from the NBRF protein sequence database. Upon observing the topological patterns present in these proteins, two interesting properties emerged: First, the number of reducible patterns was significantly greater than that expected by chance. If one disulfide pattern were as likely as another, the fraction of reducible patterns would be expected to decrease as the number of disulfides increased. However, the fraction of reducible patterns remained relatively constant, or even increased, with increasing number of disulfides. Second, the number of symmetric patterns was also greater than expected by chance (Table 1).

number of disulfides	reducible expected	reducible found	symmetric expected	symmetric found
2	0.33	0.61 (20/33)		
3	0.33	0.52 (22/42)	0.47	0.52 (22/42)
4	0.30	0.60 (12/20)	0.24	0.60 (12/20)
5	0.25	0.53 (10/19)	0.09	0.26 (5/19)
6	0.21	0.67 (6/9)	0.03	0.44 (4/9)
>6	< 0.2	0.74 (14/19)	< 0.01	0.16 (3/19)

Table 1. Topological properties of disulfide-containing proteins.

The relevance of these observations is difficult to evaluate. It would be easy to dismiss these results as predictable, since disulfide bonds were already shown to preferentially form between close cysteines (Thornton, 1981). A disulfide between two distant cysteines is likely to make a disulfide pattern irreducible; so, if these distant disulfides are rare, then the disulfide pattern distribution would be expected to more reducible patterns than expected by chance. On the other hand, this discovered distribution of reducible patterns might be taken as evidence for

the tendency for disulfide bonds to form modularly. Also, the abundance of symmetric disulfide pattern distributions may be misleading, as it could simply reflect the abundance of reducible patterns and a tendency for reducible patterns to also be symmetric.

In a further analysis, each reducible pattern was broken down into its irreducible components (Table 2). Interestingly, the subpattern *abab* is significantly more prevalent than the subpattern *abba*. While the authors did not pursue this trend further, it would might be useful to study this further, as it may reflect the process by which disulfide bonds are formed.

pattern	occurrences	pattern	occurrences
aa	75	*abcbac*	1
abab	48	*abbcca*	3
abba	15	*abccab*	1
abcbca	15	*abcacb*	1
abbcac	5	*abcbcdaede*	1
abaccb	2	*abcdefefcggbda*	1
abcabc	4	*abccdadeeffggb*	1

Table 2. Occurrence of reducible topological patterns in disulfide-bonded proteins.

Entropic and Diffusional Models of Disulfide Connectivity

One interesting effort to analyze and classify disulfide connectivity was based on two competing theoretical models of the entropic stabilization caused by disulfide bonds (Harrison and Sternberg, 1994). Disulfide bonds contribute to the thermodynamic stability of a protein by reducing the conformational space available to the protein. If this thermodynamic stabilization is the dominant role for disulfide bonds, it seems plausible that evolutionary pressure would favor proteins whose disulfides provide the maximum stabilization. This is considered the entropic model in this study.

On the other hand, a random diffusive folding process will also favor certain disulfide connectivities. If disulfide bonds form between any two cysteines that come in close contact, disulfide bonds would be more likely to form between nearby cysteines on the polypeptide chain. Although the disulfides can be later reshuffled by protein disulfide isomerase, there may be an advantage in proteins that tend to form the correct disulfide connectivity in a quick manner, with minimal assistance from PDI. This is the basis for the diffusional model.

Luckily, the mathematical treatment of the entropic model and the diffusional model are intimately related. A full derivation is beyond the scope of this thesis, but a brief summary follows. To simplify, the polypeptide chain is modeled as a chain of identical monomers with no steric hindrance, and thus no restraints on bond angles. It can be shown that the probability of two cysteines being close enough to form a disulfide is given by:

$$P_{pair} = \Delta V \left(\frac{3}{2\pi b^2} \right)^{3/2} N^{-3/2}$$

where ΔV is the "volume of tolerance", or the volume within which two cysteines are considered close enough to form a disulfide (estimates range between 5 and 58 Å^3); b is the distance between monomers, or the distance of the virtual C_α-C_α bond (3.8 Å); and N is the number of monomers (amino acids) between the two cysteines.

This formula can be extended to describe the probability of multiple cysteines being close enough to form a given disulfide connectivity:

$$P = (\Delta V)^n \left(\frac{3}{2\pi b^2} \right)^{3n/2} |\mathbf{A}|^{-3/2}$$

where n is the number of disulfide bonds, and $\mathbf{A}$ is an nxn matrix with elements defined by

$$a_{ij} = \sum_{h=1}^{l} \psi_{ih} \psi_{jh}$$

where ψ_{mk} is equal to one if residue k is inside the loop formed by disulfide bond m, or zero otherwise; and l is the length of the protein sequence. With the probability of a disulfide connectivity defined, the entropic stabilization is simply $\Delta S = R\ln(P)$.

The probability P can be used to calculate the likelihood of a disulfide connectivity. For the diffusion model, the likelihood of a given connectivity is simply proportional to P. For the entropic model, the likelihood of a connectivity is proportional to $1/P$.

A data set was produced from the Swiss-Prot database, using only sequences with disulfide connectivities firmly established by experimental means. Homologous sequences were removed using FASTA with a 25% sequence homology cutoff, and the BLOCKS database was used to remove proteins sharing a common sequence motif. This data set, consisting of 186 protein sequences, was divided into three groups according to sequence length: SL1 (length < 72), SL2 (72 ≤ length < 193), and SL3 (length ≥ 193). For each protein, the diffusional

probability *P* was calculated for each possible connectivity. The possible connectivities were sorted in order of increasing *P* and split into three groups of equal size, P1 (connectivities with low *P*), P2 (medium *P*), and P3 (high *P*). The true connectivity of each protein will be found in either P1, P2, or P3. Thus, a protein is classified as P1, P2, or P3 (Table 3).

	SL1	SL2	SL3	total
P1	27	14	3	44
P2	8	9	3	20
P3	3	13	34	50
total	48	36	40	124

Table 3. The classification of proteins according to sequence length and category of connectivity.

A correlation was observed, such that the shorter proteins (SL1) were more likely to have a true connectivity following the entropic model (P1), and longer proteins (SL2) were more likely have a connectivity following the diffusional model (P3). Furthermore, the distribution of the SL1 proteins coincide well with the expected distribution according to the entropic model, and the distribution of SL2 proteins coincide well with the expected distribution according to the diffusional model (data not shown).

Despite the simplifications involved in calculating the probability of a disulfide connectivity, these results show a remarkable pattern that may be applicable to the prediction of connectivity of an uncharacterized protein.

Sequence-Based Disulfide Connectivity Prediction

To this date, only attempt has been made to predict disulfide connectivity (Fariselli and Casadio, 2001). The predictions are based on the model that the four amino acid residues immediately flanking both cysteines in a disulfide bond interact. An amino acid "contact potential" can be calculated from known structures and then used to predict the connectivity of novel proteins.

First, a data set was created from the set of all Swiss-Prot entries containing a reference to the PDB (and thus have an experimentally determined structure) and at least one annotated disulfide bond. Proteins with disulfide bonds labeled "probable", "potential", or "by similarity" were removed. This produced 726 protein entries, which were grouped into four sets such that

the sequence homology among the different sets was ≤ 30%. Thus, they effectively created four data sets, and they claim that this allows them to cross-validate their method.

The basic procedure used to make a prediction is that of *maximum weight perfect matching*, a well-studied problem in graph theory. A fully-oxidized disulfide connectivity is an example of a *perfect matching*, because every cysteine is paired with exactly one other cysteine. If every possible cysteine-cysteine pair can be assigned a *weight*, then the optimal disulfide connectivity is the one where the the sum of weights is maximized. They take advantage of a published algorithm for efficiently finding the maximum weight perfect matching in a manner that guarantees the maximum weight without sampling every possible connectivity. It should be noted that this algorithm only makes their search in connectivity space quicker; it does not affect the accuracy of their predictions.

The challenge is to find an effective definition of the weights between each possible pair of cysteines. They use a contact potential *U(k, l)* for every possible pair of amino acids *k* and *l* and a five-residue long window centered on each cysteine. Thus, the weight for the connection between cysteine *i* and cysteine *j* is

$$w_{i,j} = \sum_{k \in Si} \sum_{l \in Sj} U(k,l)$$

where *Si* is the set of five residues immediately flanking and including cysteine *i*, and *Sj* is the set of five residues immediately flanking and including cysteine *j*. Thus, the interaction of 25 potential residue-residue contacts are considered, even though a large fraction of these residues are not in close contact in the folded protein. Four derivations of the contact potentials *U(k, l)* were used, and the prediction quality of each was evaluated separately.

To evaluate the prediction quality, the connectivity of every protein in the data set was predicted. Since the number of possible connectivities is a function of the number of disulfides, the data set was subdivided according to the number of disulfides. Only chains containing two, three, four, or five disulfides were considered. For each subset, and for each of the four contact potentials, the performance of the predictor was compared against a random predictor in two ways (Table 4).

The first contact potential tested, the Mirny-Shakhnovich potential (MS) was a potential previously developed for protein folding problems (Mirny and Shakhnovich, 1996). Its prediction accuracy was indistinguishable from a random predictor (Table 4). Since the potential

was derived from the whole protein structure, the authors suggest that a contact potential derived specifically from disulfide bond environments might perform better.

So, the second contact potential was derived by constrained optimization (CO) on the data set. The *U(k, l)* was computed that maximizes the difference in weight sums between the correct connectivities and the incorrect ones, while constraining the mean and dispersion. This approach performed somewhat better than the random predictor.

# of disulf. bonds	# of chains	Q_p					Q_c				
		(RP)	MS	CO	OR	EG	(RP)	MS	CO	OR	EG
2	158	0.333	0.34	0.42	0.46	0.56	0.333	0.35	0.42	0.46	0.56
3	153	0.067	0.05	0.09	0.17	0.21	0.200	0.19	0.21	0.29	0.36
4	103	0.010	0.09	0.05	0.11	0.17	0.142	0.14	0.21	0.31	0.37
5	44	0.001	0.0	0.0	0.0	0.02	0.111	0.12	0.14	0.17	0.21

Table 4. The prediction accuracy of four contact potentials, compared with a random predictor (RP). Q_p is defined as the fraction of total connectivities correctly predicted, while Q_c is the fraction of individual disulfide bonds correctly predicted. MS = Mirny-Shakhnovich, CO = constrained optimization, OR = odd-ratio, EG = Monte Carlo optimization.

The third contact potential was a straightforward odd-ratio (OR) computation. Each *U(k, l)* was calculated as the ratio of true disulfides containing *k* and *l* in their environment to possible but false disulfides containing *k* and *l*. This potential was consistently more accurate than the random predictor or the CO potential (Table 4).

The final contact potential was computed with a Monte Carlo simulated annealing optimization (EG) of the OR potential. The algorithm was designed to create a contact potential that maximizes the total number of correctly predicted connectivities. It is the most accurate of the four potentials, with a Q_p that is an order of magnitude better than the random predictor for chains containing four or five disulfides.

The major flaw with this study is that there is no way to estimate the performance of their prediction algorithms on a novel, uncharacterized protein. These results are declared in a context where the “learning” data set is the same as the “testing” data set, except with the MS potential, which performed poorly. Since each contact potential *U(k, l)* has 210 parameters, it is likely that the better-than-random predictions are based more on memorization of the correct disulfides. A much more accurate estimate of prediction accuracy could be attained with a “jacknife”

procedure, where a separate set of contact potentials is derived from all the chains in the data set except the one to be predicted.

Patterns and Classification of Disulfide Bonds

This thesis describes a search for conserved patterns in the sequence environment of disulfide bonds. Previous work described a weak correlation between the oxidation state of a cysteine and the presence or absence of certain residues at certain positions relative to the cysteine (Fiser *et al*., 1992; Fariselli *et al*., 1999). Perhaps the correlation could be improved by considering the possiblity that there are different classes of cysteines or disulfide bonds, each of which might have its own characteristic sequence environment. This is an attractive possibility because it could explain the roles of the multiple PDI homologs present in eukaryotic cells: each homolog could preferentially form disulfides from certain classes of cysteines.

Several possible classification schemes were explored. First, the inherent asymmetry of disulfide bond formation was considered. Disulfide transfer from PDI to a protein substrate is necessarily a two-step, asymmetrical process. First, a thiolate anion on the substrate nucleophilically attacks a disulfide on an oxidized PDI, forming a mixed disulfide. Second, a different thiolate on the substrate attacks the disulfide, leaving the substrate oxidized and PDI reduced (Figure 3).

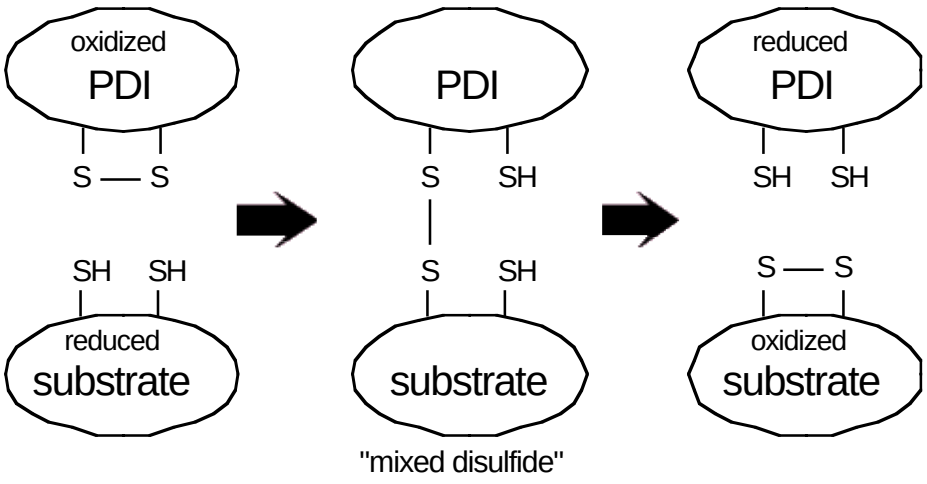


Figure 3. Steps involved in electron transfer from a reduced substrate to PDI.

So, the cysteines can be divided into two classes: those that form the mixed disulfide with PDI, and those that resolve the mixed disulfide to form the native disulfide bond.

The sequential distance between the cysteines was also used as a means of classifying disulfide bonds. Among well-characterized proteins, there is a preference for disulfide bonds formed between cysteines separated by 12-20 residues. Perhaps one PDI homolog catalyzes these nearby disulfide bonds, while another less abundant homolog catalyzes more distant disulfides.

Another way of classifing disulfide bonds is based on their relationship to other disulfides or cysteines in the protein. It is relatively uncommon to find an intervening cysteine between the two involved in the disulfide bond. Since the nascent polypeptide enters the ER linearly, a cysteine that is not meant to be connected to the next available cysteine might be associated with an identifiable motif that is recognized by a PDI homolog with a slower resolution time.

Finally, an entire protein can be classified based on the number of disulfide bonds relative to its size. Disulfide-rich proteins may require the presence of more or different PDIs than disulfide-sparse proteins (Rietsch *et al*., 1996).

Each of these classification schemes was used on the disulfide bonds from the set of nonhomologous proteins. The sequence environment of the cysteines from each class were compared, and any strongly overrepresented residues in the sequence environments were studied further to determine if they corresponded to a conserved sequence.

Since it would be very useful to be able to predict the disulfide connectivity of an uncharacterized protein, several prediction methods were explored. The prediction methods were based upon either the sequence environment surrounding the cysteines or the number of residues separating disulfide-bonded cysteines.

METHODS

Assembly of Data Set

Ideally, the data used in this study is the largest possible set of nonhomologous proteins with well-defined structures. The procedure used to produce the data set is automated and reproducible, so that newly solved protein structures can be easily incorporated into the analysis.

The starting point for the data set is the complete Protein DataBank (PDB) (Berman *et al*., 2000). As of January 2001, there were over 13000 protein structures in the database, most of which are redundant (e.g. there are 705 structures of lysozyme with various mutations). In order to analyze only a representative sample of unrelated proteins, the PDB-Select.25 list was used (Hobohm and Sander, 1994). The PDB-Select.25 list is a subset of the PDB in which each protein structure file is first divided into separate polypeptide chains. Each chain is aligned against each other chain, and if the alignment distance is positive, the chain of lower quality is removed. The alignment distance is based on a scoring mechanism designed to recognize sequences that are likely to share the same fold (Abagyan and Batalov, 1997). The October 2000 release of PDB-Select.25 contains 1427 nonhomologous chains.

For this study, the PDB structure file corresponding to each chain was parsed with a computer program ("PDBparse"), and information about the type of protein, its amino acid sequence, and its disulfide bonds was extracted into a FileMaker database (Table 5).

HEADER	contains title, date, and unique PDB ID
COMPND	description of compound crystallized; usually name of protein
SOURCE	source of protein, and expression vehicle if different
SEQRES	amino acid sequence of each chain
SSBOND	the locations of disulfide bonds

Table 5. Fields extracted by PDBparse.

