SUPPLEMENTARY DATA

Yosshi: a web-server for protein disulfide engineering by bioinformatic analysis of diverse protein families

Dmitry Suplatov*, Daria Timonina, Yana Sharapova and Vytas Švedas

Lomonosov Moscow State University, Belozersky Institute of Physicochemical Biology and Faculty of Bioengineering and Bioinformatics, Vorobjev hills 1-73, Moscow 119991, Russia

* To whom correspondence should be addressed. Tel: +74959394653; Email: d.a.suplatov@belozersky.msu.ru

1. Libraries of 3D-motifs of disulfide bonds

A set of PDB structures downloaded on October 28th, 2018 with a resolution of ≤2.5 Å [1] was prepared and filtered for redundancy at the 95% pairwise sequence identity threshold by the CD-HIT tool [2]. Analysis of closely located pairs of cysteines with $\mathrm{dist}(S\gamma - S'\gamma) \leq 2.5$ Å in the selected structures showed that $\mathrm{dist}(S\gamma - S'\gamma)$ has a normal distribution with a mean 2.05 Å and a standard deviation 0.06 Å, in agreement with previous studies of the disulfide connectivity in proteins [1, 3]. Only the pairs of cysteines with 1.88 Å $\leq \mathrm{dist}(S\gamma - S'\gamma) \leq 2.22$ Å (i.e., the $\mu \pm 3\sigma$ interval) were further considered. The statistics for $\mathrm{dist}(C\alpha - C'\alpha)$ and $\mathrm{dist}(C\beta - C'\beta)$ in the selected S-S bonds is provided in Table S1. The coordinates of all heavy side-chain and backbone atoms of each selected disulfide were saved as separate objects to form the initial pool of 16956 disulfide bond variations which were further subjected to the similarity analysis and clustering based on the values of all angles (Cα-Cβ-Sγ, Cβ-Sγ-S'γ, Sγ-S'γ-C'β, S'γ-C'β-C'α) and dihedral angles (C-Cα-Cβ-Sγ, Cα-Cβ-Sγ-S'γ, Cβ-Sγ-S'γ-C'β, Sγ-S'γ-C'β-C'α, S'γ-C'β-C'α-C') within each S-S bond that define its particular configuration (Fig. S1). For each disulfide *i* the phase space of the corresponding nine angles $P_i = \{\varphi_1, \ldots, \varphi_9\}$ was transformed to $P_i \ \varphi \ = \{\cos\varphi_1\,, \sin\varphi_1\,, \ldots, \cos\varphi_9\,, \sin\varphi_9\}$, similarly to what has been recently discussed [4]. This representation constitutes a local isometry with preservation of distance measure up to a constant [5]. The similarity between every two S-S bonds *i* and *j* was measured as a distance between the corresponding vectors $P_i(\varphi)$ and $P_j(\varphi)$, with smaller values indicating closer resemblance, and used to cluster disulfide configurations by the DBSCAN machine-learning technique [6]. The *eps* parameter for that method (i.e., a step size for expanding clusters) was optimized in the range [0.01; 5] with a 0.01 increment based on the 'silhouette estimator' technique. The 'silhouette' values for each proposed classification representing how similar S-S bond configurations are within their own clusters and different compared to other clusters were combined into a single plot, followed by analysis of this graphical representation to select the most appropriate classification as previously discussed [7]. Clusters with a diameter of more than 1.5 (i.e., indicating on average a poor in-group similarity) were re-grouped into smaller clusters with higher in-group resemblance. The DBSCAN produced 273 clusters which represented 72% of the currently known disulfide bond variability and 4748 outliers which corresponded to unique S-S bond configurations. Only one representative disulfide configuration *i* was selected from each cluster by minimizing its distances d_{ij} to all other configurations in that cluster, i.e., $argmin_i \sum_j d_{ij}$. Finally, two non-redundant collections of 3D-motifs were produced which incorporated the current knowledge of the disulfide bond geometry: the "Core collection" which contained 273 most typical configurations (i.e., the selected representatives of each cluster), and the "Complete collection" which in addition to that contained 4748 unique disulfide bonds (i.e., 5021 3D-motifs in total). The current version of the 3D-motif libraries can be downloaded from https://biokinet.belozersky.msu.ru/yosshiserver/3D-motif-libraries.tar.gz.

Table S1. Statistics for $\mathrm{dist}(\mathrm{C}\alpha - \mathrm{C}'\alpha)$ and $\mathrm{dist}(\mathrm{C}\beta - \mathrm{C}'\beta)$ in the selected S-S bonds obtained from a non-redundant set of high-quality PDB structures with a resolution of ≤2.5 Å

	μ, Å	σ, Å	$\mu + 3\sigma$, Å
$\mathrm{dist}(\mathrm{C}\alpha - \mathrm{C}'\alpha)$	5.51	0.73	7.70
$\mathrm{dist}(\mathrm{C}\beta - \mathrm{C}'\beta)$	3.85	0.21	4.48

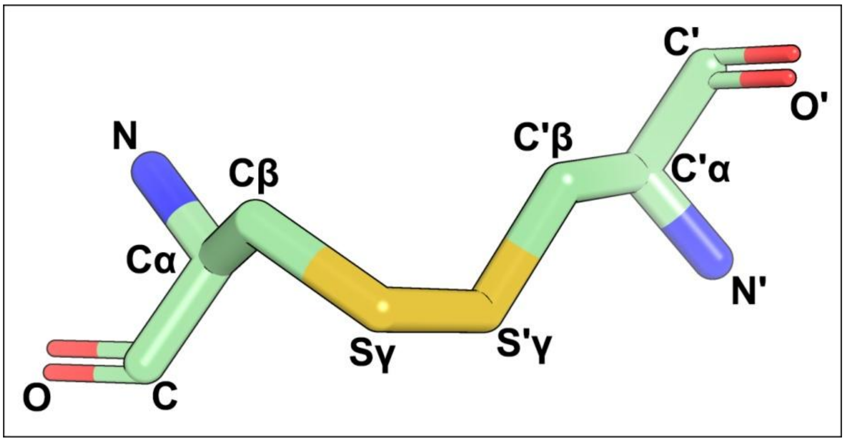


Fig. S1. A 3D-motif of a disulfide bond contains 12 atoms of the backbone and side-chain

2. Structural filtration

The Yosshi workflow consists of bioinformatic analysis to search for pairs of cysteine residues in sequences of homologs, and structural filtration to evaluate whether introduction of the selected cysteines at corresponding positions of the query structure can result in a formation of a disulfide bond. The previously reported disulfide engineering experiments have demonstrated that introduction of a crosslink into a protein structure is likely to result in a considerable shift of the backbone atoms, thus the candidate hot-spot positions cannot be expected to match the strict geometric constraints of an S-S bond (i.e., of the two covalently connected cysteines) [8, 9]. Therefore, to perform such a structural filtration it is important to consider the flexibility of a pair of non-bonded amino acid residues which can form an S-S bridge upon mutation to cysteines. Numerous attempts to embed disulfide bridges in proteins with a known PDB structure were reported in the literature (see section "Introduction" in the Main text). Systematic analysis of the published data could be done to characterize the conformational changes provoked by a disulfide bond formation; however, such data is likely to be confined and biased, as disulfide engineering performed so far was mainly focused on a limited set of industrially relevant enzymes and a narrow range of structural folds (e.g., lipases/esterases/peptidases of the α/β-hydrolases superfamily). To avoid bias in this study a large non-redundant set of disulfide bonds (C1 and C2, Fig. S2) and their non-bonded equivalences in structures of homologous proteins (E1 and E2, Fig. S2) was collected by bioinformatic analysis of domain superfamilies according to the CATH classification [10]. The CATH v. 4.2 data was downloaded on October 28th, 2018. Geometry constraints of the backbone atoms in the collected non-bonded residue pairs were further studied and used to set the structural filtration thresholds within the Yosshi workflow.

This section is organized as follows. First, we describe the procedure to collect disulfide bonds and their non-bonded equivalences in homologous proteins by bioinformatic analysis of the CATH superfamilies (subsection 2.1). Then, we discuss the selection of default thresholds for the “First tier filter” (subsection 2.2) and parameters for the “Flexible” and “Rigid” statistical models for the 3D-motif analysis (subsection 2.3). Finally, we validate the structural filtration with different settings on a large dataset (subsection 2.4).