From the SEQRES and SSBOND fields, the location of free cysteines was determined. A cysteine connectivity "map" was calculated for each chain. The map is a one-dimensional representation of a single polypeptide, from N-terminus to C-terminus (Figure 4). A number inside a pair of brackets represents the serial number (as defined in the PDB file) of a disulfide-bonded cysteine, so that two cysteines with the same serial number form a disulfide bond. The

letter “C” represents a cysteine with a free thiol. The numbers outside the brackets represent the number of non-cysteine residues between the cysteines.

A 55[1]136[2]13[2]9[3]6[4]8[4]6[3]21[5]5[5]29[1]42[C]80

B

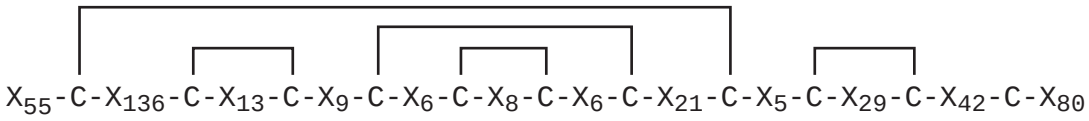


Figure 4. Representations of disulfide connectivity. A) The internal representation of DSBMax. B) A standard graphical representation, where “X” indicates any amino acid.

Each chain was classified as eukaryotic or prokaryotic based on the SOURCE field in the PDB entry and the NCBI taxonomy database (Wheeler *et al*., 2000). Since viruses do not have their own oxidative folding resources, viral proteins were included in their host’s taxon. Of the 1341 proteins in PDB-Select, 822 were eukaryotic, 493 were prokaryotic, and 26 were *de novo* synthetic peptides. The synthetic peptides were not included in the subsequent analysis.

In general, the disulfide bonds of eukaryotic proteins are formed in the ER, so a protein without the appropriate signal sequence would not be expected to contain disulfide bonds. From 50 randomly selected chains from eukaryotic organisms, 19 were targeted to the ER, as determined by their subcellular localization or by the presence of a signal sequence. Of the 31 nonsecretory proteins, only 1A9N (human spliceosomal U2A’) contained a disulfide bond, and this is an interchain disulfide that is probably an artifact of crystallization or expression. Also, a protein that folds in the ER will usually have its cysteines oxidized to the greatest extent possible; that is, it usually won’t have more than one free thiol. Of the 19 secreted proteins, 1A8E (human serin transferrin) and 1AK0 (penicillium P1 nuclease), and 1AFP (aspergillus antifungal protein) contained more than a single free thiol. Inspection of their structures revealed that 1A8E, 1AK0, and 1AFP actually contain only disulfide bonded cysteines, but for some reason these disulfides weren’t explicitly listed in the PDB files. While the data set may contain a few misclassified cysteines, these seem to be in the minority, and any conserved sequence motifs should be identifiable despite these errors.

DSBMax, a Pattern Finding Program

"DSBMax" is a program written to help search for common themes in the amino acid sequence environment of disulfide bonds. Previous research has shown that the presence of certain residues at certain positions relative to a cysteine can be used to predict the oxidation state of that cysteine (Fiser *et al*.,1992; Fariselli *et al*., 1999). DSBMax builds on this research by accounting for the inherent asymmetry of disulfide bond formation. Since disulfides are necessarily formed in a two-step reaction, it is plausible that the two cysteines play different roles in disulfide formation: the first cysteine may be part of a motif that is able to form a mixed disulfide with an oxidase (e.g. PDI), while the second cysteine may be part of a motif that is able to resolve the mixed disulfide to complete the transfer of the disufide bond to the protein. Unfortunately, there is no way to know *a priori* which cysteine is acting in the first step and which is acting in the second step. By assuming that each disulfide bond contains a cysteine in a motif from each class, DSBMax attempts to divide the cysteines into two classes such that each class has the highest possible conservation of the cysteine's flanking residues.

As input, DSBMax reads a list of amino acid 11-mers, each with a cysteine in the center (6th) position. The window size was set at 11 residues somewhat arbitrarily, but previous work (Muskal *et al*., 1990) found that performance of oxidation state prediction algorithms was not improved by using a window size larger than 11. DSBMax divides the 11-mers into two arrays: one contains the set of flanking sequences of cysteines with free thiols; the other contains two sets of flanking sequences for each disulfide bond.

DSBMax Scoring

Any set of 11-mers can be assigned a conservation score. DSBMax simply counts the occurrences $N_{obs}(r,p)$of each residue r in each position p and compares to the number expected based on the total abundance of the residue in all proteins being considered $N_{exp}(r)$. If the distribution of residues were random, N_{obs} would be expected to follow a binomial distribution centered at N_{exp}. So, the local conservation score $s(r,p)$ is defined as the deviation of N_{obs} from N_{exp} divided by the standard deviation. A positive (negative) local conservation score represents a residue that is found more (less) often at a given position than expected by chance. For a given set of data, the total conservation score S is defined as the sum of the absolute values of the local conservation score. Thus, S is a measure of how "unexpected" is a given set of data. Since

$N_{obs}(r,p)$ and $s(r,p)$ define two matrices, DSBMax displays them as such: $N_{obs}(r,p)$ is displayed numerically, and $s(r,p)$ is converted to a color displayed in the background.

The distribution of amino acids that defines $N_{exp}(r)$ was calculated from the sequences of all eukaryotic chains that contain at least one disulfide bond. This set contains nearly 40,000 residues in 256 chains. To account for the differing number of disulfides found in each chain, each sequence were weighted by the number of disulfide bonds it contains. This distribution defines the baseline amino acid composition for the disulfide-containing chains being studied. This distribution is shown along with the unweighted distribution and the distribution of all eukaryotic chains in the PDB-select set (Table 6).

residue	disulfide-containing chains (weighted)	disulfide-containing chains (unweighted)	all eukaryotic chains
A	0.0613	0.0642	0.0695
C	0.0570	0.0411	0.0233
D	0.0549	0.0559	0.0568
E	0.0513	0.0540	0.0642
F	0.0374	0.0392	0.0412
G	0.0801	0.0757	0.0694
H	0.0209	0.0226	0.0251
I	0.0452	0.0478	0.0534
K	0.0524	0.0548	0.0650
L	0.0677	0.0756	0.0876
M	0.0159	0.0178	0.0219
N	0.0569	0.0541	0.0463
P	0.0538	0.0526	0.0478
Q	0.0391	0.0405	0.0404
R	0.0477	0.0471	0.0499
S	0.0784	0.0754	0.0668
T	0.0645	0.0631	0.0563
V	0.0577	0.0613	0.0655
W	0.0184	0.0180	0.0146
Y	0.0391	0.0392	0.0350

Table 6. Amino acid abundance in disulfide-bonded proteins compared with all proteins.

The DSBMax conservation scores are difficult to interpret in an absolute sense, but controls can be used to estimate their significance. For a typical data set an equivalent number of artificial “disulfides” can be created from random bits of sequence from the same chains from which the disulfides came. Multiple artificial control sets can be created in order to establish a mean and standard deviation.

DSBMax and Simulated Annealing

For a set of n disulfide bonds, there are 2^{n-1} ways of dividing the cysteine-flanking sequences into two sets, because each cysteine in the disulfide bond can placed into either set, and then the other cysteine must be placed in the opposite set. The sequences corresponding to a single disulfide bond can be pulled from the two sets and replaced in the opposite sets, which I define as a *swap*. Two swaps of the same disulfide therefore cancel out. DSBMax attempts to find the combination of swaps that maximizes the sum of the total conservation scores for each set. This is basically a maximization problem over n-1 variables. Large dimensional maximization problems are usually tricky to solve, but this one should be tractable for two reasons: first, the limitation of each variable to binary values greatly simplifies the maximization. Second, any division of sequences is at least half correct; that is, the number of swaps necessary to maximize a random division is always less than $n/2$.

The algorithm used by DSBMax is straightforward. First, the initial sets are defined from the list of disulfide bonds by putting the N-terminal cysteine in one set and the C-terminal cysteine in the other. This is arbitrary, but it defines a convenient reference point. Next, the sets are randomly "shuffled" by swapping each disulfide bond with a probability of 1/2. Then, a simulated annealing (SA) algorithm is used to maximize the total conservations score.

SA is a heuristic local search algorithm in which unfavorable transitions are accepted with a probability derived from the Boltzmann distribution, $\exp(\Delta E/k_B T)$. In a manner analagous to the annealing of a solid, the temperature T is slowly lowered until the the system reaches a low energy state (Aarts and Korst, 1989). In the case of DSBMax, a "transition" means the swapping of one disulfide bond. The algorithm can be summed up as follows: The temperature T is initially set to a level *startTemp*. A disulfide pair is picked at random. If swapping that pair results in a higher score, the swap is made. If swapping the pair results in a lower score, the swap is made with probability $p=\exp(\Delta S/T)$, where ΔS is the difference in score incurred by the swap. This is done *numRep* times, each time with a disulfide pair picked at random. The temperature T is lowered incrementally by a factor of *tempMultiplier*, and again *numRep* disulfides are picked at random and swapped if they meet the proper conditions. This procedure is repeated until T is below a final temperature *stopTemp*.

The parameters *startTemp*, *stopTemp*, *numRep*, and *tempMultiplier* were determined by trial and error. *startTemp* was set so that 95% of unfavorable transitions were accepted.

stopTemp was set so that less than 0.05% of unfavorable transitions were accepted. Thus, in analogy to the annealing of a solid, the disulfide system is initially in a completely random "molten" state, and at the end of the SA procedure the system is in a completely ordered, low-energy state. The parameters *numRep* and *tempMultiplier* were set to values that were capable of allowing the system to reach its highest possible score without requiring excessive computer time. The values used are summarized in Table 7.

startTemp	10
stopTemp	0.01
numRep	2000
tempMultiplier	0.99

Table 7. Parameters used in DSBMax simulated annealing.

Predictors of Disulfide Connectivity

All prediction programs were written in the Perl programming language. A modular, object-oriented approach was used for maximum flexibility and reusability. Most modules should work under any operating system, but some modules that read or write to disk use MacPerl-specific functions. All predictors follow the following algorithm:

1) Get amino acid sequence (must be one-letter code).
2) Generate a list of all possible disulfide connectivities.
3) For each possible connectivity, generate a score.
4) Output the connectivity with the highest score.

The means of generating a score is unique to each predictor; these are described below.

"PredNResBtwn" is a predictor that evaluates a possible connectivity in terms of the number of residues between disulfide-bonded cysteines (the "sequential distance"). Upon initialization, it expects to read a file "DistanceScores" containing probabilities for each possible sequential distance. DistanceScores was created by fitting the observed sequential distances (Figure 17) in the data set to a simple function with a minimal number of parameters. Specifically, the probability p of a disulfide bond of sequential distance n is determined by $p=(n-0.5)\exp(-0.072n)+3$. The score assigned to a connectivity is equal to the product of the p for each disulfide bond.

"PredSeqClassDist" is a predictor that evaluates a given connectivity based on an assumed correlation between the flanking sequence and the sequential distance between two disulfide-bonded cysteines. Upon initialization, it reads the entire data set of proteins with known connectivity and classifies each disulfide as Close (sequential distance < 20 residues) or Distant (sequential distance ≥ 20 residues). Each cysteine is further classified as N-terminal or C-terminal (relative to its disulfide-bonded partner), thus creating four classes of cysteines. For each class, an amino acid composition matrix is created for the ten flanking residues. To score a possible connectivity, each cysteine in an unknown protein is compared to the four class matrices, and it is predicted to belong to the class with the best fit. The score assigned to a possible connectivity is equal to the number of cysteines with a class (determined by the connectivity being analyzed) that is the same as the predicted class (determined by its flanking sequence).

"PredSeqClassDistFair" is a modification of the above predictor designed only for testing purposes. It works the same way, but the four class matrices are dynamically adjusted, so that the chain being analyzed is removed from the matrices. This predictor is a "jackknife" test because the training set is effectively separated from the testing set. Unlike PredSeqClassDist, the performance of this predictor on the test set is likely to reflect its performance on a novel protein not in the test set.

"PredEntropy" is a predictor that evaluates a given connectivity based on its theoretical entropic stabilization of the protein structure (Harrison and Sternberg, 1994). To compare different possible connectivities within the context of a single chain of length l containing n disulfide bonds, it is sufficient to calculate $|\mathbf{A}|$, where $\mathbf{A}$ is the nxn matrix with elements

$$a_{ij} = \sum_{h=1}^{l} \psi_{ih} \psi_{jh}$$

ψ_{mk} is 1 if residue k is inside the loop formed by disulfide bond m, or 0 otherwise. A higher value of $|\mathbf{A}|$ corresponds to a greater entropic stabilization, so $|\mathbf{A}|$ is used as the score assigned to a connectivity.

"PredDiffusion" was designed to predict connectivity based on the diffusional model (Harrison and Sternberg, 1994). It is the opposite of PredEntropy; the score assigned to a connectivity is $-|\mathbf{A}|$.

"PredEntDiff" is based on PredEntropy and PredDiffusion, but takes into account the observation that smaller proteins are frequently found with connectivities that maximize entropic stabilization, while larger proteins are often found with connectivities that are more diffusionally accessible (Harrison and Sternberg, 1994). The sequence length is used to determine whether the PredEntropy model or the PredDiffusion model is used.

"PredAdjacent" is a predictor based on the observation that a large fraction of disulfide bonds are formed between two cysteines without any sequentially intervening cysteines (Figure 21). The score assigned to a connectivity is equal to the number of disulfides that have no intervening cysteines.

Evaluation of Predictors

The evaluation of a prediction method is a tricky problem in itself, as there are several possible measures of its success. Given a predictor and a data set of chains with known connectivities, we require a method of generating a numerical score to compare the quality of predictions against each other and against a random predictor.

The data set used to evaluate the predictors was derived from the data set used with DSBMax. Chains with interchain disulfides and chains with more than one non-disulfide cysteine were removed from the data set. Without these measures, the predictors would have be considerably more complicated. Also, chains with more than 10 cysteines were removed from the data set to minimize the computation time needed to consider a large number of connectivities. Trivial chains with fewer than three cysteines were removed. Thus, the data set contained 171 protein chains.

For a given predictor, two scores are assigned for each protein chain in the data set, Q_p and Q_c (Fariselli and Casadio, 2001). Q_p is 1 if the overall predicted connectivity is correct, and 0 if it is incorrect. Q_c is the fraction of cysteines that have been correctly assigned. The average Q_p and Q_c were calculated for each subset of chains with the same number of cysteines. The perfomance of predictors can be compared by comparing the overall Q_p and Q_c on the same data set.

RESULTS AND DISCUSSION

DSBMax: the Entire Data Set

DSBMax was first used to study the entire set of 751 eukaryotic intrachain disulfide bonds. Before SA optimization, DSBMax shows the distribution of each amino acid residue in each position relative to the cysteine. The sequence environment of disulfide-bonded cysteines is compared to that of cysteines with free thiols (Figure 5). The relatively small number of free thiols in this data set makes it difficult to draw conclusions about their preferred sequence environment. On the other hand, several potentially interesting features are apparent in the set of disulfide-bonded cysteines. Overall, there is an underrepresentation of the larger hydrophobic residues (Phe, Ile, Leu, Val, Trp). Some residues near the cysteine seem especially prevalent, especially Glu at -3, Gly at -2, Lys at -1, Pro at +1, Arg at -1 and +1, and Thr at -1. While these residues are overrepresented in these positions, they are far from a consensus, representing at best 10% of the observed residues in a given position.

Each disulfide-bonded cysteine can be classified according to its orientation in the disulfide bond (Figure 6). By comparing the DSBMax outputs from the N-terminal cysteines and the C-terminal cysteines, one can find several interesting differences between the two sets. For example, a Cys at +2, a Lys at +3, or a Tyr at -2 is more than twice as common in C-terminal cysteines as in N-terminal cysteines. Also, a Gln at -5 or -1 or a Gly at +1, +3, +4, or +5 is much more common in the N-terminal cysteines. The scores generated by DSBMax indicate that the sequence environment of the N-terminal cysteines is slightly less random than the sequence environment of the C-terminal cysteines.