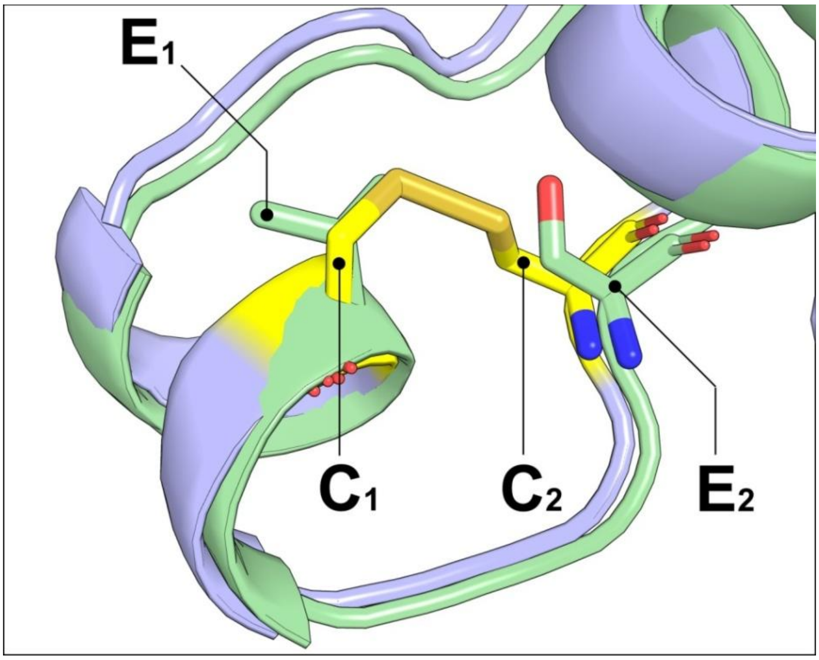


Fig. S2. Cysteines C_1 and C_2 capable of a disulfide bond formation, and the corresponding non-bonded equivalences (i.e., amino acid residues E_1 and E_2) in a homologous protein structure collected by the bioinformatic analysis. The illustration was prepared using structural superimposition of 5'-nucleotidases from *Thermus thermophilus* (PDB 2Z1A, Val348 and Ser353, chain A) and *Homo sapiens* (PDB 4H1S, Cys353 and Cys358, chain B).

2.1 Analysis of disulfide bonds and their non-bonded equivalences in homologous proteins

In this subsection we describe the procedure to collect disulfide bonds and their non-bonded equivalences in homologous proteins to be further used to select the default thresholds for the “First tier filter” (see subsection 2.2) and parameters for the “Flexible” statistical model for the 3D-motif analysis (see subsection 2.3).

Non-redundant sets of protein domain structures (at least two structures per set and at most 40% pairwise sequence identity with each other) with a resolution of ≤2.5 Å corresponding to superfamilies of the CATH v. 4.2 classification were collected and independently processed in three steps as follows: (1) the core protein was selected in each superfamily; (2) all pairwise structural alignments between proteins within a superfamily and the selected core protein were constructed; (3) all disulfide bonds occurring in protein structures within a superfamily were mapped onto the core protein structure and further subjected to a set filters to select pairs of non-bonded amino acid residues in that protein capable to form an S-S bridge assuming both positions are mutated to cysteines. The details are provided below.

In the first step, a core protein was selected in each superfamily. All pairwise structural superimpositions between all proteins within a superfamily were carried out using the *superpose* algorithm and corresponding Q-scores were estimated [11]. A Q-score is a quality function of Cα-alignment which takes both the alignment length and RMSD into account. Q-score reaches 1 only in the case of identical structures, and drops down with increasing RMSD or decreasing alignment length. For every protein *i* the sum of $Q_{i,j}$-scores of pairwise alignments with all other proteins in the set was calculated and used as a measure of an overall structural similarity to this set. The protein *i* with the highest $\sum_j Q_{i,j}$ score (i.e., indicating on average the largest structural similarity of that protein to all other proteins currently in the set) was selected as the core.

In the second step, all pairwise structural alignments between proteins within a superfamily and the selected core protein were created. The *superpose* program used on the previous step is perfect for a high-throughput database search and overall similarity estimation, but the exact superimposition of protein structures produced by that algorithm can be dubious, especially when aligning evolutionary distant homologs. Therefore, three different 3D-alignment algorithms were used on this step to independently create high-quality pairwise structural superimpositions – MUSTANG [12], MATT [13], and mTM-align [14]. Both MATT and MUSTANG were noted in a recent study for a good performance aligning distant relationships and length variations [15], and the mTM-align was considered here because it seems to be the most recent software for structural comparison available.

In the final step, all disulfide bonds in protein structures within a superfamily were identified using the previously discussed structural criteria (see section 1 of this Supplementary Data), and then their non-bonded equivalences in the core protein were selected as follows. Each pair of residues C_1 and C_2 forming a disulfide bond in a protein structure was mapped to positions E_1 and E_2 in the selected core protein (Fig. S2) if the corresponding superimposition of C_1-E_1 and C_2-E_2 was supported by all three 3D-alignment algorithms, i.e., MUSTANG, MATT, and mTM-align all agreed about how these two pairs of residues in protein structures should be aligned, thus reducing the likelihood of an alignment error. The E_1-E_2 in the core protein structure was further confirmed as a non-bonded equivalent of the disulfide bond C_1-C_2 in a homologous protein structure if all the following criteria were met:

- At least one of the two amino acid residues at positions E_1 and E_2 was not a cysteine or the two cysteines were not connected by a covalent bond in the selected core protein structure;
- Pairs of amino acid residues C_1-E_1 and C_2-E_2 were located in regions of the same secondary structure assigned by the STRIDE algorithm [16], e.g. this filter was passed if C_1/E_1 were

located in α-helixes, and C_2/E_2 were located in extended β-strands in both structures. The contrary could indicate significant structural differences between the corresponding positions in homologs making it hard to determine without experimental evaluation whether E_1-E_2 in the selected core protein would form a disulfide bond upon mutation to cysteines;

- Neither E_1 nor E_2 in the core protein structure were occupied by a proline. The contrary could indicate significant structural differences between the corresponding positions in homologs. Although some disulfide engineering experiments were reported that successfully improved stability by substituting prolines for an S-S bond [e.g., 17], such equivalences are dubious without experimental evaluation and were dismissed in this study for the sake of clarity;
- Neither E_1 nor E_2 in the core protein structure were located in regions of high mobility indicated by: (1) a high B-factor of the backbone atoms C, Cα, N, and O at corresponding positions (i.e., exceeding $\mu + 2\sigma$ compared to the whole protein), or (2) incomplete/missing backbone atoms located within ±3 amino acid residues from E_1 or E_2, or (3) presence of E_1 or E_2 at the C-/N-terminus of a protein chain. The contrary could mean that positions E_1 and E_2 have inaccurate or alternative coordinates in the corresponding PDB records;
- Finally, the automatically collected pairs C_1-E_1/C_2-E_2 were subjected to manual curation as follows. In total 11% of pairs were dismissed at this step as visual expert inspection of the automatically created 3D alignments revealed a controversy in the residue superimposition in the two protein structures, or significant structural differences between the corresponding positions in homologs, e.g., caused by insertions or deletions that led to a significantly increased gap between E_1 and E_2 compared to C_1 and C_2, making it hard to determine without experimental evaluation whether E_1-E_2 in the selected core protein would form a disulfide bond upon mutation to cysteines.

These strict criteria were passed by 187 equivalences C_1-E_1/C_2-E_2 (123 unique residue pairs E_1-E_2 in the core protein structures) located in the regions of high structure similarity between the selected core proteins and their homologs in 103 CATH superfamilies. It was further assumed that the collected non-bonded residue pairs E_1 and E_2 in the corresponding core proteins would form a disulfide bond if both are mutated to cysteines, as observed in the structures of homologous proteins. Geometry constraints of the backbone atoms in the collected non-bonded residue pairs E_1 and E_2 were further studied and used to set the structural filtration thresholds within the Yosshi workflow, as further discussed in subsections 2.2 and 2.3 below.

Table S2. Structural and statistical analysis of the non-bonded equivalences E_1 and E_2 of S-S bonds in homologous proteins

Characteristic	μ (Å)	σ (Å)	P(x > X) = p	
			X (Å)	p
$\mathrm{dist}(C\alpha - C'\alpha)$	5.70	0.96	8.58	0.00135
$\mathrm{dist}(C\beta - C'\beta)$	4.86	0.70	6.96	0.00135
$RMSD_{C_i-E_i}^{backbone}$	0.39	0.19	0.70	0.05

The p-value of 0.00135 corresponds to the $\mu + 3\sigma$ quantile of a normal distribution.

2.2 Thresholds for the “First tier filter”

The “First tier filter” represents first of the two structural filtration steps and is applied to the query protein structure to dismiss all pairs of positions the least likely to form a crosslink even if both were mutated to cysteines using simple $\mathrm{dist}(\mathrm{C\alpha} - \mathrm{C'\alpha})$ and $\mathrm{dist}(\mathrm{C\beta} - \mathrm{C'\beta})$ distance criteria. This step is casually performed by disulfide engineering algorithms, but the threshold selection is usually based on the restrictive model of two covalently bound cysteines, e.g., all amino acid pairs are considered as potential hot-spots if their Cβ atoms are within 4.5 Å [18], 5.21 Å [19], or 5.5 Å [8, 20]. There are numerous reports of successfully engineered disulfide bonds that violate geometric constraints of these rigid computational models [8, 9]. Thus, in this web-server a flexible geometry model is applied – the cut-off values are set to $\mathrm{dist}(\mathrm{C\alpha} - \mathrm{C'\alpha}) \leq 8.58$ Å and $\mathrm{dist}(\mathrm{C\beta} - \mathrm{C'\beta}) \leq 6.96$ Å, what corresponds to $\mu + 3\sigma$ of the values calculated over 123 non-bonded equivalences E_1-E_2 of disulfide bonds in homologous proteins collected as discussed in subsection 2.1 above (Table S2, Fig. S2). The aim of the “First tier filter” is to accelerate further processing by removing the least appropriate candidate positions. Therefore, a low first-order error of 0.135% (i.e., $\mathrm{P}(\mathrm{x} > \mu + 3\sigma) = 0.00135$ assuming the distances are normally distributed) is considered for threshold selection at this step to minimize the risk of rejecting two candidate positions which may, in fact, form a correct disulfide bond at a cost of failing to reject two candidate positions incapable of a disulfide bond formation. This initial selection will be evaluated by two more steps of the algorithm, in particular, a more rigorous analysis of the geometry of each pair of candidate positions will be performed at the last step of the workflow by the 3D-motif analysis (see subsection 2.3 below).