To search for a potential conserved sequence present near only one of the two cysteines in a disulfide bond, the SA optimization routine of DSBMax was used on this set of disulfide bonds (Figure 7). The SA-optimized groupings of cysteines showed higher conservation scores (357 and 341) than the division of cysteines according to their orientation (212 and 198). Still, no highly conserved residues were observed; the most conserved residue was Gly at +2 in the first grouping, which was only present in 12% of the sequences in that group.

disulfide-bonded cysteines

disulfide-bonded cysteines

	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5
A	85	95	95	94	80	0	73	88	92	83	83
C	94	89	54	86	671502		61	85	56	89	93
D	89	73	81	79	76	0	87	92	78	75	75
E	69	77	105	74	81	0	96	78	74	79	73
F	48	41	47	54	38	0	51	49	46	50	48
G	113	128	129	149	102	0	117	140	129	131	123
H	33	23	34	29	32	0	34	26	36	29	30
I	55	64	56	46	55	0	67	46	61	43	57
K	91	80	93	88	103	0	91	96	86	96	87
L	79	82	74	101	90	0	82	84	89	82	76
M	23	23	22	24	16	0	14	18	21	23	18
N	92	95	87	92	87	0	78	91	82	89	86
P	79	79	79	78	69	0	111	99	80	71	78
Q	66	49	55	50	73	0	59	50	63	52	57
R	69	65	78	66	95	0	103	70	74	64	71
S	103	121	108	111	115	0	114	129	129	114	108
T	83	92	100	112	136	0	101	80	94	92	98
V	75	76	75	59	90	0	70	64	77	67	67
W	23	28	22	20	26	0	15	23	20	30	22
Y	52	60	64	67	65	0	59	51	48	51	40

Score: 204.494172410703

cysteines with free thiols

free-thiol cysteines

	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5
A	8	8	6	7	10	0	6	7	11	6	4
C	4	2	7	6	0	127	1	8	7	3	5
D	8	8	10	12	8	0	10	8	4	4	8
E	6	5	3	7	5	0	10	7	10	11	11
F	7	7	6	4	8	0	5	2	6	10	2
G	9	14	14	13	8	0	11	11	12	11	11
H	3	4	3	0	5	0	6	4	2	5	6
I	8	6	9	7	7	0	7	5	7	5	4
K	7	14	4	9	8	0	6	4	8	3	9
L	11	11	13	11	10	0	13	13	13	13	9
M	4	3	1	4	1	0	1	3	4	3	5
N	7	4	7	6	5	0	4	10	5	5	8
P	8	7	2	7	8	0	8	3	5	10	3
Q	4	2	4	3	2	0	3	6	5	1	9
R	4	3	6	4	7	0	5	6	3	10	4
S	10	5	9	6	12	0	7	12	5	6	9
T	7	5	11	7	7	0	10	5	7	9	5
V	9	5	8	5	10	0	11	7	4	9	10
W	0	5	0	1	1	0	2	3	4	1	3
Y	2	8	3	7	5	0	1	3	5	2	2

Score: 177.985260035779

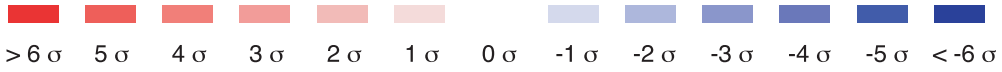


Figure 5. The sequence environment of cysteines from eukaryotic proteins containing at least one disulfide bond. Cysteines are classified as disulfide-bonded or free-thiol. Each matrix element displays the number of times each residue was found at each position relative to the central cysteine. The color indicates the deviation from the expected value, in standard deviation units.

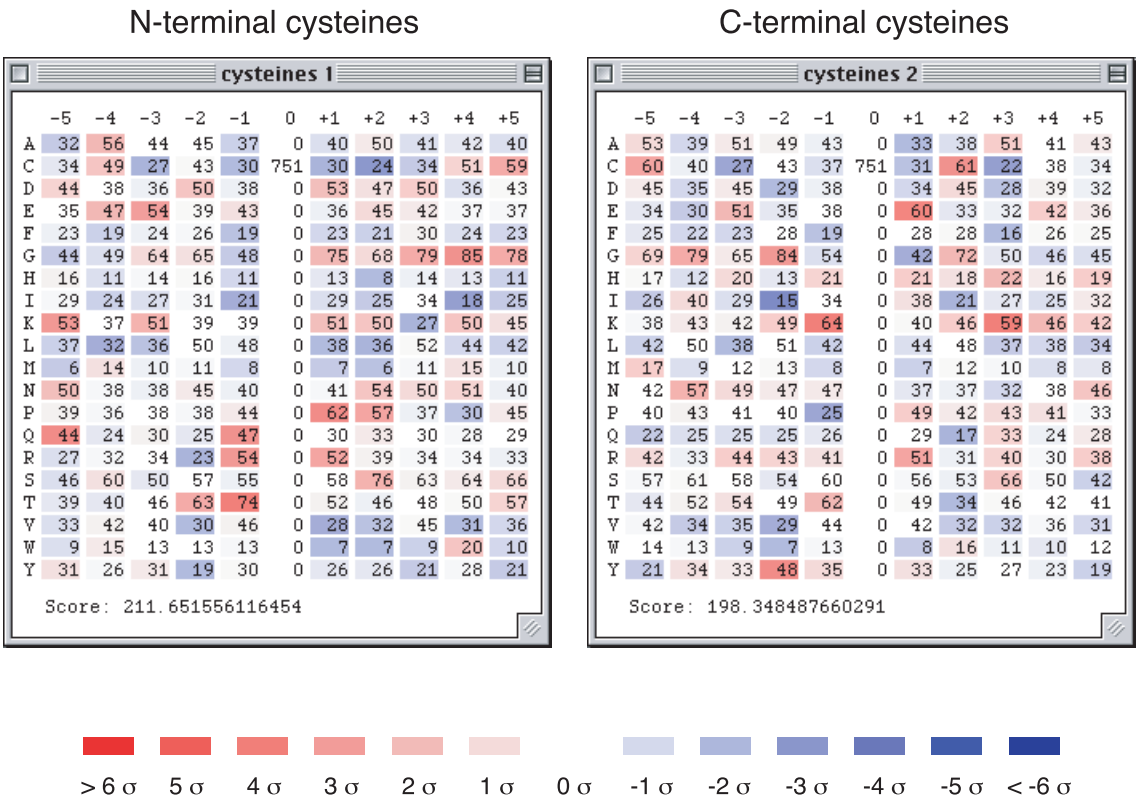


Figure 6. The sequence environment of disulfide-bonded cysteines from eukaryotic proteins. Cysteines are subclassified according to which cysteine in the pair is closest to the N-terminus or to the C-terminus.

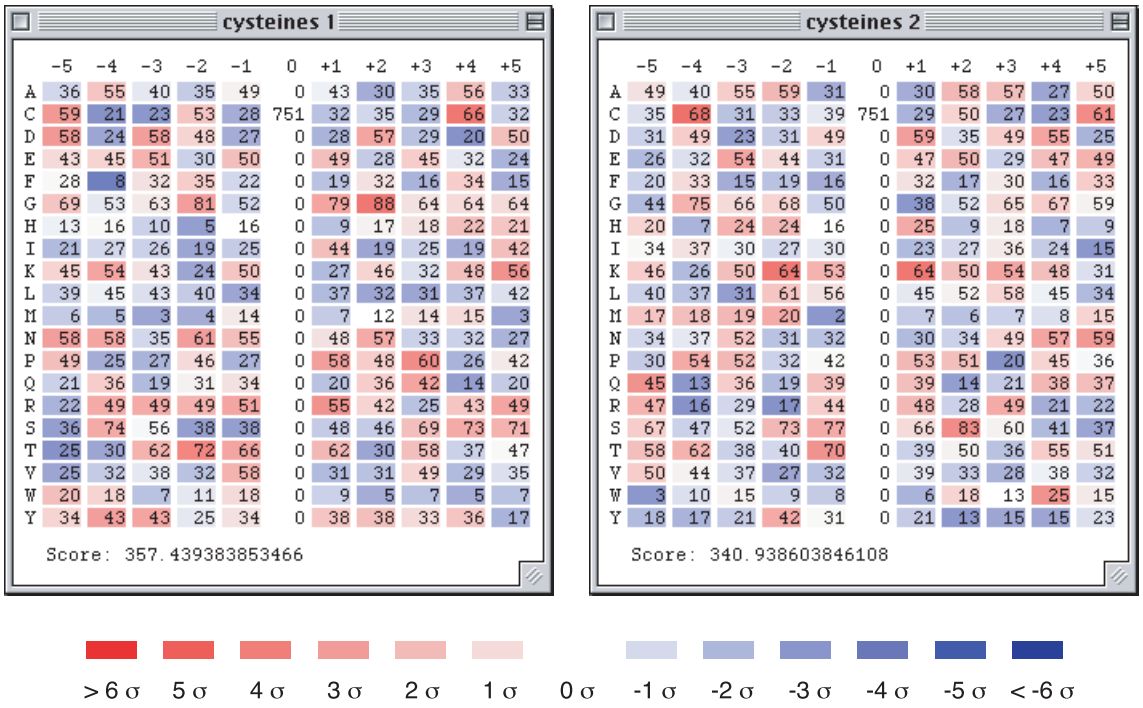


Figure 7. Simulated Annealing optimization of eukaryotic disulfide-bonded cysteine pairs.

Trends and Classification in the Data Set

In a search for broad trends and natural divisions in the distribution of disulfide bonds, 868 eukaryotic and 533 prokaryotic chains were analyzed in various ways. First, the number of cysteines per chain was plotted as a histogram (Fig. 8). Both eukaryotes and prokaryotes had a preference for even numbers of cysteines, which is due to the bias of secreted proteins against free thiols. Next, the distributions of intrachain and interchain disulfide bonds was plotted (Figures 9, 10). On the average, the eukaryotic chains were significanly more likely to contain disulfide bonds. The prokaryotic chains had no more than three intrachain disulfide bonds, with an average of 0.16 bonds per chain. The eukaryotic chains had up to 12 intrachain disulfide bonds, with an average of 0.9 bonds per chain. Interchain disulfide bonds were significantly less common in both eukaryotes and prokaryotes, with an average of 0.009 per chain in prokaryotes and 0.04 in eukaryotes.

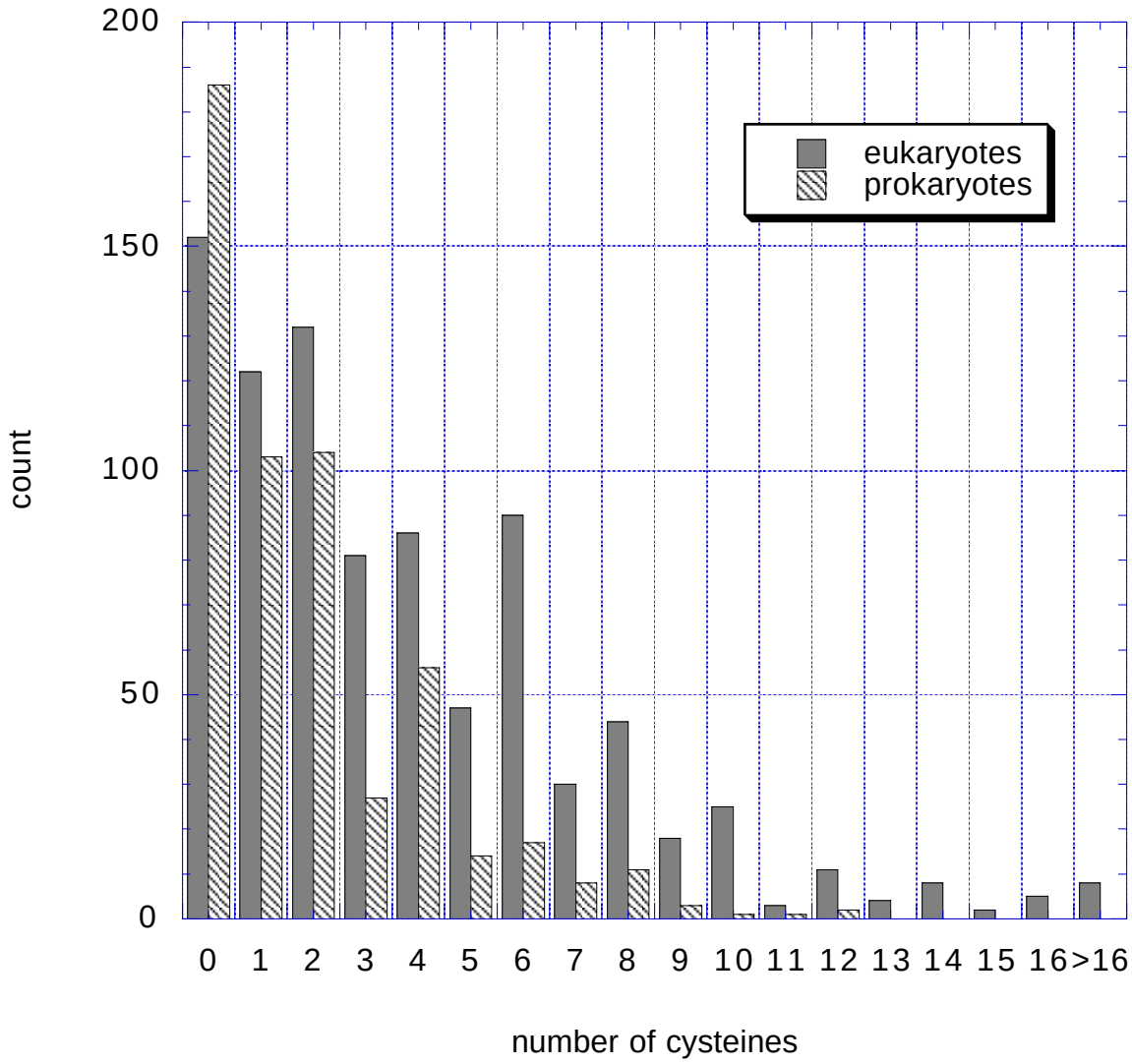


Figure 8. Histogram of the total number of cysteines per chain in eukaryotic and prokaryotic proteins.

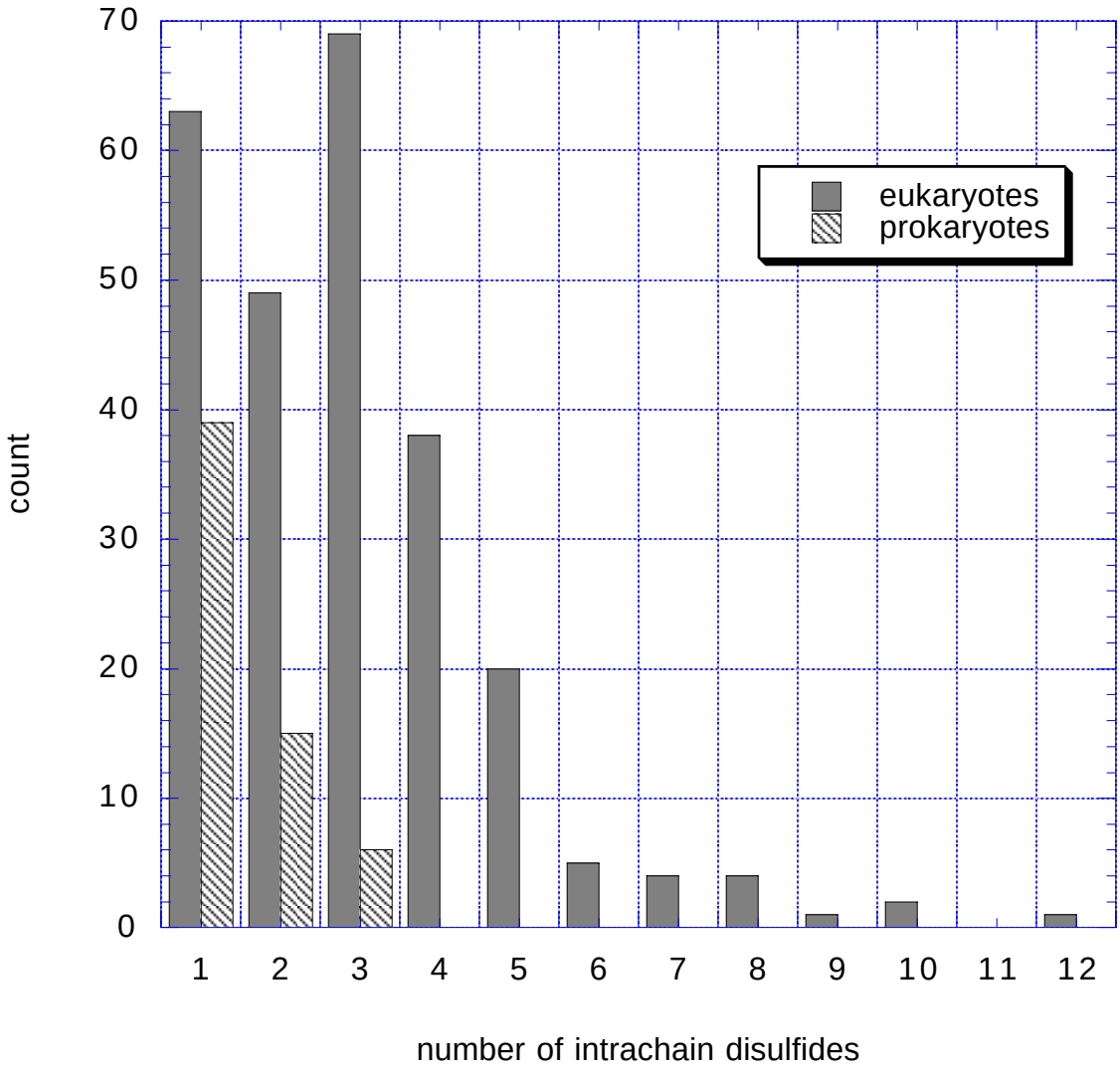


Figure 9. Histogram of the number of intrachain disulfide bonds per chain in eukaryotic and prokaryotic proteins. Not shown: 612 eukaryotic and 473 prokaryotic chains containing no intrachain bonds.