2.3 Statistical models for the 3D-motif analysis

In the context of this study, a 3D-motif of a disulfide bond contains 12 atoms of the backbone and side-chain of two covalently linked cysteines observed in a high-quality crystallographic structure (Fig. S1). At the 3D-motif analysis step a pair of candidate positions in the query protein structure which passed the “First tier filter” and was further selected by the bioinformatic analysis (i.e., both positions were occupied by cysteines in at least one homolog) would be confirmed as a promising site for S-S bond formation if it matches with at least one 3D-motif from the selected collection (see section 1 above). Two statistical models were proposed to evaluate whether a 3D-motif matches a pair of positions in the query protein structure.

The **“Flexible” statistical model** for the 3D-motif analysis and selection of the most promising sites in the query structure is based on bioinformatic analysis of homologous proteins with different disulfide connectivity to take into account the flexibility of a pair of non-bonded amino acid residues that can form an S-S bridge upon mutation to cysteines. This statistical model was trained by comparing the disulfide bonds C_1-C_2 with their non-bonded equivalences E_1-E_2 in homologous proteins (Fig. S2), collected as discussed in subsection 2.1 above. The atoms corresponding to C_1-C_2 and E_1-E_2 were cut out from corresponding PDB structures and structurally superimposed for best-fit of the backbone. Such a superimposition was used to calculate two values of the root mean square deviation of the backbone atoms $RMSD_{C_1-E_1}^{backbone}$ and $RMSD_{C_2-E_2}^{backbone}$. From each best-fit comparison only the largest value $RMSD_{C_i-E_i}^{backbone}$ was preserved. If a non-bonded pair of positions E_1-E_2 was equivalent to multiple disulfide bonds C_1-C_2 in homologs within a superfamily, then only the largest RMSD value was selected among all pairwise best-fit comparisons. This analysis was performed for all collected 187 equivalences C_1-E_1/C_2-E_2 and resulted in 123 independent $RMSD_{C_i-E_i}^{backbone}$ values with μ=0.39 Å and σ=0.19 Å which describe the flexibility of a pair of non-bonded amino acid residues capable of a disulfide bond formation upon mutation to cysteines. Assuming the normal distribution of $RMSD_{C_i-E_i}^{backbone}$ the corresponding model was used for the 3D-motif analysis in the Yosshi workflow. At this last step of the workflow each 3D-motif of a disulfide bond from the selected collection is structurally superimposed with the provided pair of candidate positions in the query protein structure for best-fit of the backbone atoms. A pair of positions in the query protein structure would be confirmed as a promising site for S-S bond formation if it matches with at least one 3D-motif so that both RMSD values between two pairs of superimposed backbone atoms are within X=0.70 Å, what corresponds to a p-value of $\mathrm{P(x > X) = 0.05}$ of a normal distribution with μ=0.39 Å and σ=0.19 Å, or rejected otherwise.

Alternatively, the **"Rigid" statistical model** has been proposed based on the analysis of 273 clusters incorporating known S-S bond configurations produced by the DBSCAN (see section 1 in this Supplementary Data). Within each cluster, all pairwise structural superimpositions between the representative disulfide configuration (selected as described in section 1) and all other members of the cluster were performed. For each best-fit pairwise comparison two RMSD values were calculated for each pair of backbone atoms, and then only the largest RMSD value was selected over a cluster. Finally, the produced 273 independent values were used to estimate the statistical model. Under this "Rigid" model, a pair of positions in the query protein structure would be confirmed as a promising site for S-S bond formation if it matches with at least one 3D-motif so that both RMSD values between two pairs of superimposed backbone atoms are within X=0.28 Å, what corresponds to a p-value of $\mathrm{P(x > X) = 0.05}$ of a normal distribution with μ=0.16 Å and σ=0.07 Å, or rejected otherwise. Evaluation of the structural filtration with these parameters is further discussed in subsection 2.4 below.

2.4 Benchmarking of the structural filtration

The error rate of computational methods for structure-based disulfide engineering is not well characterized [8]. The disulfide bond formation in proteins is well-known to depend on various details such as pH, ionic strength, temperature, time, presence of oxidizing/reducing agents in the solution, choice of cysteine-protecting groups and corresponding deblocking conditions, etc. [21]. Therefore, it is hard to determine the connectivity status of non-bonded cysteines solely from protein crystallographic structures, i.e., if two cysteines are not connected by an S-S bond in a particular PDB record but are spatially close, they could still form a covalent bond under different experimental conditions. Consequently, an accurate estimation of the specificity (i.e., $1 - False\ Positive\ Rate$) of computational methods for disulfide engineering is implausible, as it would require a large number of predictions over a wide range of proteins, followed by a thorough experimental insight [8]. As a result, only sensitivity is reported for such algorithms, i.e., the ability to correctly predict known S-S bonds. E.g., probably the most popular disulfide engineering software "Disulfide by Design" was reported to correctly identify 706 out of the 710 true disulfides, for a sensitivity of 99.4%; however, specificity of the algorithm was not estimated at that time [22].

Table S3. Results of benchmarking of the structural filtration on “True bonds” and “True non-bonds” collections of pairs of cysteine residues in protein structures

Stat. model	3D-motif library	TP	FN	FP	TN	Sen, %	Spec, %
Flexible	**Core**	13291	15	223	759	**99.89**	**77.29**
	Complete	13306	0	867	115	100	11.71
Rigid	Core	10453	2853	0	982	78.56	100
	Complete	12985	321	52	930	**97.59**	**94.71**

TP – true positives, TN – true negatives, FP – false positives, FN – false negatives; TP and FN – the number of pairs from the “True bonds” collection which were confirmed as promising sites for S-S bond formation or rejected as such, respectively. FP and TN – the number of pairs from the “True non-bonds” collection which were confirmed as promising sites for S-S bond formation or rejected as such, respectively. For each statistical model the 3D-motif library that provided the best outcome (i.e., the balance of sensitivity and specificity) is highlighted in bold

Here we attempted a broader study to evaluate the performance of the Yosshi’s structural filtration workflow. First, we validated the 3D-motif analysis on a large dataset by showing that the method can discriminate between a true disulfide bond (i.e., “True bonds”) and a pair of cysteine residues located at a close range but unlikely to form an S-S bridge (i.e., “True non-bonds”). Second, were used the 3D-motif analysis to study pairs of non-bonded cysteines in protein structures whose connectivity status cannot be determined solely from the corresponding PDB record (i.e., “Connectivity status unknown”). In all cases, the 3D-motif analysis was tested with different setup, i.e., the two 3D-motif libraries (“Core collection” and “Complete collection”, see section 1 of this Supplementary Data) and the two statistical models (“Flexible” and “Rigid”, see subsection 2.3), and the observed difference in outcomes is discussed at the end of this subsection. To collect a set of cysteine pairs for this evaluation the X-Ray structures downloaded on October 28th, 2018 with a resolution within the (2.5; 4] Å interval (i.e., structures which were considered neither for construction of the 3D-motif libraries nor for the bioinformatic analysis of the CATH superfamilies, but still had a reasonable quality) were filtered for redundancy at the 95% pairwise sequence identity threshold by the CD-HIT [2].

First, we validated the 3D-motif analysis on a large dataset of true disulfide bonds and pairs of cysteines located at a close range but unlikely to form an S-S bridge – to estimate sensitivity and specificity of the new algorithm at different setup. All pairs of cysteines which had their respective $S\gamma$ atoms within $\mathrm{dist}(S\gamma - S'\gamma) \leq 2.5$ Å were considered as “True bonds” like in [3]. Then, the “First tier filter” with the default parameters was applied to each protein structure, and positions occupied by non-bonded cysteines so that $dist(\mathrm{C\alpha} - \mathrm{C'\alpha}) > 7.70$ Å and $dist(\mathrm{C\beta} - \mathrm{C'\beta}) > 4.48$ Å (i.e., larger than $\mu + 3\sigma$ for corresponding distances calculated from S-S bonds observed in high-resolution PDB records; see section 1 and Table S1 above) were considered as “True non-bonds”. The two distance criteria were applied to dismiss pairs of free cysteines which can, in principle, form an S-S bond, but were captured in a reduced state in a particular crystal structure, and select only those pairs of free cysteine residues which were located at a close range but characterized by unusual distances between key backbone atoms and thus might not be capable of forming a disulfide bond. The “True bonds” and “True non-bonds” collections contained 13306 and 982 pairs of positions, respectively. The “First tier filter” and the 3D-motif analysis were applied to the compiled benchmark set of positions. The results are summarized in Table S3. Under the “Flexible” statistical model the $Sensitivity = \frac{TP}{TP+FN}$ (i.e., TP – true positives and FN – false negatives) of the structural filtration was 99.89% when using the “Core collection” which contained only 273 most typical configurations of S-S bonds, and 100% when using the “Complete collection” of 5021 3D-motifs. The highly similar outcomes of the search versus the two libraries can be explained by the fact that RMSD between many of the rare 3D-motifs in the “Complete collection” was within the default threshold (i.e., 0.70 Å under the “Flexible” statistical model) with at least one typical 3D-motif in the “Core collection”. The $Specificity = \frac{TN}{TN+FP}$ (i.e., TN – true negatives, FP – false positives) of the structural filtration under same conditions was 77.29% and 11.71% for the two libraries, respectively (i.e., the search using the “Core collection” was significantly more specific, McNemar's Chi-squared test [23] p-value=0.008). Alternatively, under the “Rigid” statistical model the sensitivity and specificity were 78.56% and 100% when using the “Core collection” library, or 97.59% and 94.71% when using the “Complete collection” library, respectively (Table S3).