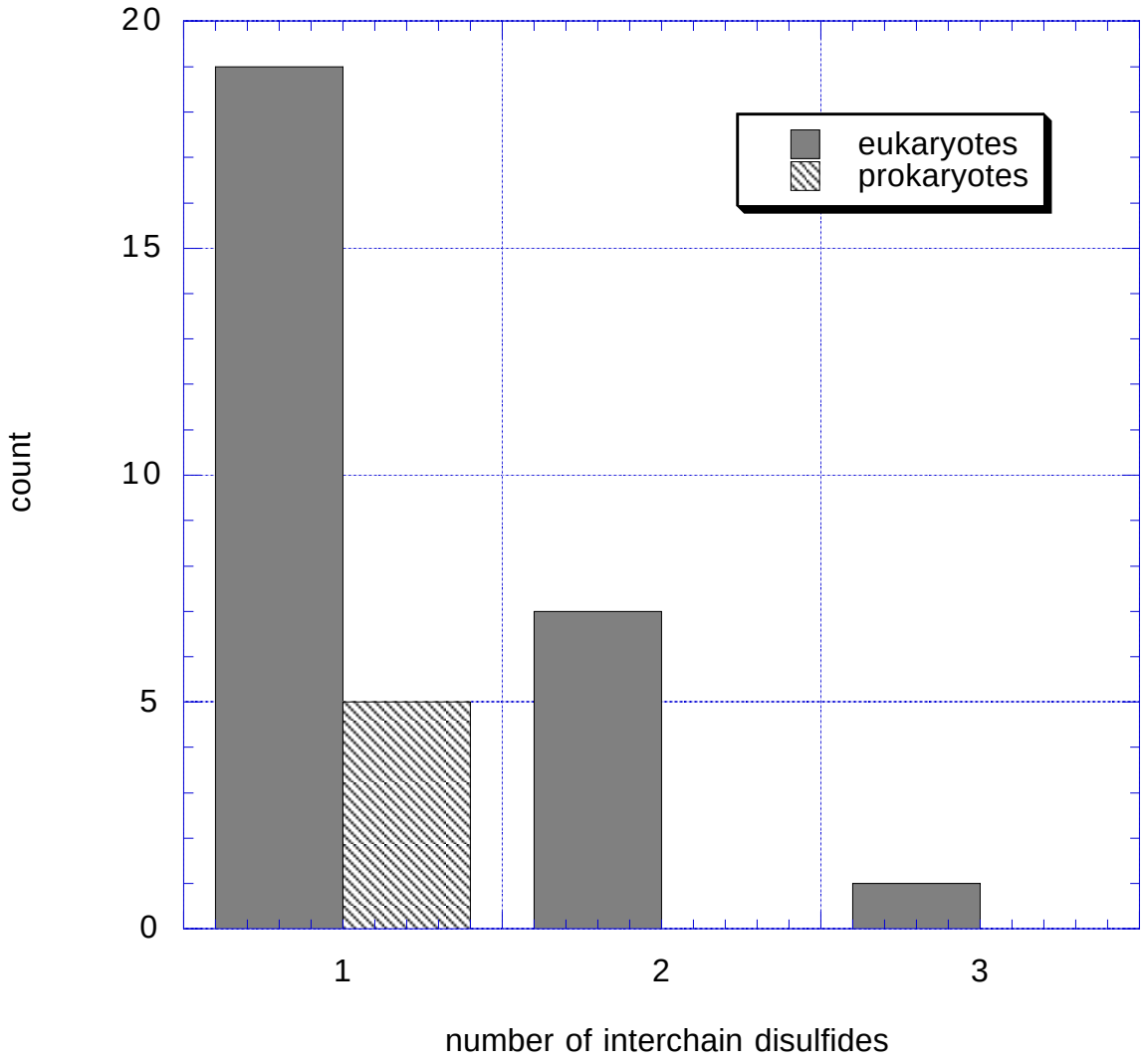


Figure 10. Histogram of the number of interchain disulfide bonds per chain in eukaryotic and prokaryotic proteins. Not shown: 841 eukaryotic and 528 prokaryotic chains containing no interchain bonds.

DSBMax: Disulfide Density

An interesting comparison can be made by plotting the average number of intrachain disulfide bonds per residue for each chain (Figure 11). While the prokaryotic chains exhibit the expected decaying exponential, the eukaryotic chains show a distinct bimodal character. The eukaryotic chains also have a predominant decaying exponential, but there are also a significant number of chains with a higher density of disulfide bonds. This suggests a a natural division of eukaryotic chains: those with a high density of disulfide bonds, and those with a low density.

The flanking sequences from 300 disulfides from proteins with a low disulfide density (< 0.05 disulfides/residue) and 451 disulfides from proteins with a high disulfide density (> 0.05 disulfides/residue) were analyzed with the scoring capabilities of DSBMax. First, all disulfide-bonded cysteines from chains in each class were compared (Figure 12). Interestingly, the high-disulfide-density cysteines show a tendency to be flanked by the posively-charged residues Lys and Arg. The high-disulfide-density cysteines also tend to have fewer hydrophobic residues in their surrounding environment. The low-disulfide-density cysteines have a high concentration of Glu at -3 and Thr at -1.

The low-disulfide-density cysteines and the high-disulfide-density cysteines were then subgrouped by their orientation relative to the N- and C-termini (Figure 13). Several noteworthy differences between the four subgroups are apparent. For example, in the low-disulfide-density disulfides, Tyr at -2 is five times as abundant in the C-terminal cysteines as in the N-terminal cysteines. Also, in the high-disulfide-density disulfides, Arg at -1 is favored by the N-terminal cysteines and Lys at -1 is favored by the C-terminal cysteines.

SA optimization was run on the two groups of disulfides (Figure 14). In the group of disulfides from low-disulfide-density chains, the second optimized subgroup has a large fraction of Glu at -3 and Tyr at -2 and -1, Thr at -1, and Ser at +2. The data was searched to check if these abundant residues are often found in the same sequence. All combinations of any two of the above residues plus a cysteine were checked, and none were found to be exceptionally common in the data set (data not shown). The most common of these sequences was EXTC, found 11 times (expectation = 6) in the set of low-disulfide-density disulfides and 17 times in the set of all eukaryotic disulfides.

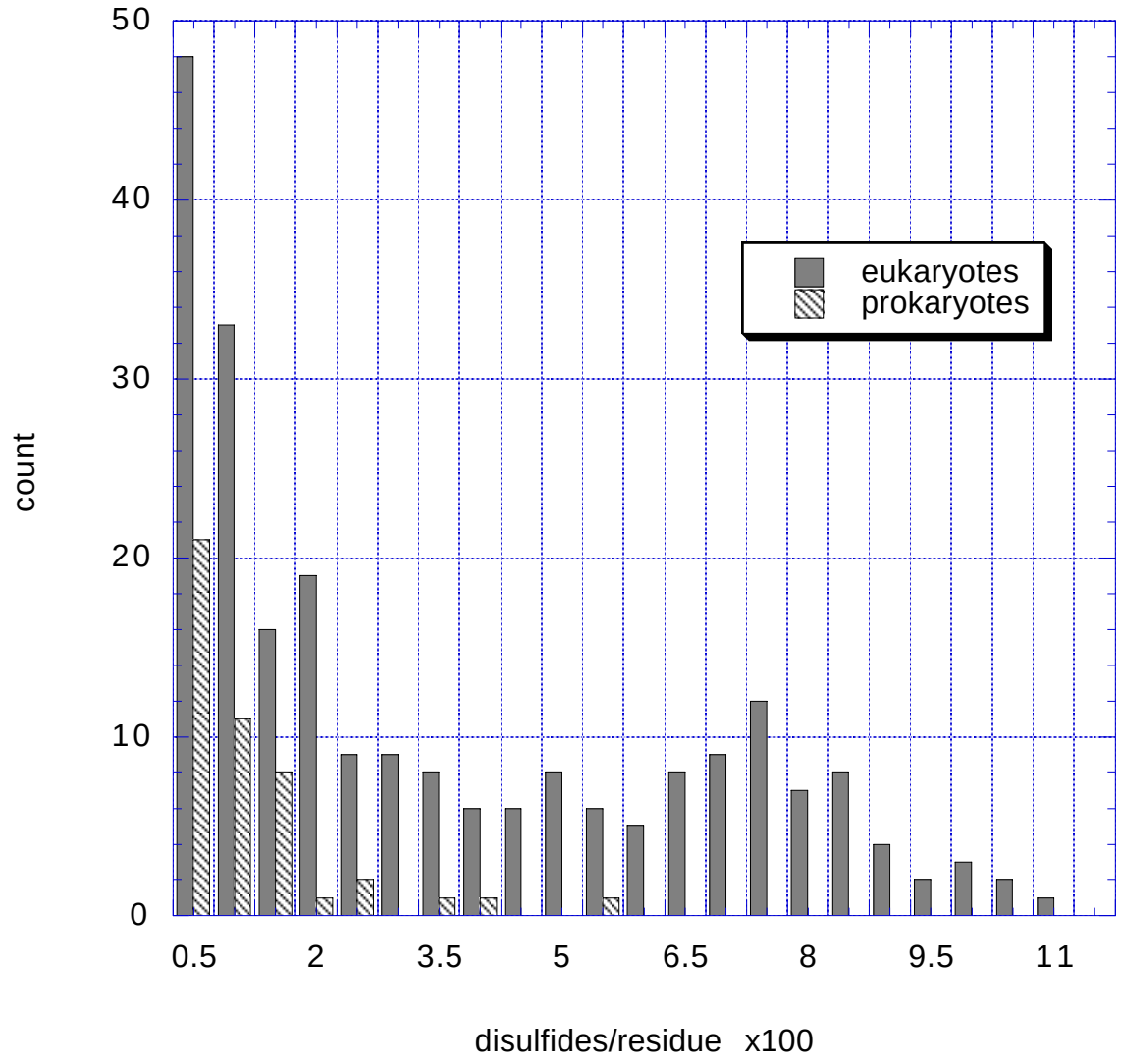


Figure 11. Histogram of the number of intrachain disulfide bonds per residue in eukaryotic and prokaryotic chains. Not shown: 612 eukaryotic and 473 prokaryotic chains containing no intrachain bonds.

low disulfide density

disulfide-bonded cysteines

	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5
A	35	46	30	36	31	0	29	41	36	32	44
C	17	12	6	8	6	600	5	8	7	12	15
D	35	31	37	31	32	0	43	38	30	38	30
E	34	29	52	30	27	0	44	32	30	29	35
F	20	19	23	33	15	0	25	23	20	19	30
G	45	56	47	52	45	0	46	62	45	53	51
H	13	13	16	17	17	0	14	12	16	17	15
I	20	27	24	19	14	0	24	22	26	17	25
K	38	27	34	26	34	0	20	21	30	34	24
L	44	38	36	59	51	0	42	43	46	43	43
M	12	11	12	6	8	0	7	7	11	9	9
N	44	35	29	31	31	0	31	39	40	35	26
P	24	28	33	30	30	0	45	34	26	32	29
Q	27	24	22	22	21	0	29	18	31	26	22
R	25	18	31	22	30	0	37	25	25	17	29
S	39	56	42	55	60	0	56	63	53	53	54
T	40	39	51	47	65	0	38	34	41	40	40
V	36	41	34	30	39	0	34	39	38	32	27
W	10	13	7	6	8	0	5	10	10	13	8
Y	23	21	24	33	33	0	19	18	18	26	16

Score: 208.801647890575

high disulfide density

disulfide-bonded cysteines

	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5
A	50	49	65	58	49	0	44	47	56	51	39
C	77	77	48	78	61	902	56	77	49	77	78
D	54	42	44	48	44	0	44	54	48	37	45
E	35	48	53	44	54	0	52	46	44	50	38
F	28	22	24	21	23	0	26	26	26	31	18
G	68	72	82	97	57	0	71	78	84	78	72
H	20	10	18	12	15	0	20	14	20	12	15
I	35	37	32	27	41	0	43	24	35	26	32
K	53	53	59	62	69	0	71	75	56	62	63
L	35	44	38	42	39	0	40	41	43	39	33
M	11	12	10	18	8	0	7	11	10	14	9
N	48	60	58	61	56	0	47	52	42	54	60
P	55	51	46	48	39	0	66	65	54	39	49
Q	39	25	33	28	52	0	30	32	32	26	35
R	44	47	47	44	65	0	66	45	49	47	42
S	64	65	66	56	55	0	58	66	76	61	54
T	43	53	49	65	71	0	63	46	53	52	58
V	39	35	41	29	51	0	36	25	39	35	40
W	13	15	15	14	18	0	10	13	10	17	14
Y	29	39	40	34	32	0	40	33	30	25	24

Score: 254.362532351798

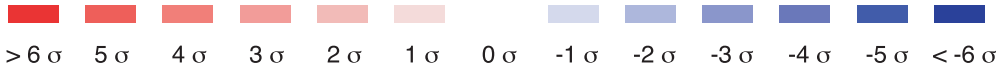
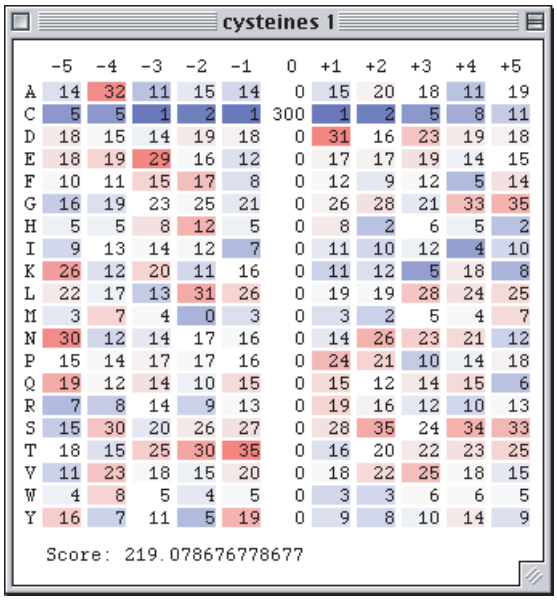
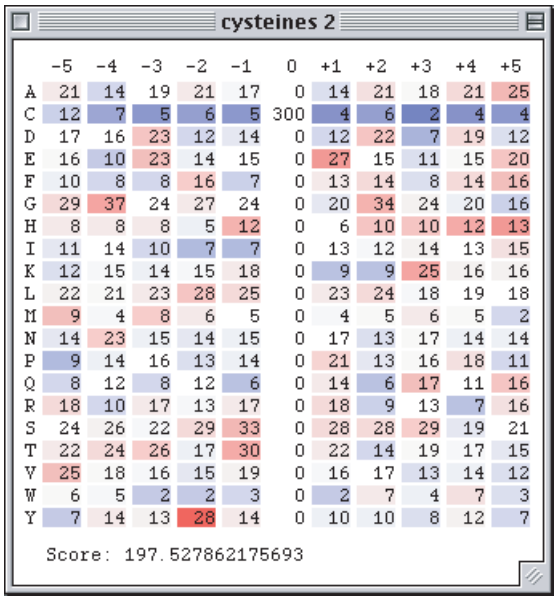


Figure 12. Comparison between disulfides from low-disulfide-density chains (< 0.05 disulfides/residue) and disulfides from high-disulfide-density chains (> 0.05 disulfides/residue).

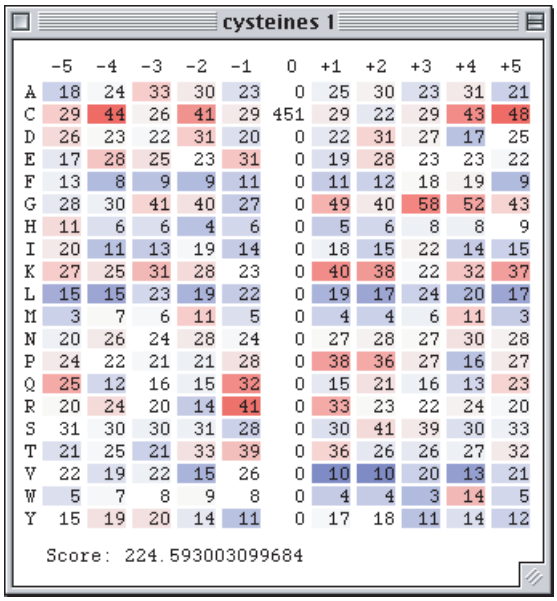
low density, N-terminal



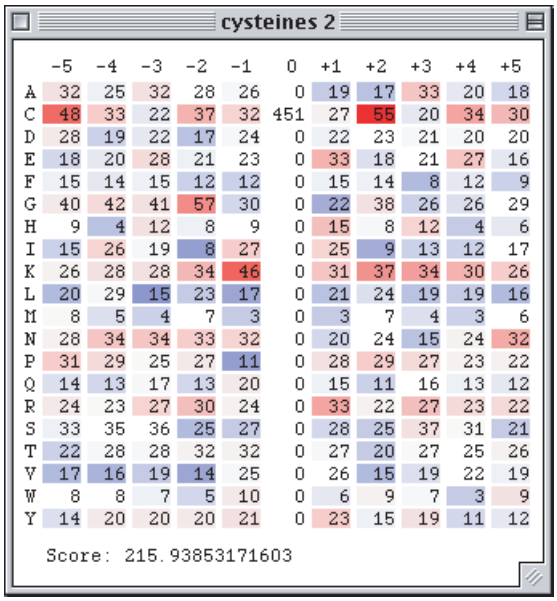
low density, C-terminal



high density, N-terminal



high density, C-terminal



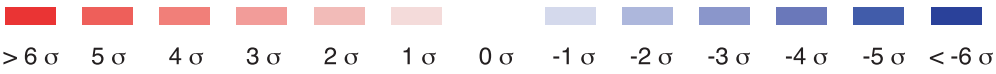
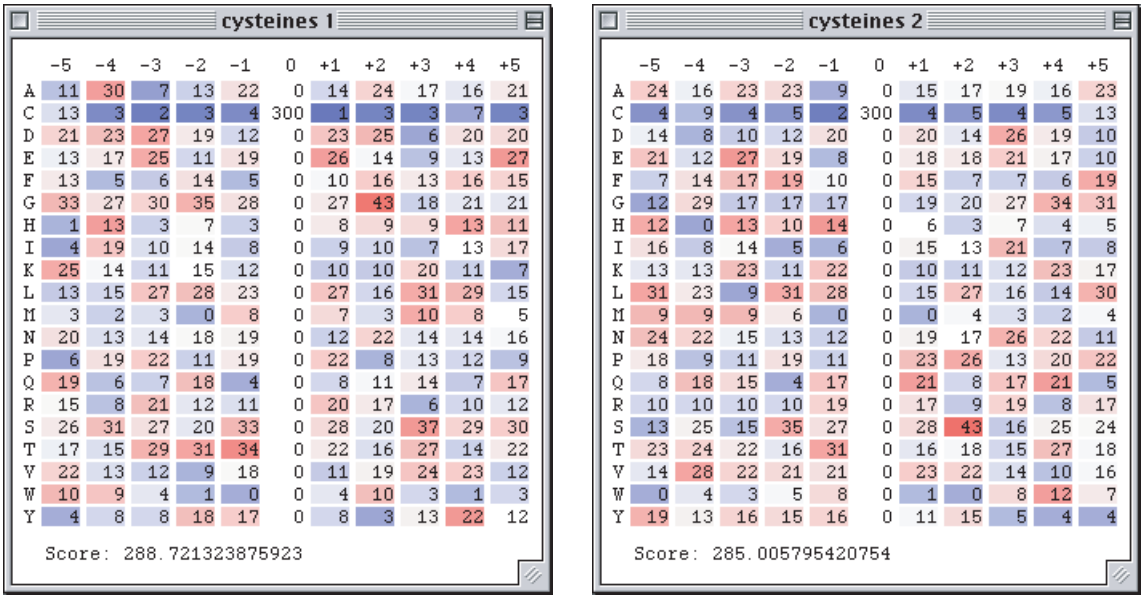
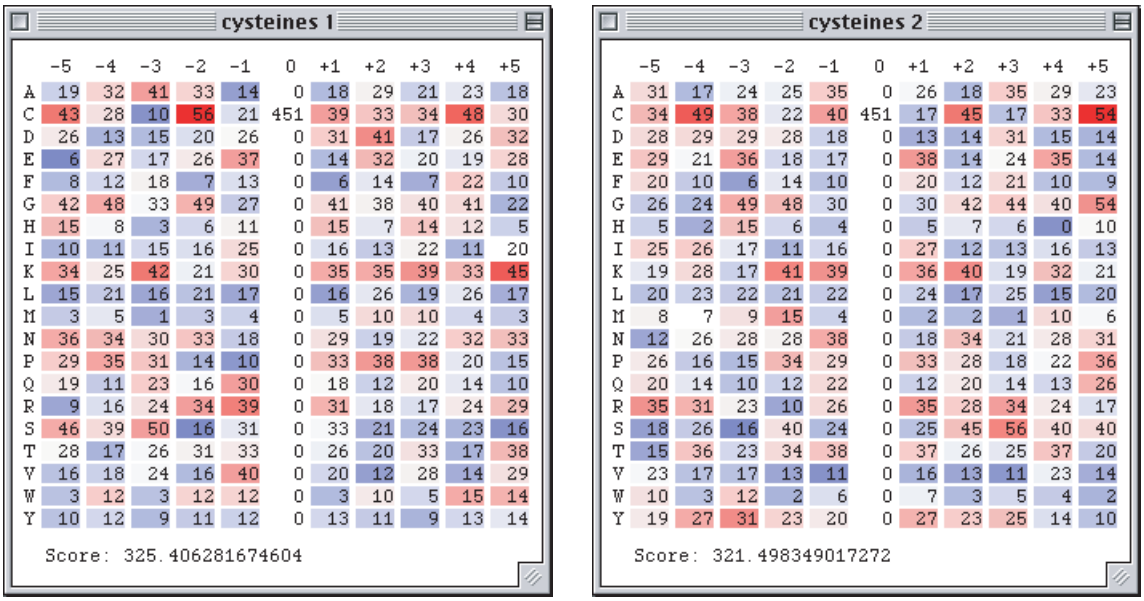


Figure 13. Comparison between N-terminal cysteines and C-terminal cysteines from low-disulfide-density and high-density chains.

low-disulfide-density



high-disulfide-density



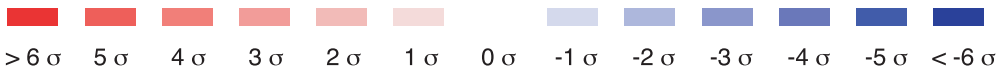


Figure 14. SA optimization of disulfides from low-disulfide-density chains and from high-disulfide-density chains.

DSBMax: Sequential Distance

Next, the number of residues separating the two cysteines involved in intrachain disulfide bonds (the "sequential distance") was calculated (Figures 15, 16). There was a strong bias towards sequential distances between four and 20 residues, as previously observed (Thornton, 1981). This is especially apparent when the distance distribution is compared to the distribution of all possible cysteine-cysteine distances in disulfide-bonded chains (Figure 17). From this distribution, a disulfide bond can be classified as sequentially-close (≤ 20 residues separation) or sequentially-distant (> 20 residues separation).

From the data set of eukaryotic intrachain disulfides, 449 sequentially-close disulfides and 414 sequentially-distant disulfides were compared with DSBMax (Figure 18). Overall, the sequentially-close disulfides had a higher abundance of Gly and Lys and a lower abundance of hydrophobic residues in their sequence environments. The sequentially-distant disulfides display no especially interesting features.

The cysteines in the sequentially-close and sequentially-distant disulfides were then subclassified according their orientation relative to the N- and C-termini (Figure 19). Interestingly, the N-terminal sequentially-close cysteines had an abundance of glycine upstream of the cysteine, while the C-terminal cysteines had an abundance of glycine downstream of the cysteine. This may suggest that a glycine is common between the cysteines of sequentially-close disulfides. The sequentially-distant disulfides again displayed no remarkable features.

SA optimization was run on the two groups of disulfides (Figure 20). Among the sequentially-close disulfides, the first subgroup has an abundance of Cys at -5; Gly at -3, -2, and +5; and Lys at +3. However, there was no significant correlation between any of these residues (data not shown). Among the sequentially-distant disulfides, His at -5, Lys at -3, and Pro at +1 were common in the first subgroup. However, the combination of the three or any two did not appear significantly often in the data set (data not shown).

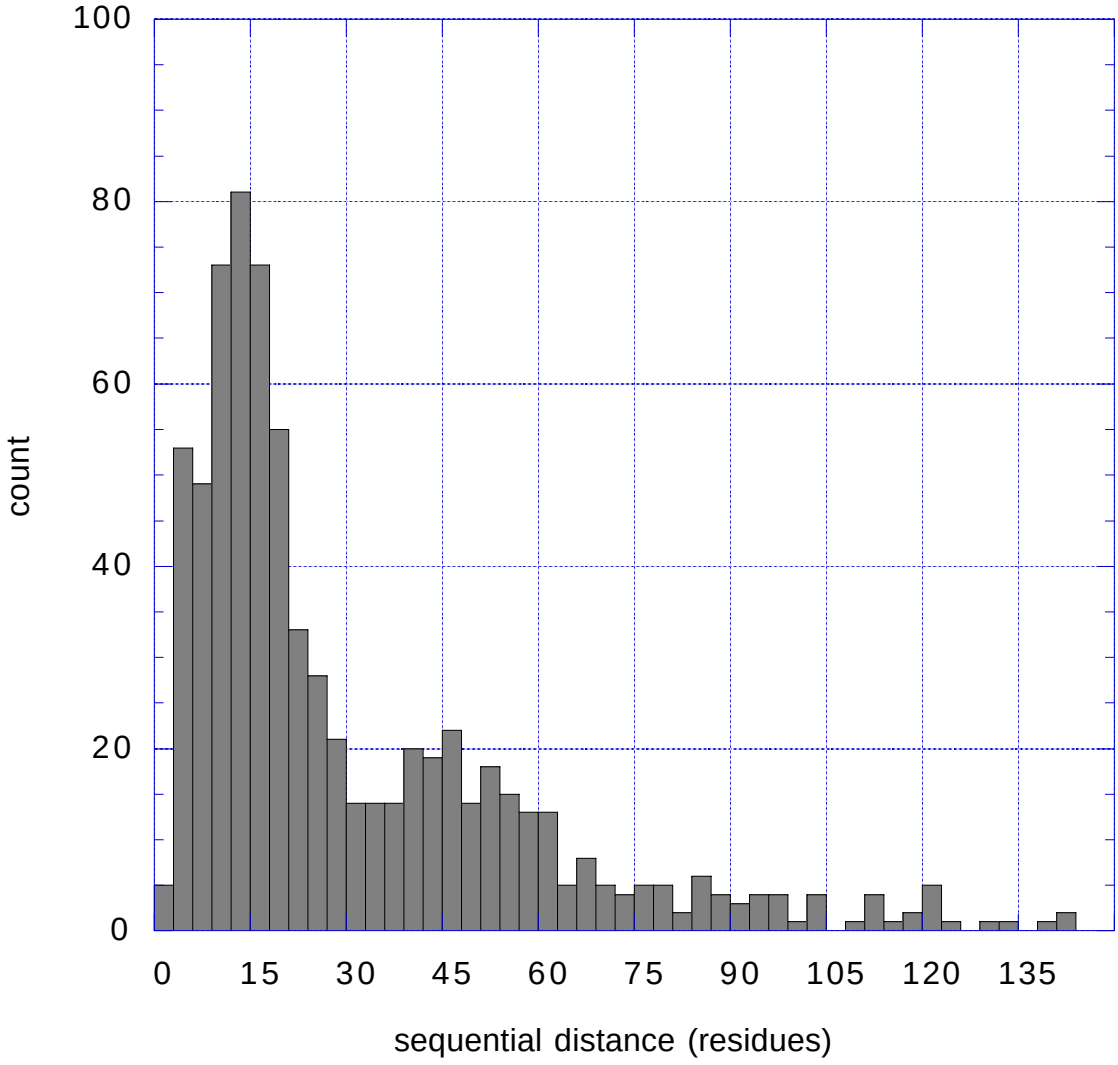


Figure 15. Histogram of the sequential distance between intrachain disulfides in eukaryotic proteins. Bin size=3. Not shown: 25 disulfides >150 residues apart.

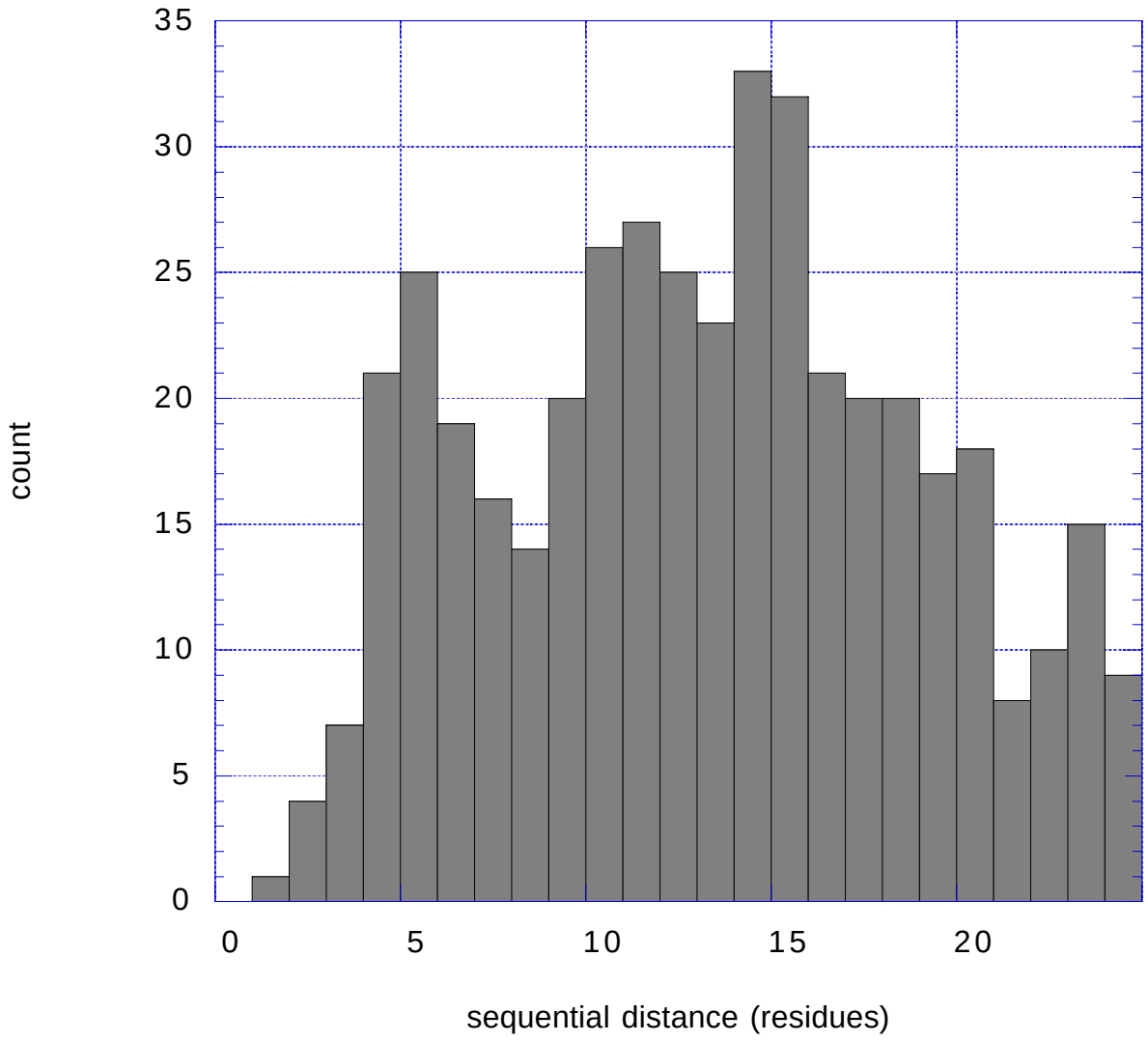


Figure 16. Close-up of histogram of the sequential distance between intrachain disulfides in eukaryotic proteins. Bin size=1.