Table S4. The 3D-motif analysis of the “Connectivity status unknown” collection

Stat. model	3D-motif library	Confirmed	Rejected	Confirmed vs total, %
Flexible	Core	4819	3222	59.93
	Complete	7327	714	91.12
Rigid	Core	887	7154	11.03
	Complete	2977	5064	37.02

“Confirmed”, and “Rejected” – the number of pairs from the “Connectivity status unknown” collection which were confirmed as promising sites for S-S bond formation or rejected as such, respectively.

Second, all pairs of cysteines which passed the “First tier filter” but were previously assigned neither to the “True bonds” nor to the “True non-bonds” collections were evaluated. The corresponding residues did not form a crosslink in the selected protein structures, but had $dist(\mathrm{C\alpha - C'\alpha})$ and/or $dist(\mathrm{C\beta - C'\beta})$ within a range typical for known S-S bonds and thus could potentially form a covalent bond under appropriate experimental conditions. Therefore, as explained above, without a thorough experimental insight into each particular case it is hard to learn whether these pairs of cysteines in the selected protein structures are capable of a disulfide bond formation either *in vivo* or *in vitro*. The corresponding 8041 cases were assigned to the “Connectivity status unknown” collection and subjected to the 3D-motif analysis. The results are provided in Table S4. Cysteine is the most rarely occurring amino acid residue in the known protein Universe with only 1.2% frequency in the latest release of the TrEMBL database (https://www.uniprot.org/statistics/TrEMBL), and thus the probability to observe two closely located cysteines just by chance is the lowest among all other pairs of amino acid residues in a protein structure. The 3D-motif analysis showed that many of these closely located non-bonded cysteine pairs may, in fact, form a disulfide bond which can affect protein stability, function, and regulation under appropriate environmental/experimental conditions. Further experimental insight is required to learn their connectivity status and its implication to protein function.

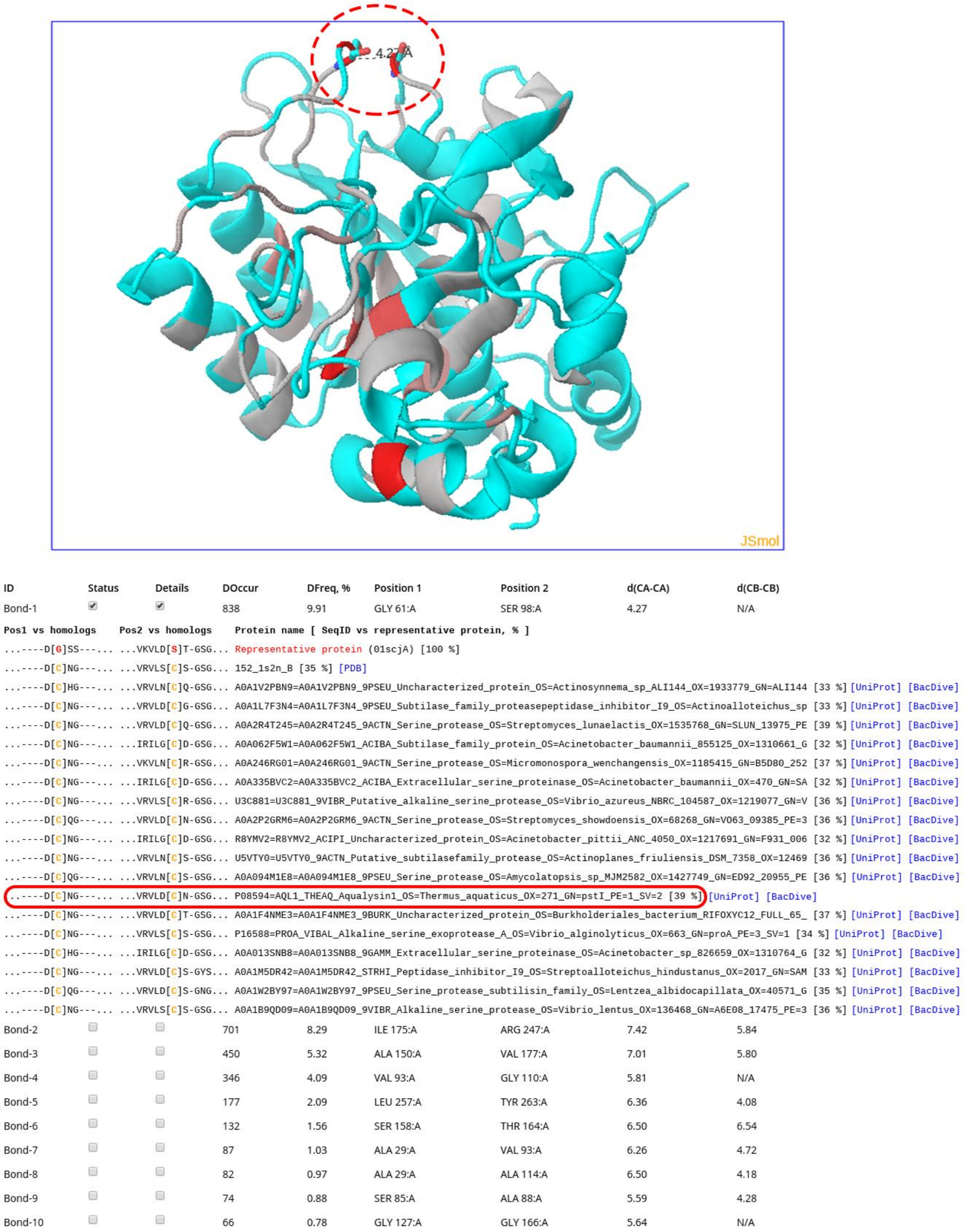
The obtained results indicate that the “Rigid” statistical model which was trained on a set of known S-S bridges can better discriminate between pairs of bonded and non-connected cysteines. However, as previously discussed, there are reports of successfully engineered S-S bonds that violate geometric constraints of such rigid computational models trained on covalently connected cysteines [8, 9]. On the contrary, the “Flexible” statistical model was based on the bioinformatic analysis of disulfide bonds and their non-bonded equivalences in homologous protein structures to take into account the flexibility of backbone atoms. The benchmarking showed that implementation of the “Core collection” library of the 273 3D-motifs of S-S bonds under the “Flexible” statistical model provides a superb sensitivity and a promising specificity for the 3D-motif analysis, thus this combination of parameters is used as the default setup in the Yosshi web-server. Alternatively, the user can choose to perform the 3D-motif analysis with the “Rigid” statistical model to select only those sites for S-S bond engineering whose geometry in the query protein structure is highly similar to the geometry of the known covalently connected cysteines. In such a case, the “Complete collection” of 3D-motifs will be automatically activated as the default setup, due to a good performance of this combination of parameters at the benchmarking.

3. Examples

In addition to the bioinformatic analysis of subtilisin (PDB 1SCJ, chain A) and myoglobin (PDB 1JP6, chain A) and corresponding superfamilies discussed in the Main text (see section “Results” in that file, and also Fig. S3) the Yosshi+Mustguseal integrated tool was used to study disulfide connectivity in different lipases, carbonic anhydrases, xylanases, and members of the ribonuclease A superfamily to reproduce the previously reported disulfide engineering experiments. In all case-studies, Yosshi and Mustguseal were used with the default setup (see section “Materials and Methods” in the Main text). The results obtained with the 3D-motif analysis set to the “Flexible” mode are discussed below, and the output in the “Rigid” mode is summarized in Table 1 (see the Main text). The input and output data regarding all case-studies are available on-line at https://biokinet.belozersky.msu.ru/yosshi-example.

The PDB structure 4GW3 (chain A) of lipase from *Proteus mirabilis* was submitted as a query to the Mustguseal web-server to automatically construct an alignment of a non-redundant set of 1680 homologs with high structure similarity but low sequence identity to the query protein. Further analysis of the alignment by Yosshi provided 5 pairs of hot-spots for disulfide engineering ranked based on the expected occurrence of the corresponding crosslinks in homologs. The pair #1 corresponded to positions Gly181 and Ser238 both occupied by cysteines in 1092 proteins (65%). Introduction of this naturally occurring crosslink into the structure of lipase from *Proteus mirabilis* by a double mutation Gly181Cys/Ser238Cys was previously reported to increase the half-inactivation temperature from 37°C to 48°C compared to the wild-type protein [24].