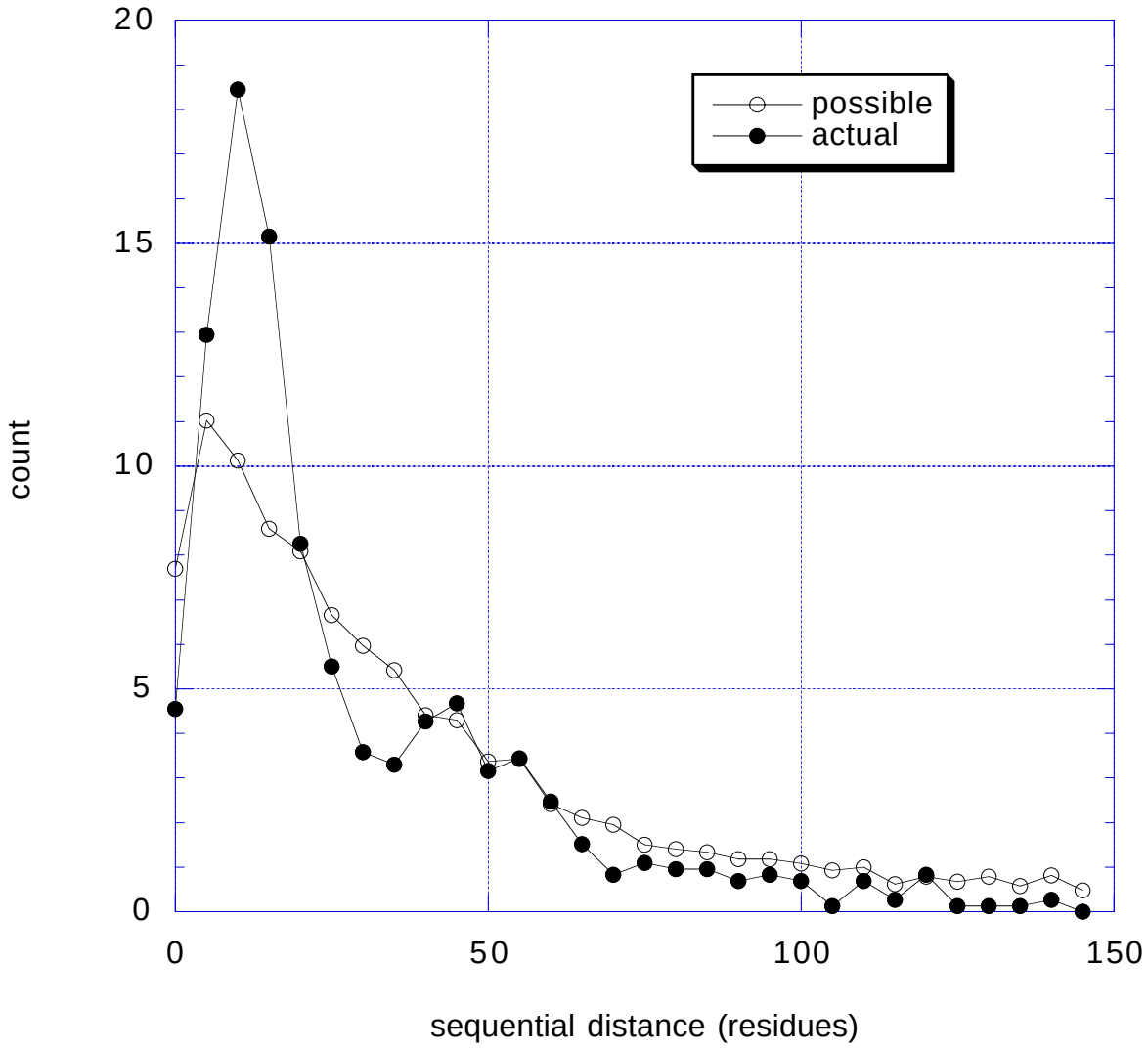


Figure 17. Comparison of the actual sequential distances between intrachain disulfides with the normalized possible distances, assuming equal likelihood of disulfide formation between any two cysteines. Bin size = 5.

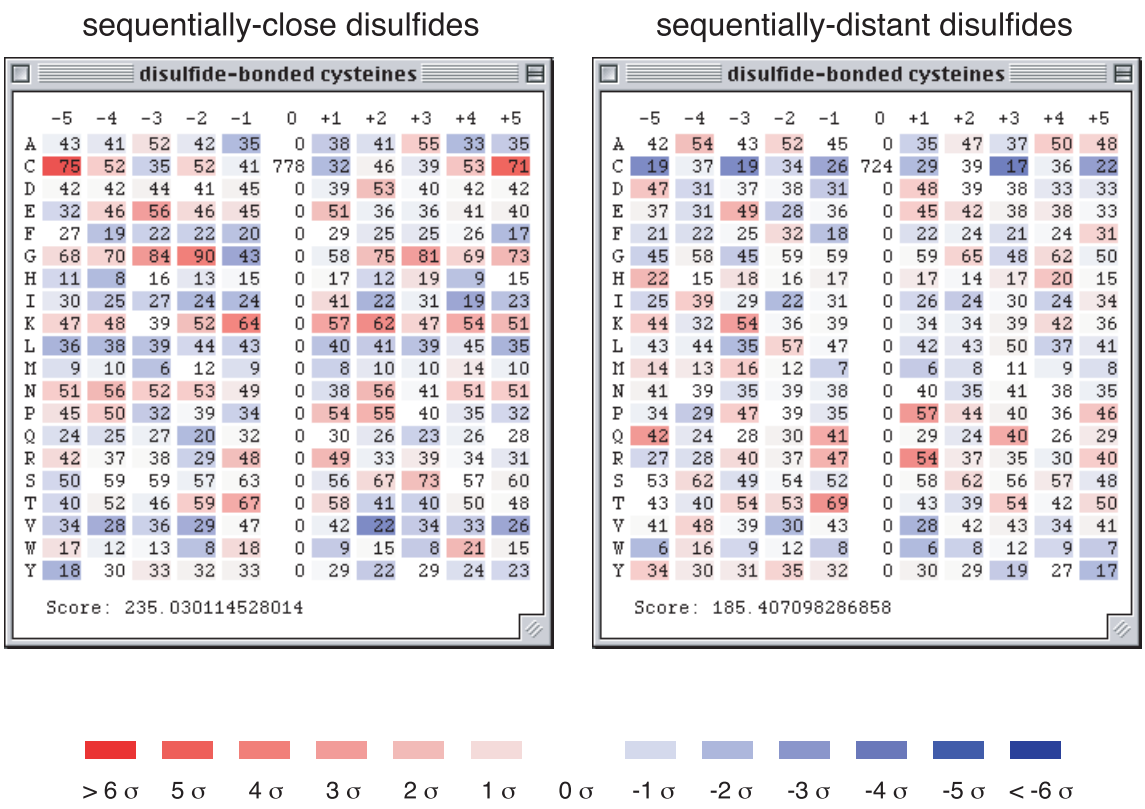


Figure 18. Comparison between sequentially-close (<= 20 residues) disulfides and sequentially-distant (> 20 residues) disulfides.

sequentially-close, N-terminal

cysteines 1

	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5
A	16	22	22	18	18	0	23	25	27	21	17
C	26	27	20	28	17	389	13	9	24	32	48
D	22	22	17	26	23	0	22	25	23	16	23
E	18	27	27	27	25	0	19	24	21	15	17
F	14	9	16	10	10	0	12	9	15	14	7
G	26	27	40	37	22	0	39	38	58	43	49
H	8	6	8	7	5	0	4	4	6	4	7
I	19	13	16	15	11	0	20	9	19	7	10
K	29	24	21	23	22	0	29	33	15	28	33
L	18	13	18	22	26	0	19	16	21	24	18
M	3	4	1	6	4	0	5	2	5	9	5
N	28	22	22	25	23	0	20	31	25	29	23
P	16	22	16	17	23	0	37	32	17	14	16
Q	15	13	13	9	22	0	15	19	11	16	12
R	19	18	16	10	31	0	25	19	13	19	10
S	23	31	27	29	24	0	27	44	37	28	35
T	16	23	19	32	36	0	31	24	19	28	27
V	18	14	20	19	21	0	17	9	18	13	14
W	7	9	8	4	9	0	2	3	3	14	4
Y	12	13	20	11	14	0	10	13	12	15	13

Score: 217.202684174065

sequentially-close, C-terminal

cysteines 2

	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5
A	27	19	30	24	17	0	15	16	28	12	18
C	49	25	15	24	24	389	19	37	15	21	23
D	20	20	27	15	22	0	17	28	17	26	19
E	14	19	29	19	20	0	32	12	15	26	23
F	13	10	6	12	10	0	17	16	10	12	10
G	42	43	44	53	21	0	19	37	23	26	24
H	3	2	8	6	10	0	13	8	13	5	8
I	11	12	11	9	13	0	21	13	12	12	13
K	18	24	18	29	42	0	28	29	32	26	18
L	18	25	21	22	17	0	21	25	18	21	17
M	6	6	5	6	5	0	3	8	5	5	5
N	23	34	30	28	26	0	18	25	16	22	28
P	29	28	16	22	11	0	17	23	23	21	16
Q	9	12	14	11	10	0	15	7	12	10	16
R	23	19	22	19	17	0	24	14	26	15	21
S	27	28	32	28	39	0	29	23	36	29	25
T	24	29	27	27	31	0	27	17	21	22	21
V	16	14	16	10	26	0	25	13	16	20	12
W	10	3	5	4	9	0	7	12	5	7	11
Y	6	17	13	21	19	0	19	9	17	9	10

Score: 218.719735324292

sequentially-distant, N-terminal

cysteines 1

	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5
A	16	34	22	27	19	0	17	25	14	21	23
C	8	22	7	15	13	362	17	15	10	19	11
D	22	16	19	24	15	0	31	22	27	20	20
E	17	20	27	12	18	0	17	21	21	22	20
F	9	10	8	16	9	0	11	12	15	10	16
G	18	22	24	28	26	0	36	30	21	42	29
H	8	5	6	9	6	0	9	4	8	9	4
I	10	11	11	16	10	0	9	16	15	11	15
K	24	13	30	16	17	0	22	17	12	22	12
L	19	19	18	28	22	0	19	20	31	20	24
M	3	10	9	5	4	0	2	4	6	6	5
N	22	16	16	20	17	0	21	23	25	22	17
P	23	14	22	21	21	0	25	25	20	16	29
Q	29	11	17	16	25	0	15	14	19	12	17
R	8	14	18	13	23	0	27	20	21	15	23
S	23	29	23	28	31	0	31	32	26	36	31
T	23	17	27	31	38	0	21	22	29	22	30
V	15	28	20	11	25	0	11	23	27	18	22
W	2	6	5	9	4	0	5	4	6	6	6
Y	19	13	11	8	16	0	16	13	9	13	8

Score: 187.974982903476

sequentially-distant, C-terminal

cysteines 2

	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5
A	26	20	21	25	26	0	18	22	23	29	25
C	11	15	12	19	13	362	12	24	7	17	11
D	25	15	18	14	16	0	17	17	11	13	13
E	20	11	22	16	18	0	28	21	17	16	13
F	12	12	17	16	9	0	11	12	6	14	15
G	27	36	21	31	33	0	23	35	27	20	21
H	14	10	12	7	11	0	8	10	9	11	11
I	15	28	18	6	21	0	17	8	15	13	19
K	20	19	24	20	22	0	12	17	27	20	24
L	24	25	17	29	25	0	23	23	19	17	17
M	11	3	7	7	3	0	4	4	5	3	3
N	19	23	19	19	21	0	19	12	16	16	18
P	11	15	25	18	14	0	32	19	20	20	17
Q	13	13	11	14	16	0	14	10	21	14	12
R	19	14	22	24	24	0	27	17	14	15	17
S	30	33	26	26	21	0	27	30	30	21	17
T	20	23	27	22	31	0	22	17	25	20	20
V	26	20	19	19	18	0	17	19	16	16	19
W	4	10	4	3	4	0	1	4	6	3	1
Y	15	17	20	27	16	0	14	16	10	14	9

Score: 176.603990773238

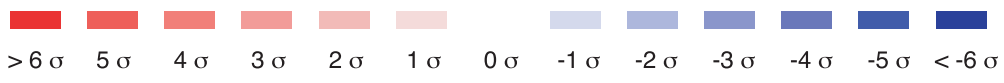


Figure 19. Comparison of the N-teminal and C-terminal cysteines from sequentially-close disulfides and from sequentially-distant disulfides.

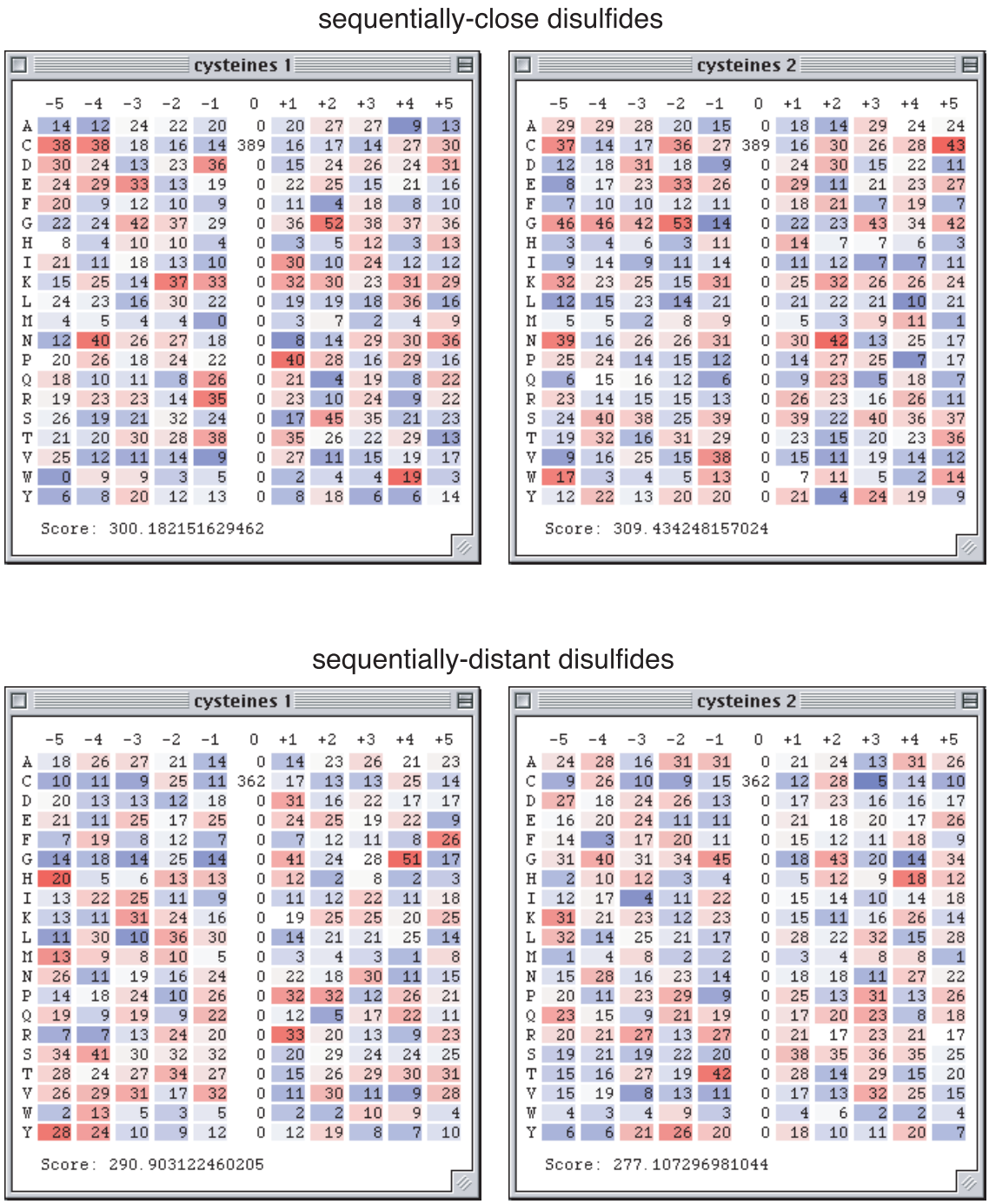


Figure 20. SA optimization of sequentially-close disulfides and sequentially-distant disulfides.

DSBMax: Intervening Cysteines

Next, the number of intervening cysteines between the cysteines in a disulfide bond was calculated (Figure 21). The histogram follows a decaying exponential, with a slight bias towards even numbers of intervening cysteines. The data set was divided into 273 disulfides with no intervening cysteines and 478 disulfides with at least one intervening cysteine and scored with DSBMax (Figure 22). The sequence environment of the disulfides without intervening cysteines was fairly similar to the sequence environment of disulfides from low-disulfide-density chains; this reflects that 75% of the disulfide set falls into the same category under classification by disulfide-density or by number of intervening cysteines.

The cysteines from the two sets of disulfides were subgrouped according to their orientation relative to the N- and C-termini (Figure 23). Again, the groupings yielded sequence environments similar to those when grouped by disulfide density. The DSBMax scores were lower overall for intervening-cysteine groupings than for disulfide-density groupings, suggesting that classification by disulfide density is the more natural one.

Finally, SA optimization was run on the two groups of disulfides (Figure 24). In the first subgroup of disulfides with no intervening cysteines, the more overrepresented residues are Asn at -4, Glu at -3, Tyr at -2, and Asp at +4. No combination of any two of these was represented more than five times. In the second subgroup of disulfides with no intervening cysteines, Pro was overrepresented at +1 and Ser was overrepresented at +2 and +3. The sequences CPS and CPXS were both found five times in the data set, which is not remarkable. In the first subgroup of disulfides with intervening cysteines, the overrepresented residues are Arg at +1, Pro at +2, Cys at +4, and Lys at +5. Interestingly, the sequence CXXXCK was found 13 times, when five would be expected by chance. In the second subgroup of disulfides with intervening cysteines, Cys at -4 and +2, Arg at -1, Thr at -1, and Lys at +1 are the most overrepresented. With the exception of CXXXCK, combinations of any two of these yielded no notable results. Even the CXXXCK sequence, while found more than twice as often as expected by chance, reflects less than 2% of the disulfide-bonded cysteines in this group.

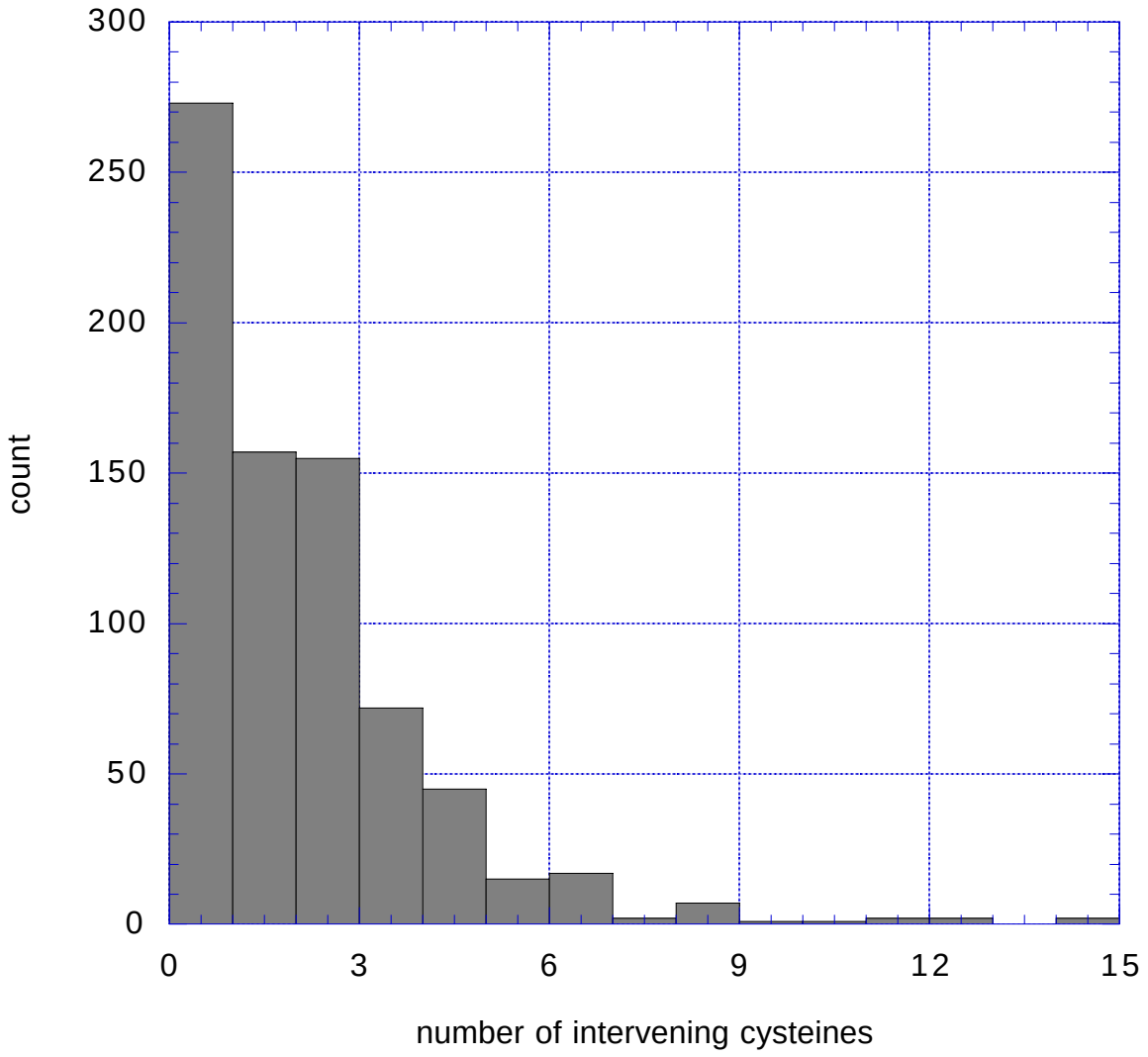


Figure 21. Histogram of the number of cysteines intervening between eukaryotic disulfide bonds.

disulfides without intervening cysteines

disulfide-bonded cysteines

	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5
A	34	37	33	41	27	0	21	29	35	17	31
C	30	18	13	20	8	546	11	12	9	15	23
D	27	37	35	32	33	0	33	40	24	41	32
E	25	23	46	28	22	0	38	31	31	30	33
F	23	15	20	26	16	0	25	19	18	22	20
G	42	52	46	50	36	0	35	56	49	45	39
H	9	6	13	13	15	0	13	9	13	9	17
I	24	23	21	18	13	0	23	15	24	13	24
K	32	30	26	26	47	0	27	31	27	38	26
L	35	32	35	46	38	0	36	34	31	40	38
M	8	13	10	9	8	0	6	7	10	8	9
N	37	41	28	31	33	0	31	38	30	33	27
P	21	28	26	27	22	0	46	25	28	28	26
Q	22	20	23	14	17	0	21	19	15	17	19
R	24	20	23	16	29	0	32	24	18	18	24
S	40	44	43	44	55	0	40	53	65	46	52
T	38	38	38	38	52	0	45	38	36	41	40
V	29	33	34	23	34	0	33	28	36	27	19
W	9	9	4	4	10	0	5	9	7	10	8
Y	23	17	23	35	28	0	19	17	20	24	14

Score: 190.594721605924

disulfides with intervening cysteines

disulfide-bonded cysteines

	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5
A	51	58	62	53	53	0	52	59	57	66	52
C	64	71	41	66	59	956	50	73	47	74	70
D	62	36	46	47	43	0	54	52	54	34	43
E	44	54	59	46	59	0	58	47	43	49	40
F	25	26	27	28	22	0	26	30	28	28	28
G	71	76	83	99	66	0	82	84	80	86	84
H	24	17	21	16	17	0	21	17	23	20	13
I	31	41	35	28	42	0	44	31	37	30	33
K	59	50	67	62	56	0	64	65	59	58	61
L	44	50	39	55	52	0	46	50	58	42	38
M	15	10	12	15	8	0	8	11	11	15	9
N	55	54	59	61	54	0	47	53	52	56	59
P	58	51	53	51	47	0	65	74	52	43	52
Q	44	29	32	36	56	0	38	31	48	35	38
R	45	45	55	50	66	0	71	46	56	46	47
S	63	77	65	67	60	0	74	76	64	68	56
T	45	54	62	74	84	0	56	42	58	51	58
V	46	43	41	36	56	0	37	36	41	40	48
W	14	19	18	16	16	0	10	14	13	20	14
Y	29	43	41	32	37	0	40	34	28	27	26

Score: 224.735625968918

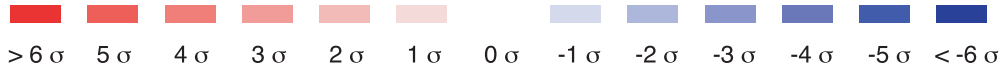


Figure 22. Comparison between disulfides with no intervening cysteines and cysteines with at least one intervening cysteine.

no intervening cysteines, N-terminal

cysteines 1

	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5
A	13	29	15	18	12	0	10	12	12	8	16
C	10	11	9	19	8	273	0	1	4	7	20
D	15	18	11	16	16	0	23	18	20	18	18
E	13	13	25	15	11	0	14	21	18	13	11
F	13	10	16	12	9	0	11	10	11	9	10
G	16	13	22	20	18	0	20	25	30	25	27
H	4	2	7	7	2	0	7	2	5	2	4
I	13	12	13	13	7	0	10	4	12	5	10
K	20	17	15	11	18	0	14	17	6	23	13
L	18	14	15	25	21	0	17	16	20	24	22
M	2	7	3	3	2	0	4	2	5	4	6
N	22	12	14	14	15	0	13	26	17	19	9
P	14	11	11	13	10	0	32	13	12	11	15
Q	15	13	13	7	13	0	12	12	7	12	6
R	9	12	6	6	18	0	17	17	6	11	6
S	16	22	18	22	21	0	20	31	34	28	33
T	15	14	16	21	28	0	24	19	18	22	24
V	12	17	21	14	19	0	16	16	22	14	11
W	5	7	3	3	8	0	2	2	3	5	3
Y	15	9	14	9	14	0	7	8	11	13	9

Score: 196.637039480584

no intervening cysteines, C-terminal

cysteines 2

	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5
A	21	8	18	23	15	0	11	17	23	9	15
C	20	7	4	1	0	273	11	11	5	8	3
D	12	19	24	16	17	0	10	22	4	23	14
E	12	10	21	13	11	0	24	10	13	17	22
F	10	5	4	14	7	0	14	9	7	13	10
G	26	39	24	30	18	0	15	31	19	20	12
H	5	4	6	6	13	0	6	7	8	7	13
I	11	11	8	5	6	0	13	11	12	8	14
K	12	13	11	15	29	0	13	14	21	15	13
L	17	18	20	21	17	0	19	18	11	16	16
M	6	6	7	6	6	0	2	5	5	4	3
N	15	29	14	17	18	0	18	12	13	14	18
P	7	17	15	14	12	0	14	12	16	17	11
Q	7	7	10	7	4	0	9	7	8	5	13
R	15	8	17	10	11	0	15	7	12	7	18
S	24	22	25	22	34	0	20	22	31	18	19
T	23	24	22	17	24	0	21	19	18	19	16
V	17	16	13	9	15	0	17	12	14	13	8
W	4	2	1	1	2	0	3	7	4	5	5
Y	8	8	9	26	14	0	12	9	9	11	5

Score: 202.705356835787

with intervening cysteines, N-terminal

cysteines 1

	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5
A	19	27	29	27	25	0	30	38	29	34	24
C	24	38	18	24	22	478	30	23	30	44	39
D	29	20	25	34	22	0	30	29	30	18	25
E	22	34	29	24	32	0	22	24	24	24	26
F	10	9	8	14	10	0	12	11	19	15	13
G	28	36	42	45	30	0	55	43	49	60	51
H	12	9	7	9	9	0	6	6	9	11	7
I	16	12	14	18	14	0	19	21	22	13	15
K	33	20	36	28	21	0	37	33	21	27	32
L	19	18	21	25	27	0	21	20	32	20	20
M	4	7	7	8	6	0	3	4	6	11	4
N	28	26	24	31	25	0	28	28	33	32	31
P	25	25	27	25	34	0	30	44	25	19	30
Q	29	11	17	18	34	0	18	21	23	16	23
R	18	20	28	17	36	0	35	22	28	23	27
S	30	38	32	35	34	0	38	45	29	36	33
T	24	26	30	42	46	0	28	27	30	28	33
V	21	25	19	16	27	0	12	16	23	17	25
W	4	8	10	10	5	0	5	5	6	15	7
Y	16	17	17	10	16	0	19	18	10	15	12

Score: 205.647502006461

with intervening cysteines, C-terminal

cysteines 2

	-5	-4	-3	-2	-1	0	+1	+2	+3	+4	+5
A	32	31	33	26	28	0	22	21	28	32	28
C	40	33	23	42	37	478	20	50	17	30	31
D	33	16	21	13	21	0	24	23	24	16	18
E	22	20	30	22	27	0	36	23	19	25	14
F	15	17	19	14	12	0	14	19	9	13	15
G	43	40	41	54	36	0	27	41	31	26	33
H	12	8	14	7	8	0	15	11	14	9	6
I	15	29	21	10	28	0	25	10	15	17	18
K	26	30	31	34	35	0	27	32	38	31	29
L	25	32	18	30	25	0	25	30	26	22	18
M	11	3	5	7	2	0	5	7	5	4	5
N	27	28	35	30	29	0	19	25	19	24	28
P	33	26	26	26	13	0	35	30	27	24	22
Q	15	18	15	18	22	0	20	10	25	19	15
R	27	25	27	33	30	0	36	24	28	23	20
S	33	39	33	32	26	0	36	31	35	32	23
T	21	28	32	32	38	0	28	15	28	23	25
V	25	18	22	20	29	0	25	20	18	23	23
W	10	11	8	6	11	0	5	9	7	5	7
Y	13	26	24	22	21	0	21	16	18	12	14

Score: 195.704139969342

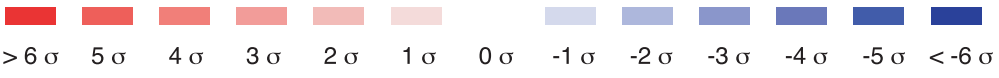
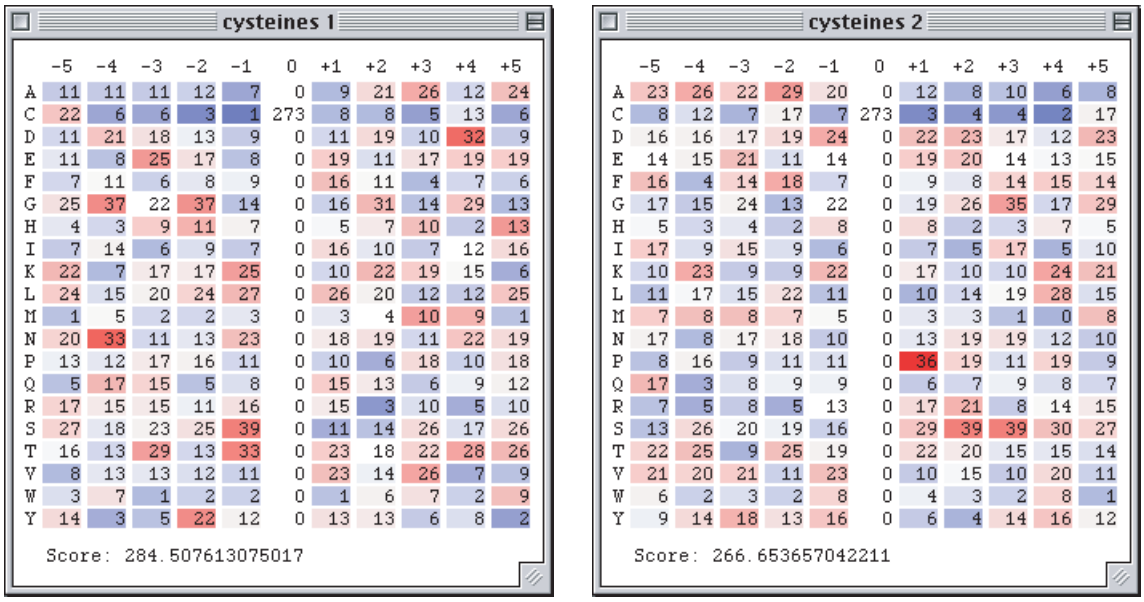
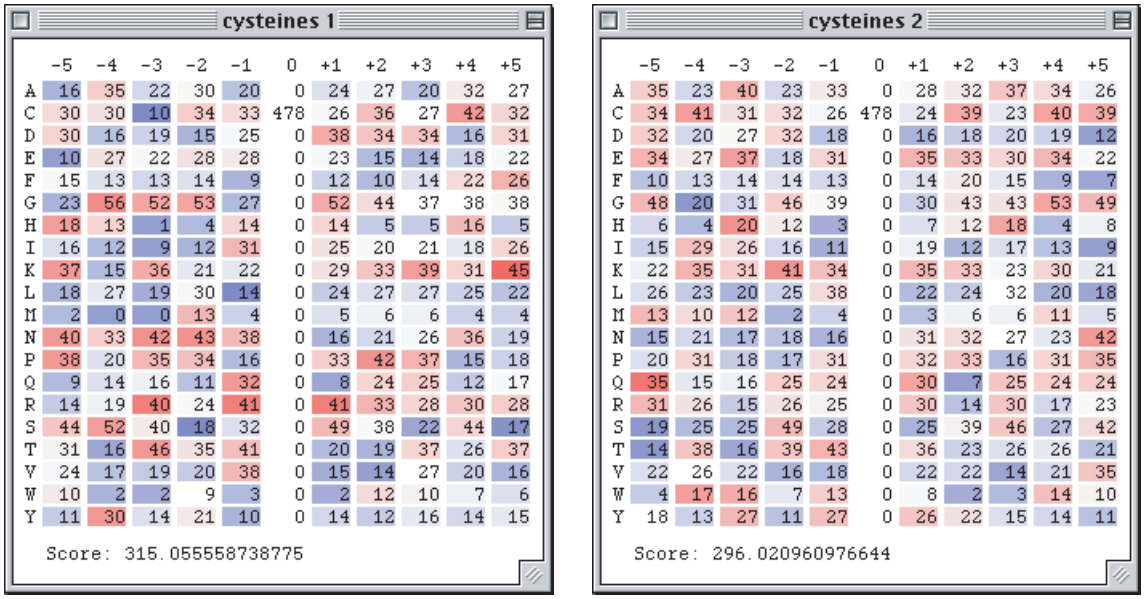


Figure 23. Comparison of the N-teminal and C-terminal cysteines from disulfides without intervening cysteines and from disulfides with intervening cysteines.

disulfides with no intervening cysteines



disulfides with intervening cysteines



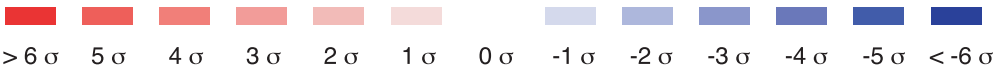


Figure 24. SA optimization of disulfides without intervening cysteines and of disulfides with intervening cysteines.

Predictors of Disulfide Connectivity

Several programs were written to predict the disulfide connectivity of an uncharacterized protein. Their peformance was analyzed by predicting the connectivity of 171 proteins of known structure. First, the expected results of a random predictor are shown to establish a baseline for evaluating the predictors (Table 8). Overall, a random predictor should pick the correct connectivity in 12% of the proteins in the data set, and it should pick the correct partner for 19% of the cysteines in the data set.

number of cysteines	Q_p (random predictor)	Q_c (random predictor)
3	0.333 (1/3)	0.333 (1/3)
4	0.333 (1/3)	0.333 (1/3)
5	0.067 (1/15)	0.200 (1/5)
6	0.067 (1/15)	0.200 (1/5)
7	0.010 (1/105)	0.143 (1/7)
8	0.010 (1/105)	0.143 (1/7)
10	0.001 (1/945)	0.111 (1/9)
overall	0.121 (20.8/171)	0.191 (201.3/1056)

Table 8. Prediction quality scores expected from a random predictor.