In a separate case-study the Yosshi+Mustguseal tool was used to study disulfide connectivity in the Mono- and diacylglycerol lipase (MDLA) from *Penicillium cyclopium* (PDB 5CH8, chain A). The corresponding structure-guided sequence alignment constructed by the Mustguseal contained 853 proteins. Further analysis of the alignment by Yosshi provided 14 pairs of hot-spots for disulfide engineering. The pair #2 corresponded to positions Tyr22 and Gly269 both occupied by cysteines in 598 proteins (70% of the set). Introduction of this naturally occurring crosslink into the structure of MDLA from *Penicillium camembertii* (Uniprot:P61870) which was 100% identical to the query protein MDLA from *Penicillium cyclopium* (Uniprot:P61869) both in length and sequence by a double mutation Tyr22Cys/Gly269Cys was previously reported to increase the melting temperature from 51°C to 63°C compared to the wild-type protein [25]. The MDLA from *Penicillium cyclopium* was used as a query in this case-study because the full-size crystallographic structure of MDLA from *Penicillium camembertii* was not available.



ID	Status	Details	DOccur	DFreq, %	Position 1	Position 2	d(CA-CA)	d(CB-CB)
Bond-1	☑	☑	838	9.91	GLY 61:A	SER 98:A	4.27	N/A

```
Pos1 vs homologs   Pos2 vs homologs   Protein name [ SeqID vs representative protein, % ]
...----D[G]SS---... ...VKVLD[S]T-GSG... Representative protein (01scjA) [100 %]
...----D[C]NG---... ...VRVLS[C]S-GSG... 152_1s2n_B [35 %] [PDB]
...----D[C]HG---... ...VRVLN[C]Q-GSG... A0A1V2PBN9=A0A1V2PBN9_9PSEU_Uncharacterized_protein_OS=Actinosynnema_sp_ALI144_OX=1933779_GN=ALI144 [33 %] [UniProt] [BacDive]
...----D[C]NG---... ...VRVLD[C]G-GSG... A0A1L7F3N4=A0A1L7F3N4_9PSEU_Subtilase_family_proteasepeptidase_inhibitor_I9_OS=Actinoalloteichus_sp [33 %] [UniProt] [BacDive]
...----D[C]NG---... ...VRVLD[C]Q-GSG... A0A2R4T245=A0A2R4T245_9ACTN_Serine_protease_OS=Streptomyces_lunaelactis_OX=1535768_GN=SLUN_13975_PE [39 %] [UniProt] [BacDive]
...----D[C]NG---... ...IRILG[C]D-GSG... A0A062F5W1=A0A062F5W1_ACIBA_Subtilase_family_protein_OS=Acinetobacter_baumannii_855125_OX=1310661_G [32 %] [UniProt] [BacDive]
...----D[C]NG---... ...VKVLN[C]R-GSG... A0A246RG01=A0A246RG01_9ACTN_Serine_protease_OS=Micromonospora_wenchangensis_OX=1185415_GN=B5D80_252 [37 %] [UniProt] [BacDive]
...----D[C]NG---... ...IRILG[C]D-GSG... A0A335BVC2=A0A335BVC2_ACIBA_Extracellular_serine_proteinase_OS=Acinetobacter_baumannii_OX=470_GN=SA [32 %] [UniProt] [BacDive]
...----D[C]NG---... ...VRVLS[C]R-GSG... U3C881=U3C881_9VIBR_Putative_alkaline_serine_protease_OS=Vibrio_azureus_NBRC_104587_OX=1219077_GN=V [36 %] [UniProt] [BacDive]
...----D[C]QG---... ...VRVLD[C]N-GSG... A0A2P2GRM6=A0A2P2GRM6_9ACTN_Serine_protease_OS=Streptomyces_showdoensis_OX=68268_GN=V063_09385_PE=3 [36 %] [UniProt] [BacDive]
...----D[C]NG---... ...IRILG[C]D-GSG... R8YMV2=R8YMV2_ACIPI_Uncharacterized_protein_OS=Acinetobacter_pittii_ANC_4050_OX=1217691_GN=F931_006 [32 %] [UniProt] [BacDive]
...----D[C]NG---... ...VRVLN[C]S-GSG... U5VTY0=U5VTY0_9ACTN_Putative_subtilasefamily_protease_OS=Actinoplanes_friuliensis_DSM_7358_OX=12469 [36 %] [UniProt] [BacDive]
...----D[C]QG---... ...VRVLN[C]S-GSG... A0A094M1E8=A0A094M1E8_9PSEU_Serine_protease_OS=Amycolatopsis_sp_MJM2582_OX=1427749_GN=ED92_20955_PE [36 %] [UniProt] [BacDive]
...----D[C]NG---... ...VRVLD[C]N-GSG... P08594=AQL1_THEAQ_Aqualysin1_OS=Thermus_aquaticus_OX=271_GN=pstI_PE=1_SV=2 [39 %] [UniProt] [BacDive]
...----D[C]NG---... ...VRVLD[C]T-GSG... A0A1F4NME3=A0A1F4NME3_9BURK_Uncharacterized_protein_OS=Burkholderiales_bacterium_RIFOXYC12_FULL_65_ [37 %] [UniProt] [BacDive]
...----D[C]NG---... ...VRVLS[C]S-GSG... P16588=PROA_VIBAL_Alkaline_serine_exoprotease_A_OS=Vibrio_alginolyticus_OX=663_GN=proA_PE=3_SV=1 [34 %] [UniProt] [BacDive]
...----D[C]HG---... ...IRILG[C]D-GSG... A0A013SNB8=A0A013SNB8_9GAMM_Extracellular_serine_proteinase_OS=Acinetobacter_sp_826659_OX=1310764_G [32 %] [UniProt] [BacDive]
...----D[C]NG---... ...VRVLD[C]S-GYS... A0A1M5DR42=A0A1M5DR42_STRHI_Peptidase_inhibitor_I9_OS=Streptoalloteichus_hindustanus_OX=2017_GN=SAM [33 %] [UniProt] [BacDive]
...----D[C]QG---... ...VRVLD[C]S-GNG... A0A1W2BY97=A0A1W2BY97_9PSEU_Serine_protease_subtilisin_family_OS=Lentzea_albidocapillata_OX=40571_G [35 %] [UniProt] [BacDive]
...----D[C]NG---... ...VRVLS[C]S-GSG... A0A1B9QD09=A0A1B9QD09_9VIBR_Alkaline_serine_protease_OS=Vibrio_lentus_OX=136468_GN=A6E08_17475_PE=3 [36 %] [UniProt] [BacDive]
```

Bond-2	□	□	701	8.29	ILE 175:A	ARG 247:A	7.42	5.84
Bond-3	□	□	450	5.32	ALA 150:A	VAL 177:A	7.01	5.80
Bond-4	□	□	346	4.09	VAL 93:A	GLY 110:A	5.81	N/A
Bond-5	□	□	177	2.09	LEU 257:A	TYR 263:A	6.36	4.08
Bond-6	□	□	132	1.56	SER 158:A	THR 164:A	6.50	6.54
Bond-7	□	□	87	1.03	ALA 29:A	VAL 93:A	6.26	4.72
Bond-8	□	□	82	0.97	ALA 29:A	ALA 114:A	6.50	4.18
Bond-9	□	□	74	0.88	SER 85:A	ALA 88:A	5.59	4.28
Bond-10	□	□	66	0.78	GLY 127:A	GLY 166:A	5.64	N/A

Fig. S3. Interactive on-line analysis page to study the Yosshi output. The structure of subtilisin E from *Bacillus subtilis* is shown in the 3D viewer (see section "Results" in the Main text for details). The backbone of the query protein is gradient painted from red to grey according to the expected abundance of an S-S bond at the corresponding positions in homologs, with intensive red indicating pairs of positions most frequently hosting disulfides. A pair of residues Gly61 and Ser98 was ranked #1 in the Yosshi output (shown as sticks and indicated by a dashed oval) as 838 proteins were discovered to contain cysteines in the equivalent positions. The corresponding sub-sequences of some of these homologs are shown (cysteines are colored in yellow). The entry corresponding to a homolog from a thermophilic organism (Aqualysin-1 from *Thermus aquaticus*) is indicated by a red oval. For each protein the HTML links to the respective pages in PDB, UniProt, and BacDive databases are highlighted in blue.

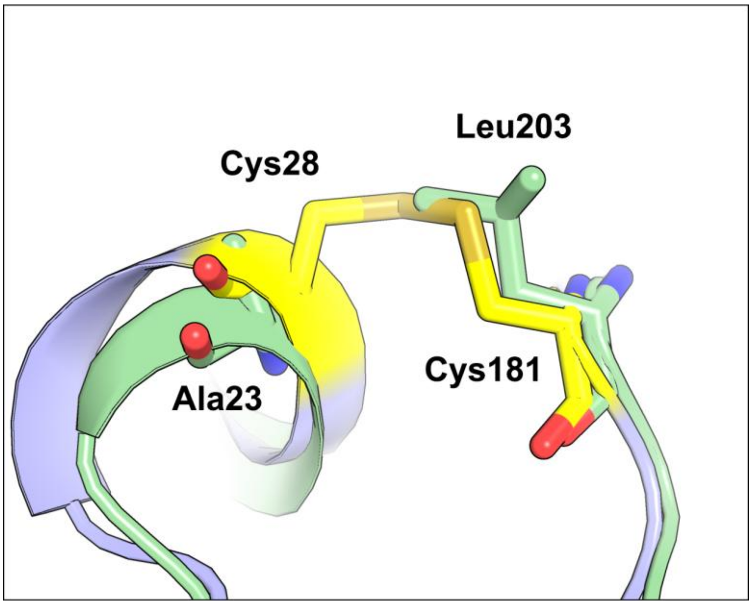


Fig. S4. 3D-superimposition of the human carbonic anhydrase II (PDB 2CBA, green) and carbonic anhydrase from *Neisseria gonorrhoeae* (PDB 1KOP, blue).

In a separate case-study the Yosshi+Mustguseal tool was used to study disulfide connectivity in the human carbonic anhydrase II (PDB 2CBA, chain A). The corresponding structure-guided sequence alignment constructed by the Mustguseal contained 5260 proteins. Further analysis of the alignment by Yosshi provided 40 pairs of hot-spots for disulfide engineering. The pair #1 corresponded to positions Ala23 and Leu203 both occupied by cysteines in 1377 proteins (26% of the set), in particular, S-S bond between the equivalent residues Cys28/Cys181 was found in the PDB structure 1KOP of a more stable carbonic anhydrase from *Neisseria gonorrhoeae* (Fig. S4, [26]). Introduction of this naturally occurring crosslink into the structure of human carbonic anhydrase II by a double mutation Ala23Cys/Leu203Cys was previously reported to increase the unfolding midpoint from 0.9 M guanidine HCl for the wild-type to 1.7 M guanidine HCl for the variant [27].

The homology-driven analysis of disulfide connectivity in xylanases was carried out by the Yosshi+Mustguseal tool. The PDB 1BCX (chain A) corresponding to the *Bacillus circulans* xylanase was used as a query to automatically construct the alignment of 3369 proteins which was further subjected to analysis by the Yosshi. Pairs ranked #2 and #6 corresponded to positions Ser100/Asn148 and Val98/Ala152 (Fig. S5). The corresponding double mutations Ser100Cys/Asn148Cys and Val98Cys/Ala152Cys formed S-S bonds between α-helix and adjacent β-strand in the protein core and were each shown to increase thermostability of the xylanase [28]. In a separate development, PDB 1XYP (chain A) of the *Trichoderma reesei* xylanase and the multiple alignment of 3181 proteins automatically created by the Mustguseal from that query were analyzed by the Yosshi. Introduction of an S-S bond by a double mutation at positions Ser110/Asn154 that were ranked #2 in the Yosshi’s output increased the protein half-life from less than 1 min to 14 min at 65°C [29]. It can be noted that positions Ser110/Asn154 in the *Trichoderma reesei* xylanase were equivalent to Ser100/Asn148 in the *Bacillus circulans* xylanase and occupied by an S-S bond between Cys110/Cys154 in the thermostable xylanase from *Thermomyces lanuginosus*, all proteins being members of the family GH11 (Fig. S5). We conclude, that introduction of a disulfide bond between the two respective regions (i.e., α-helix and adjacent β-strand, Fig. S5) in the protein core can be used to design robust variants of the family GH11 xylanases.

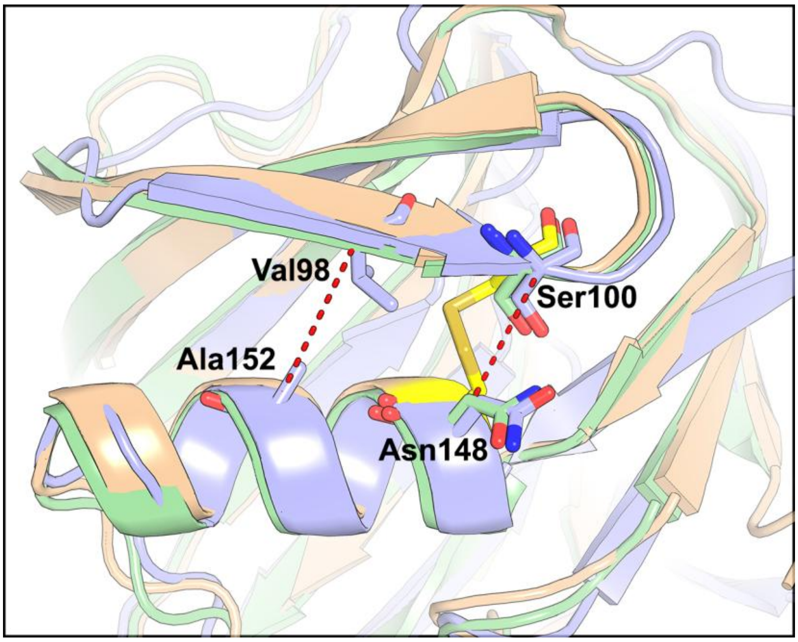


Fig. S5. 3D-superimposition of the family GH11 xylanases from *Bacillus circulans* (PDB 1BCX, blue), *Trichoderma reesei* (PDB 1XYP, green), and *Thermomyces lanuginosus* (PDB 1YNA, wheat). The two pairs of hotspots in α-helix and adjacent β-strand selected for disulfide engineering in 1BCX and 1XYP by the Yosshi are connected by dashed lines. The numbering is provided as in 1BCX.

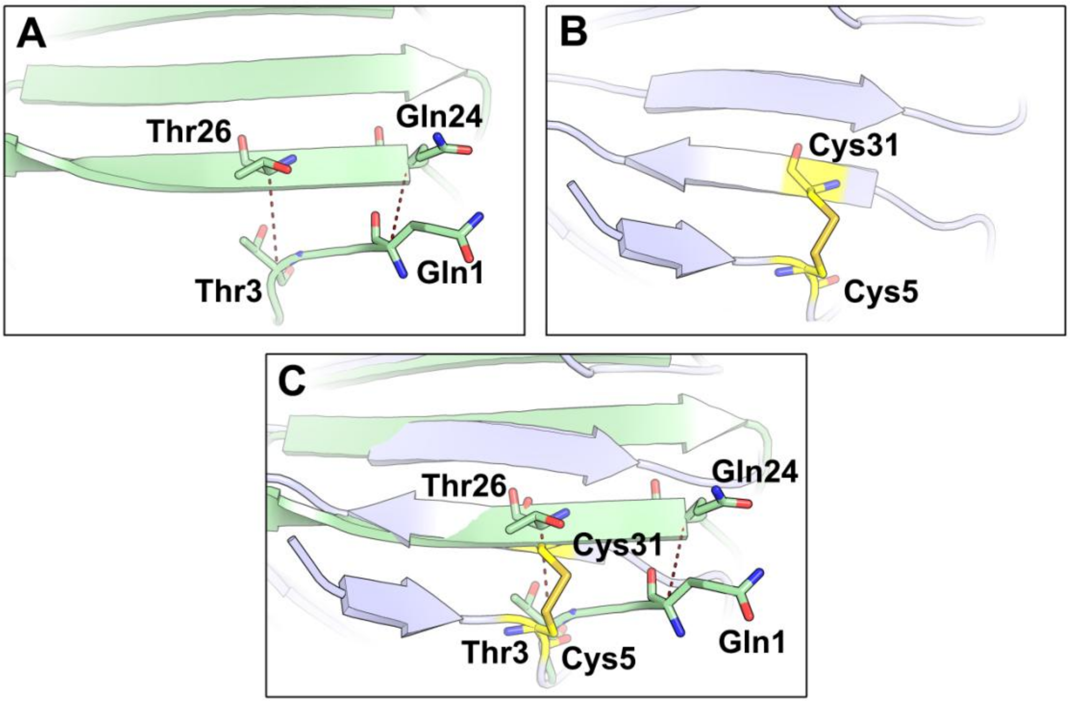


Fig. S6. Structurally equivalent regions of thermally stable xylanases from (A) *Thermomyces lanuginosus*, family GH11, PDB 1YNA (the side-chain of Gln1 was reconstructed using PyMol), and (B) *Streptomyces sp. 11AG8*, family GH12, PDB 1OA4;
(C) 3D-superimposition of the full-size structures of the two proteins.