The results of the predictor PredNResBtwn are shown in Table 9. Overall, this predictor performed about twice as well as a random predictor, with a Q_p of 25% and a Q_c of 35%. The most notable improvement is in the prediction of proteins containing six or eight cysteines.

number of cysteines	Q_p (NResBtwn)	Q_c (NResBtwn)
3	0.333 (4/12)	0.333 (12/36)
4	0.472 (17/36)	0.472 (68/144)
5	0.222 (2/9)	0.356 (16/45)
6	0.276 (16/58)	0.402 (140/348)
7	0.000 (0/5)	0.229 (8/35)
8	0.129 (4/31)	0.347 (86/248)
10	0.000 (0/20)	0.190 (38/200)
overall	0.251 (43/171)	0.348 (368/1056)

Table 9. Prediction quality scores from PredNResBtwn.

The prediction quality of PredSeqClassDistFair was unimpressive and not significantly different from the quality of a random predictor (Table 10). Initial results (not shown) showed that PredSeqClassDist (the same program but without the jacknife procedure) performed about twice as well as the random predictor. But with 800 parameters, PredSeqClassDist was essentially memorizing the sequences in the data set, which is a poor indicator of future performance on novel proteins.

number of cysteines	Q_p (SeqClassDistFair)	Q_c (SeqClassDistFair)
3	0.417 (5/12)	0.417 (15/36)
4	0.389 (14/36)	0.389 (56/144)
5	0.000 (0/9)	0.178 (8/45)
6	0.069 (4/58)	0.155 (54/348)
7	0.000 (0/5)	0.343 (12/35)
8	0.000 (0/31)	0.097 (24/248)
10	0.000 (0/20)	0.180 (36/200)
overall	0.135 (23/171)	0.194 (205/1056)

Table 10. Prediction quality scores from PredSeqClassDistFair.

When evaluated over the entire data set, the predictor PredEntropy performed slightly worse than the random predictor (Table 11). On the other hand, if the data set is reduced to include only chains of length less than 81 residues, the accuracy is significantly better than the random predictor (Table 12). In this case, only 40% of the chains in the data set are evaluated, but the results are better than the random predictor.

number of cysteines	Q_p (Entropy, all)	Q_c (Entropy, all)
3	0.250 (3/12)	0.250 (9/36)
4	0.278 (10/36)	0.278 (40/144)
5	0.000 (0/9)	0.156 (7/45)
6	0.0512 (3/58)	0.230 (80/348)
7	0.000 (0/5)	0.114 (4/35)
8	0.032 (1/31)	0.169 (42/248)
10	0.000 (0/20)	0.060 (12/200)
overall	0.099 (17/171)	0.184 (194/1056)

Table 11. Prediction quality scores from PredEntropy over the entire data set.

number of cysteines	Q_p (Entropy, length ≤ 80)	Q_c (Entropy, length ≤ 80)
3	n/a	n/a
4	1.000 (6/6)	1.000 (6/6)
5	n/a	n/a
6	0.079 (3/38)	0.316 (72/228)
7	0.000 (0/1)	0.286 (2/7)
8	0.053 (1/19)	0.224 (34/152)
10	0.000 (0/8)	0.050 (4/80)
overall	0.139 (10/72)	0.277 (136/491)

Table 12. Prediction quality scores from PredEntropy over the subset of chains of length less than 80.

The predictor PredDiffusion showed an uninteresting performance over the entire data set (Table 13). However, its accuracy was improved significantly over the subset of the data set with length greater than 164 residues (Table 14). In the reduced data set, the Q_p was 40% and the Q_c was 47%.

number of cysteines	Q_p (Diffusion, all)	Q_c (Diffusion, all)
3	0.417 (5/12)	0.417 (15/36)
4	0.389 (14/36)	0.389 (56/144)
5	0.444 (4/9)	0.533 (24/45)
6	0.052 (3/58)	0.161 (56/348)
7	0.000 (0/5)	0.229 (8/35)
8	0.000 (0/31)	0.089 (22/248)
10	0.000 (0/20)	0.110 (22/200)
overall	0.152 (26/171)	0.192 (203/1056)

Table 13. Prediction quality scores from PredDiffusion over the entire data set.

number of cysteines	Q_p (Diffusion, length ≥165)	Q_c (Diffusion, length ≥165)
3	0.600 (3/5)	0.600 (9/15)
4	0.647 (11/17)	0.647 (44/68)
5	0.444 (4/9)	0.533 (24/45)
6	0.333 (3/9)	0.519 (28/54)
7	0.000 (0/4)	0.286 (8/28)
8	0.000 (0/5)	0.350 (14/40)
10	0.000 (0/4)	0.250 (10/40)
overall	0.396 (21/53)	0.472 (137/290)

Table 14. Prediction quality scores from PredDiffusion over the subset of chains of length greater than 164.

The predictor PredEntDiff performed fairly well with a cutoff value of 100 residues, with an overall Q_p of 21% and a Q_c of 29% (Table 15). With this cutoff, about half of the chains were predicted with PredEntropy and about half were predicted with PredDiffusion. Additionally, the predictor was run on a reduced data set, such that chains below length 81 were predicted using PredEntropy, chains above length 164 were predicted with PredDiffusion, and chains from 81 to 179 residues were not predicted (Table 16). Using the reduced data set increased the prediction accuracy to a Q_p of 25% and a Q_c of 35%.

number of cysteines	Q_p (EntDiff)	Q_c (EntDiff)
3	0.417 (5/12)	0.417 (15/36)
4	0.556 (20/36)	0.556 (80/144)
5	0.444 (4/9)	0.533 (24/45)
6	0.103 (6/58)	0.328 (114/348)
7	0.000 (0/5)	0.286 (10/35)
8	0.032 (1/31)	0.194 (48/248)
10	0.000 (0/20)	0.090 (18/200)
overall	0.211 (36/171)	0.293 (309/1056)

Table 15. Prediction quality scores from PredEntDiff.

number of cysteines	Q_p (EntDiff, length ≤80 or length ≥165)	Q_c (EntDiff, length ≤80 or length ≥165)
3	0.600 (3/5)	0.600 (9/15)
4	0.739 (17/23)	0.739 (68/92)
5	0.444 (4/9)	0.533 (24/45)
6	0.128 (6/47)	0.355 (100/282)
7	0.000 (0/5)	0.286 (10/35)
8	0.042 (1/24)	0.250 (48/92)
10	0.000 (0/12)	0.117 (14/120)
overall	0.248 (31/125)	0.350 (273/781)

Table 16. Prediction quality scores from PredEntDiff over the subset of chains of length less than 81 or greater than 164.

PredAdjacent was another unimpressive predictor with a performance roughly equal to the random predictor (Table 17).

number of cysteines	Q_p (Adjacent)	Q_c (Adjacent)
3	0.250 (3/12)	0.250 (9/36)
4	0.361 (13/36)	0.361 (52/144)
5	0.111 (1/9)	0.200 (9/45)
6	0.086 (5/58)	0.184 (64/348)
7	0.000 (0/5)	0.286 (10/35)
8	0.000 (0/31)	0.113 (28/248)
10	0.000 (0/20)	0.160 (32/200)
overall	0.129 (22/171)	0.193 (204/1056)

Table 17. Prediction quality scores from PredAdjacent.

CONCLUSIONS

DSBMax

A set of nonhomologous eukaryotic proteins of known crystal structure was searched for patterns in the sequence environment flanking the cysteines involved in disulfide bonds. A program, DSBMax, was written to graphically display the bulk sequence environments of a given subset of disulfide-bonded cysteines. The program also implemented a simulated annealing algorithm to sort disulfide-bonded cysteines into two groups that had the most similar sequence environment.

DSBMax was used to study the set of all disulfide bonds, as well as three subsets chosen by easily measurable parameters. The subsets were based on the disulfide-bond-density of the polypeptide chain, the sequential distance between the disulfide-bonded cysteines, or the presence of absence of an intervening cysteine in the sequence separating the two disulfide-bonded cysteines. Each set and subset was also divided into N- and C-terminal cysteines.

The DSBMax output of each subset was scanned for overrepresented residues that might reflect a sequence bias in the formation of disulfide bonds belonging to that subset. A few sequence patterns were found that were present near disulfide-bonded cysteines more often than expected by chance, but none were present nearly often enough to predict the subgroup to which a cysteine might belong.

An extension of this approach to disulfide bond classification might prove fruitful. After the average sequence environment of a subset of disulfide-bonded cysteines is calculated, each cysteine in the subset could be compared to the average. Cysteines that are relatively close to the average would be kept in the group, while cysteines that are farther from the average would be moved to a different group. It would be desirable to have several groups to which a poorly-fitting cysteine could be moved. This approach might be able to divide disulfide-bonded cysteines into several groups which have a more conserved sequence environment than have been produced in this study.

Prediction Programs

As a test of the significance of the trends found by DSBMax, a series of sequence-based prediction programs were written. In addition to a predictor based on the flanking

sequence matrices as studied by DSBMax, other predictors were also developed. Originally designed as controls, the non-sequence-based predictors ultimately performed better than the sequence-based one.

Two predictors that performed relatively well over then entire data set are PredNResBtwn and PredEntDiff. Overall, PredNResBtwn was twice as likely as a random predictor to predict the correct connectivity. However, with a 25% accuracy its predictions are still far from reliable. The success of this predictor is based on a slight bias in sequential distance between disulfide-bonded cysteines, where disulfides tend to form between nearby cysteines preferentially. This bias may be the result of nearby cysteines being more likely to come in contact with each other during the folding process, or it might be a reflection of the tendency of proteins to fold into discreet domains.

PredEntDiff was 1.7 times as likely as a random predictor to predict the correct connectivity, when evaluated over the entire data set. This success of this predictor is particularly interesting because it is almost entirely derived from a theoretical model, with only one free parameter (the cutoff length for choosing between entropy/diffusion models) learned by analyzing real proteins.

Two of the predictors, PredEntropy and PredDiffusion, performed better over a limited subset of the data set. PredEntropy performed fairly well on short chains of length less than 80. PredDiffusion performed remarkably well on longer chains of length greater than 165.

All the predictors studied here utilize only a small portion of the available sequence data, and it is likely that a more complicated predictor would offer improved prediction quality. For example, with the increased availability of computational power it might soon be possible to analyze the structural feasibility of each possible disulfide connectivity.

.

APPENDIX

References

Aarts, A., and Korst, J. (1989). Simulated Annealing and Boltzmann Machines: Wiley.

Abagyan, R. A., and Batalov, S. (1997). Do aligned sequences share the same fold? J Mol Biol *273*, 355-68.

Anfinsen, C. B. (1973). Principles that govern the folding of protein chains. Science *181*, 223-30.

Benham, C. J., and Jafri, M. S. (1993). Disulfide bonding patterns and protein topologies. Protein Sci *2*, 41-54.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., and Bourne, P. E. (2000). The Protein Data Bank. Nucleic Acids Res *28*, 235-42.

Braakman, I., Hoover-Litty, H., Wagner, K. R., and Helenius, A. (1991). Folding of influenza hemagglutinin in the endoplasmic reticulum. J Cell Biol *114*, 401-11.

Creighton, T. E. (1997). Protein folding coupled to disulphide bond formation. Biol Chem *378*, 731-44.

Debarbieux, L., and Beckwith, J. (1999). Electron avenue: pathways of disulfide bond formation and isomerization. Cell *99*, 117-9.

Fariselli, P., Riccobelli, P., and Casadio, R. (1999). Role of evolutionary information in predicting the disulfide-bonding state of cysteine in proteins. Proteins *36*, 340-6.

Fariselli, P., and Casadio, R. (2001). Prediction of disulfide connectivity in proteins. Bioinformatics *17*, 957-64.

Ferrari, D. M., and Soling, H. D. (1999). The protein disulphide-isomerase family: unravelling a string of folds. Biochem J *339*, 1-10.

Fiser, A., Cserzo, M., Tudos, E., and Simon, I. (1992). Different sequence environments of cysteines and half cystines in proteins. Application to predict disulfide forming residues. FEBS Lett *302*, 117-20.

Frand, A. R., Cuozzo, J. W., and Kaiser, C. A. (2000). Pathways for protein disulphide bond formation. Trends Cell Biol *10*, 203-10.

Frand, A. R., and Kaiser, C. A. (1999). Ero1p oxidizes protein disulfide isomerase in a pathway for disulfide bond formation in the endoplasmic reticulum. Mol Cell *4*, 469-77.

Goldberger, R. F., Epstein, C. J., and Anfinsen, C. B. (1963). Acceleration of reactivation of reduced bovine pancreatic ribonuclease by a microsomal system from rat liver. J Biol Chem *238*, 628-35.

Harrison, P. M., and Sternberg, M. J. (1994). Analysis and classification of disulphide connectivity in proteins. The entropic effect of cross-linkage. J Mol Biol *244*, 448-63.

Hobohm, U., and Sander, C. (1994). Enlarged representative set of protein structures. Protein Sci *3*, 522-4.

Johnson, R. E., Adams, P., and Rupley, J. A. (1978). Thermodynamics of protein cross-links. Biochemistry *17*, 1479-84.

Kaiser, C. A., Gimeno, R. E., and Shaywitz, D. A. (1997). Protein Secretion, Membrane Biogenesis, and Endocytosis. In Yeast III: Cold Spring Harbor Laboratory Press).

Kemmink, J., Darby, N. J., Dijkstra, K., Nilges, M., and Creighton, T. E. (1997). The folding catalyst protein disulfide isomerase is constructed of active and inactive thioredoxin modules. Curr Biol *7*, 239-45.

Kemmink, J., Darby, N. J., Dijkstra, K., Nilges, M., and Creighton, T. E. (1996). Structure determination of the N-terminal thioredoxin-like domain of protein disulfide isomerase using multidimensional heteronuclear 13C/15N NMR spectroscopy. Biochemistry *35*, 7684-91.

Kobayashi, T., Kishigami, S., Sone, M., Inokuchi, H., Mogi, T., and Ito, K. (1997). Respiratory chain is required to maintain oxidized states of the DsbA- DsbB disulfide bond formation system in aerobically growing Escherichia coli cells. Proc Natl Acad Sci U S A *94*, 11857-62.

Lambert, N., and Freedman, R. B. (1983). Kinetics and specificity of homogeneous protein disulphide-isomerase in protein disulphide isomerization and in thiol-protein-disulphide oxidoreduction. Biochem J *213*, 235-43.

Mirny, L. A., and Shakhnovich, E. I. (1996). How to derive a protein folding potential? A new approach to an old problem. J Mol Biol *264*, 1164-79.

Morjana, N. A., and Gilbert, H. F. (1991). Effect of protein and peptide inhibitors on the activity of protein disulfide isomerase. Biochemistry *30*, 4985-90.

Muskal, S. M., Holbrook, S. R., and Kim, S. H. (1990). Prediction of the disulfide-bonding state of cysteine in proteins. Protein Eng *3*, 667-72.

Norgaard, P., Westphal, V. V., Tachibana, C., Alsoe, L., Holst, B., and Winther, J. R. (2001). Functional Differences in Yeast Protein Disulfide Isomerases. J Cell Biol *152*, 553-562.

Petersen, M. T., Jonson, P. H., and Petersen, S. B. (1999). Amino acid neighbours and detailed conformational analysis of cysteines in proteins. Protein Eng *12*, 535-48.

Raina, S., and Missiakas, D. (1997). Making and breaking disulfide bonds. Annu Rev Microbiol *51*, 179-202.

Rietsch, A., Belin, D., Martin, N., and Beckwith, J. (1996). An in vivo pathway for disulfide bond isomerization in Escherichia coli. Proc Natl Acad Sci U S A *93*, 13048-53.

Sander, C., and Schneider, R. (1991). Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins *9*, 56-68.

Thornton, J. M. (1981). Disulphide bridges in globular proteins. J Mol Biol *151*, 261-87.

Wedemeyer, W. J., Welker, E., Narayan, M., and Scheraga, H. A. (2000). Disulfide bonds and protein folding. Biochemistry *39*, 4207-16.

Wei, J., and Hendershot, L. M. (1996). Protein folding and assembly in the endoplasmic reticulum. Exs *77*, 41-55.

Weissman, J. S., and Kim, P. S. (1991). Reexamination of the folding of BPTI: predominance of native intermediates. Science *253*, 1386-93.

Westphal, V., Spetzler, J. C., Meldal, M., Christensen, U., and Winther, J. R. (1998). Kinetic analysis of the mechanism and specificity of protein-disulfide isomerase using fluorescence-quenched peptides. J Biol Chem *273*, 24992-9.

Wheeler, D. L., Chappey, C., Lash, A. E., Leipe, D. D., Madden, T. L., Schuler, G. D., Tatusova, T. A., and Rapp, B. A. (2000). Database resources of the National Center for Biotechnology Information. Nucleic Acids Res *28*, 10-4.