Finally, disulfide connectivity between the N-terminal region and adjacent β-sheet in the protein core of xylanases was studied by the Yosshi. The PDB 1YNA (chain A) corresponding to the thermostable GH11 xylanase from *Thermomyces lanuginosus* discussed in the previous example was in turn used as a query to construct an alignment of 3388 proteins by the Mustguseal followed by the Yosshi analysis. The pair #1 corresponded to positions Thr3 and Thr26 both occupied by cysteines in 414 proteins (12% of the set; Fig. S6, A). Analysis of the output showed that the disulfide bond between equivalent positions is conserved in the family GH12 xylanases, and analysis of the literature concluded that it is very important for their stability (e.g., [30]). E.g., a crosslink between Cys5 and Cys31 anchors the N-terminus to the β-strand A2 in the thermostable GH12 xylanase from alkalophilic bacterium *Streptomyces sp. 11AG8* [31] (Fig. S6, B) although the N-terminal region in that protein has a slightly different orientation compared to that in the family GH11 xylanases (Fig. S6, A and C). Expert inspection of the PDB 1YNA previously helped to select positions Gln1 and Gln24 with a higher B-factor as hotspots for disulfide engineering of xylanase from *Thermomyces lanuginosus* (Fig. S6, A). The corresponding mutation Gln1Cys/Gln24Cys led to a significant stabilization of the variant (e.g., at pH 8 and 70 °C, the disulfide bridge increased the enzyme half-life 20-fold) [32]. Qualitatively similar disulfide bonds between the N-terminus and the adjacent β-strand were successfully used to stabilize other family GH11 xylanases [33, 34, 35]. We conclude that robustness of the thermostable family GH11 xylanases (e.g., the *Thermomyces lanuginosus* xylanase) can be additionally improved by introducing N-terminal disulfide bridges that are qualitatively similar to that observed in homologs from the family GH12. These results suggest that structural integrity of the N-terminus has a role in thermostability of xylanases. For the sake of clarity, in this case-study three multiple alignments were constructed by the Mustguseal from PDB queries of three GH11 xylanases, i.e., 1BCX, 1XYP, 1YNA (as discussed above). Pairwise sequence identity and structural similarity (quantified by the percentage of secondary structure equivalence [11]) between these PDB queries were within ranges 47-61% and 89-99%, respectively. Consequently, the alignments produced by the Mustguseal from these PDB queries were qualitatively similar, but had minor differences in the content and size (i.e., 3369, 3181, and 3388 proteins, respectively) due to the fact that slightly different homologs were automatically selected at the structure and sequence similarity search steps.

Popular 3D-structure based algorithms to improve protein stability by disulfide engineering – Disulfide by Design [36], MODIP [37], and SSBOND [38] – were tested on the same case-studies. These web-servers usually predicted a large number of hot-spots to construct S-S bonds but failed to select most of the discussed double mutations (Table 1, see the Main text). Similarly, Yosshi with the 3D-motif analysis performed in the “Rigid” mode failed to select some of the mutations which were correctly identified by the bioinformatic analysis. The observed poor performance of the 3D-structure based methods as well as Yosshi in the “Rigid” mode probably can be explained by the use of strict geometric models trained on covalently connected cysteines to evaluate the candidate positions. As previously discussed (see section 2 above), introduction of a crosslink into a protein structure can result in a considerable shift of the backbone atoms. Therefore, the most promising hot-spots do not always match the strict geometric constraints of an S-S bond. E.g., there was a considerable shift between the backbone atoms of Ala23/Leu203 in the human carbonic anhydrase II (PDB 2CBA) compared to an S-S bond between the equivalent residues Cys28/Cys181 in the homologous carbonic anhydrase from *Neisseria gonorrhoeae* (PDB 1KOP), i.e., the best-fit RMSD between backbone atoms was

0.595 Å (Fig. S4). The pair of positions Ala23/Leu203 in 2CBA was identified by the Yosshi's bioinformatic analysis and then confirmed as a promising hot-spot by the "Flexible" mode of 3D-motif analysis that attempts to take the backbone mobility into account (see subsection 2.3 above). The same pair of positions was rejected in the "Rigid" mode trained on covalently connected cysteines similar to Disulfide by Design, MODIP, and SSBOND (see "2CBA:A" in Table 1 in the Main text). To conclude, different performance of Yosshi in "Flexible" and "Rigid" 3D-motif analysis modes as well as relatively poor performance of other methods can be explained by backbone flexibility at the discussed positions.

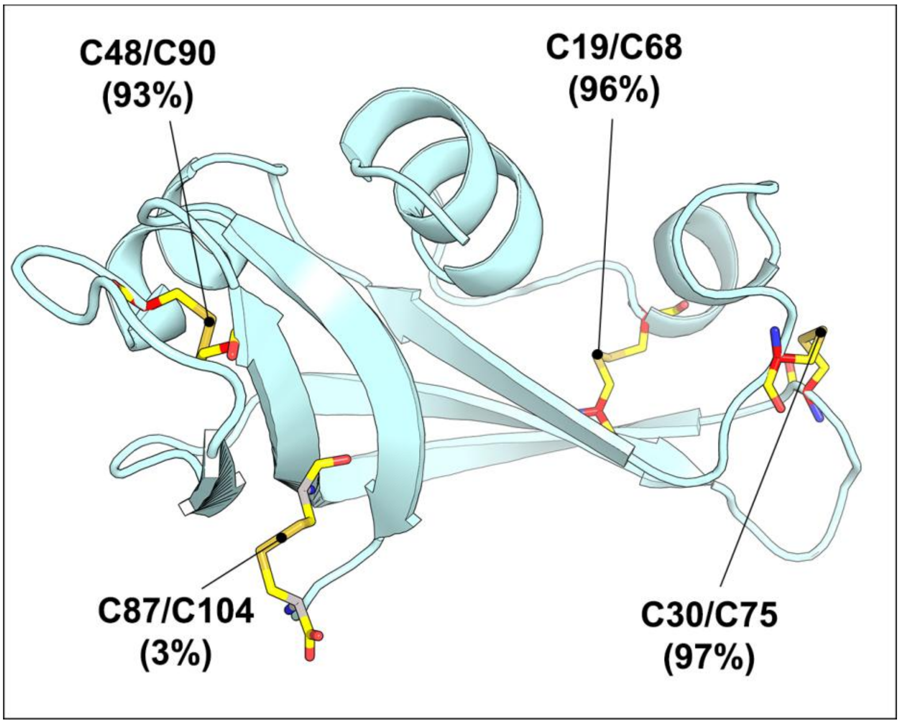


Fig. S7. Structure of onconase from *Lithobates pipiens* (PDB 1ONC). Four disulfide bonds present in the wild-type protein are shown. The occurrence of corresponding disulfides in homologs determined by the Yosshi analysis is shown in parentheses.

In the last case-study, disulfide connectivity in onconase from *Lithobates pipiens* and its homologs from the ribonuclease A (RNase A) superfamily was studied by the Yosshi+Mustguseal tool. Onconase is a protein with a great potential as tumor therapeutic [39] that contains four disulfide bonds in its structure (Fig. S7, PDB 1ONC). Three S-S bonds – i.e., Cys19/Cys68, Cys30/Cys75, and Cys48/Cys90 – were shown by the Yosshi to be highly conserved in homologs (93%-97% among 1033 proteins collected by the Mustguseal). The forth disulfide Cys87/Cys104 in onconase links the C-terminal region to the protein core and was observed only in 33 proteins (i.e., 3%), in particular, it was absent from the structure of the homologous bovine RNase A (PDB 7RSA). Such a conservation pattern suggests that the first three S-S bonds have a common role in the superfamily while Cys87/Cys104 can be responsible for specific properties of the onconase [40]. Removal of one of the three conserved S-S bonds by a double mutation Cys30Ala/Cys75Ala previously resulted in a considerable destabilization of the onconase [41]. Similarly, elimination of the conserved disulfide bonds from RNase A by Cys→Ala mutations decreased the T_m value of that protein by up to 38 °C outlining structural importance of these crosslinks in the common fold of a superfamily [42]. On the contrary, the role of the Cys87/Cys104 bond that was present only in the onconase was specific to this protein. Removal of the corresponding crosslink from onconase reduced both stability and activity of the variant and significantly affected its folding kinetics [43], while introduction of a qualitatively similar disulfide bond in the structure of a less stable RNase A by a double mutation His105Cys/Val124Cys had almost no effect on either stability or folding [44]. We conclude that Yosshi can be used not only to select hot-spots for disulfide engineering, but also can assist to study the role of native S-S bonds by a homology-driven comparative analysis of proteins with different functions and disulfide connectivity patterns.

REFERENCES

1. Pijning,A.E., Chiu,J., Yeo,R.X., Wong,J.W., and Hogg,P.J. (2018). Identification of allosteric disulfides from labile bonds in X-ray structures. *Royal Society Open Science*, 5(2), 171058.

2. Fu,L., Niu,B., Zhu,Z., Wu,S., and Li,W. (2012). CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics*, 28(23), 3150-3152.

3. Rubinstein,R., and Fiser,A. (2008). Predicting disulfide bond connectivity in proteins by correlated mutations analysis. *Bioinformatics*, 24(4), 498-504.

4. Kirilin,E.M., and Švedas,V.K. (2018). Study of the Conformational Variety of the Oligosaccharide Substrates of Neuraminidases from Pathogens using Molecular Modeling. *Moscow University Chemistry Bulletin*, 73(1), 39-45.

5. Costa,S.I., Torezzan,C., Campello,A., and Vaishampayan,V.A. (2013). Flat tori, lattices and spherical codes. In *Information Theory and Applications Workshop*, San Diego, CA.

6. Ester,M., Kriegel,H.P., Sander,J., and Xu,X. (1996). A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd*, 96(34), 226-231.

7. Rousseeuw,P.J. (1987). Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, 20, 53-65.

8. Dombkowski,A.A., Sultana,K.Z., and Craig,D.B. (2014). Protein disulfide engineering. *FEBS Letters*, *588*(2), 206-212.

9. Pellequer,J.L., and Chen,S.W.W. (2006). Multi-template approach to modeling engineered disulfide bonds. *Proteins*, *65*(1), 192-202.

10. Sillitoe, I., Lewis, T. E., Cuff, A., Das, S., Ashford, P., Dawson, N. L., ... & Lehtinen, S. (2014). CATH: comprehensive structural and functional annotations for genome sequences. *Nucleic Acids Research*, 43(D1), D376-D381.

11. Krissinel,E., and Henrick,K. (2004). Secondary-structure matching (SSM), a new tool for fast protein structure alignment in three dimensions. *Acta Crystallographica Section D*, 60(12-1), 2256-2268.

12. Konagurthu,A.S., Whisstock,J.C., Stuckey,P.J., and Lesk,A.M. (2006). MUSTANG: a multiple structural alignment algorithm. *Proteins: Structure, Function, and Bioinformatics*,

64(3), 559-574.

13. Menke,M., Berger,B., and Cowen,L. (2008). Matt: local flexibility aids protein multiple structure alignment. *PLoS Computational Biology*, 4(1), e10.

14. Dong,R., Peng,Z., Zhang,Y., and Yang,J. (2017). mTM-align: an algorithm for fast and accurate multiple protein structure alignment. *Bioinformatics*, 34(10), 1719-1725.

15. Kalaimathy,S., Sowdhamini,R., and Kanagarajadurai,K. (2011). Critical assessment of structure-based sequence alignment methods at distant relationships. *Briefings in Bioinformatics*, 12(2), 163-175.

16. Frishman, D., & Argos, P. (1995). Knowledge-based protein secondary structure assignment. *Proteins: Structure, Function, and Bioinformatics*, 23(4), 566-579.

17. Wieteska,L., Ionov,M., Szemraj,J., Feller,C., Kolinski,A., and Gront,D. (2015). Improving thermal stability of thermophilic l-threonine aldolase from Thermotoga maritima. *Journal of Biotechnology*, 199, 69-76.

18. Sowdhamini,R., Srinivasan,N., Shoichet,B., Santi,D.V., Ramakrishnan,C., and Balaram,P. (1989). Stereochemical modeling of disulfide bridges. Criteria for introduction into proteins by site-directed mutagenesis. *Protein Engineering, Design and Selection*, 3(2), 95-103.

19. Kanaya,S., Katsuda,C., Kimura,S., Nakai,T., Kitakuni,E., Nakamura,H., Katayanagi,K., Morikawa,K. and Ikehara,M. (1991). Stabilization of Escherichia coli ribonuclease H by introduction of an artificial disulfide bond. *Journal of Biological Chemistry*, 266(10), 6038-6044

20. Perry, L. J., & Wetzel, R. (1984). Disulfide bond engineered into T4 lysozyme: stabilization of the protein toward thermal inactivation. *Science*, 226(4674), 555-557.

21. Andreu,D., Albericio,F., Solé,N.A., Munson,M.C., Ferrer,M., and Barany,G. (1994). Formation of disulfide bonds in synthetic peptides and proteins. In: Pennington M.W., Dunn B.M. (eds) *Peptide Synthesis Protocols. Methods in Molecular Biology*, vol 35. Humana Press, Totowa, NJ, pp. 91-169.

22. Dombkowski,A.A. (2003). Disulfide by Design: a computational method for the rational design of disulfide bonds in proteins. *Bioinformatics*, 19(14), 1852-1853.

23. Agresti,A., and Kateri,M. (2011). Categorical data analysis. In: Lovric M. (ed.) International Encyclopedia of Statistical Science, Springer, Berlin Heidelberg, pp. 206-208.

24. Korman,T.P., Sahachartsiri,B., Charbonneau,D.M., Huang,G.L., Beauregard,M., and Bowie,J.U. (2013). Dieselzymes: development of a stable and methanol tolerant lipase for biodiesel production by directed evolution. *Biotechnology for Biofuels*, 6(1), 70

25. Yamaguchi,S., Takeuchi,K., Mase,T., Oikawa,K., McMullen,T., Derewenda,U., McElhaney,R.N., Kay,C.M., and Derewenda,Z.S. (1996). The consequences of engineering an extra disulfide bond in the Penicillium camembertii mono-and diglyceride specific lipase. *Protein Engineering, Design and Selection*, 9(9), 789-795

26. Huang,S., Xue,Y., Sauer-Eriksson,E., Chirica,L., Lindskog,S., and Jonsson,B.H. (1998). Crystal structure of carbonic anhydrase from Neisseria gonorrhoeae and its complex with the inhibitor acetazolamide. *Journal of Molecular Biology*, 283(1), 301-310.

27. Mårtensson,L.G., Karlsson,M., and Carlsson,U. (2002). Dramatic stabilization of the native state of human carbonic anhydrase II by an engineered disulfide bond. *Biochemistry*, 41(52), 15867-15875.

28. Wakarchuk,W.W., Sung,W.L., Campbell,R.L., Cunningham,A., Watson,D.C., and Yaguchi,M. (1994). Thermostabilization of the Bacillus circulans xylanase by the introduction of disulfide bonds. *Protein Engineering, Design and Selection*, 7(11), 1379-1386.

29. Turunen,O., Etuaho,K., Fenel,F., Vehmaanperä,J., Wu,X., Rouvinen,J., and Leisola,M. (2001). A combination of weakly stabilizing mutations with a disulfide bridge in the α-helix region of Trichoderma reesei endo-1, 4-β-xylanase II increases the thermal stability through synergism. *Journal of Biotechnology*, 88(1), 37-46.

30. Sandgren,M., Ståhlberg,J., and Mitchinson,C. (2005). Structural and biochemical studies of GH family 12 cellulases: improved thermal stability, and ligand complexes. *Progress in Biophysics and Molecular Biology*, 89(3), 246-291.

31. Sandgren,M., Gualfetti,P.J., Shaw,A., Gross,L.S., Saldajeno,M., Day,A.G., Jones,T.A., and Mitchinson,C. (2003). Comparison of family 12 glycoside hydrolases and recruited substitutions important for thermal stability. *Protein Science*, 12(4), 848-860.

32. Wang,Y., Fu,Z., Huang,H., Zhang,H., Yao,B., Xiong,H., and Turunen,O. (2012). Improved thermal performance of Thermomyces lanuginosus GH11 xylanase by engineering of an N-terminal disulfide bridge. *Bioresource Technology*, 112, 275-279.

33. Li,H., Kankaanpää,A., Xiong,H., Hummel,M., Sixta,H., Ojamo,H., and Turunen,O. (2013). Thermostabilization of extremophilic Dictyoglomus thermophilum GH11 xylanase by an N-terminal disulfide bridge and the effect of ionic liquid [emim] OAc on the enzymatic performance. *Enzyme and Microbial Technology*, 53(6-7), 414-419.

34. Paës, G., & O'Donohue, M. J. (2006). Engineering increased thermostability in the thermostable GH-11 xylanase from Thermobacillus xylanilyticus. *Journal of Biotechnology*, 125(3), 338-350.

35. Xiong,H., Fenel,F., Leisola,M., and Turunen,O. (2004). Engineering the thermostability of Trichoderma reesei endo-1, 4-β-xylanase II by combination of disulphide bridges. *Extremophiles*, 8(5), 393-400.

36. Craig,D.B., and Dombkowski,A.A. (2013). Disulfide by Design 2.0: a web-based tool for disulfide engineering in proteins. *BMC Bioinformatics, 14*(1), 346.

37. Dani,V.S., Ramakrishnan,C., and Varadarajan,R. (2003). MODIP revisited: re-evaluation and refinement of an automated procedure for modeling of disulfide bonds in proteins. *Protein Engineering*, 16(3), 187-193

38. Hazes,B., and Dijkstra,B.W. (1988). Model building of disulfide bonds in proteins with known three-dimensional structure. *Protein Engineering, Design and Selection*, 2(2), 119-125.

39. Schulenburg,C., Ardelt,B., Ardelt,W., Arnold,U., Shogen,K., Ulbrich-Hofmann,R., and Darzynkiewicz,Z. (2007). The interdependence between catalytic activity, conformational stability and cytotoxicity of onconase. *Cancer Biology & Therapy*, 6(8), 1244-1250.

40. Arnold,U. (2014). Stability and folding of amphibian ribonuclease A superfamily members in comparison with mammalian homologues. *The FEBS Journal*, 281(16), 3559-3575.

41. Torrent,G., Benito,A., Castro,J., Ribo,M., and Vilanova,M. (2008). Contribution of the C30/C75 disulfide bond to the biological properties of onconase. *Biological Chemistry*, 389(8), 1127-1136.

42. Klink,T.A., Woycechowsky,K.J., Taylor,K.M., and Raines,R.T. (2000). Contribution of disulfide bonds to the conformational stability and catalytic activity of ribonuclease A. *European Journal of Biochemistry*, 267(2), 566-572.

43. Schulenburg,C., Weininger,U., Neumann,P., Meiselbach,H., Stubbs,M.T., Sticht,H., Balbach,J., Ulbrich-Hofmann,R., and Arnold,U. (2010). Impact of the C-terminal Disulfide Bond on the Folding and Stability of Onconase. *ChemBioChem*, 11(7), 978-986.

44. Pecher,P., and Arnold,U. (2009). The effect of additional disulfide bonds on the stability and folding of ribonuclease A. *Biophysical Chemistry*, 141(1), 21-28.