Native and Modeled Disulfide Bonds in Proteins: Knowledge-Based Approaches Toward Structure Prediction of Disulfide-Rich Polypeptides

Ratna Rajesh Thangudu,[1] A. Vinayagam,[2] G. Pugalenthi,[2] A. Manonmani,[2] B. Offmann,[1] and R. Sowdhamini[2]*
[1]*Laboratoire de Biochimie et Génétique Moléculaire, Université de La Réunion, La Réunion, France*
[2]*National Centre for Biological Sciences (TIFR), UAS-GKVK Campus, Bangalore, India*

ABSTRACT **Structure prediction and three-dimensional modeling of disulfide-rich systems are challenging due to the limited number of such folds in the structural databank. We exploit the stereochemical compatibility of substructures in known protein structures to accomodate disulfide bonds in predicting the structures of disulfide-rich polypeptides directly from disulfide connectivity pattern and amino acid sequence in the absence of structural homologs and any other structural information. This knowledge-based approach is illustrated using structure prediction of 40 nonredundant bioactive disulfide-rich polypeptides such as toxins, growth factors, and endothelins available in the structural databank. The polypeptide conformation could be predicted in 35 out of 40 nonredundant entries (87%). Nonhomologous templates could be identified and models could be obtained within 2 Å deviation from the query in 29 peptides (72%). This procedure can be accessed from the World Wide Web (http://www.ncbs.res.in/~faculty/mini/dsdbase/dsdbase.html). Proteins 2005;58:866–879.**

Key words: SS bonds; covalent crosslinks; fold recognition; bioactive peptides; DSDBASE; disulfide database

INTRODUCTION

Disulfide bonds are covalent crosslinks that connect different parts of a protein. Due to the strong reducing environment in the interior of the cell, disulfides are associated mainly with extracellular proteins. Rigid covalent crosslinks are a convenient means of limiting the mobility of a polypeptide and thereby stabilizing the native folded conformations. It is estimated that the addition of one disulfide bond to a protein[1] can impart stabilities in the order of 2–5 kcal mol^{-1} and the stabilization increases with increase in loop size.[2] Usually, better results in enhancement of thermal stability are expected due to the introduction of new disulfide bonds of large loop sizes in a protein or crosslinks that connect protomers by site-directed mutagenesis. Several site-directed mutagenesis experiments have been designed to introduce extra disulfides with the intention of enhancing thermal stability of proteins[3–8] or across protomers[9,10] with varying extents of success. The introduction of covalent crosslinks is believed to entropically destabilize the unfolded state of a polypeptide thereby increasing the net stability of the folded form of a protein[3] as well as stabilize the native state.[11]

Protein folds containing disulfide bonds have been analyzed for preferred values of internal parameters at the crosslink.[12–14] Different computer algorithms have been designed to choose sites for strainless disulfide bond introduction.[15–17] MODIP[17] is an algorithm by which one can examine a given protein structure for all possible pairs of residues for the local stereochemical compatibility to accommodate a disulfide bridge.

The inherent conformational restraints in disulfide-rich polypeptides render them as excellent systems for knowledge-based modeling. Several small, disulfide-rich systems are bioactive peptides such as vasoconstrictors, antibacterial peptides and toxins. Despite their biomedical importance, the three-dimensional structures are not available for many such peptides. In some instances, conformational flexibility at relatively flexible parts of the molecules allows multiple conformations for the polypeptide that could be biologically relevant. Most disulfide-rich, small polypeptides contain very few secondary structures. The main stabilizing interactions are the covalent crosslinks that serve as convenient handles or constraints to explore the polypeptide conformational space. For small polypeptides, ab initio backbone conformational search with a fixed grid for altering backbone conformations to retain the observed covalent crosslink constraint(s) is feasible (for example, Hruby and coworkers[18]). Poland and Scheraga[1] had sampled backbone conformational space for an

Abbreviations: Å, Angstrom; DSDBASE, Disulfide DataBASE; PDB, Protein Data Bank; NMR, Nuclear Magnetic Resonance; SCF, sulfur could not be fixed; RMSD, root-mean-square deviation; MODIP, Modeling of Disulfide bonds in Proteins; SS, Disulfide bond; MULSS, Multiple disulfide bond search.

The Supplementary Material referred to in this article can be found at http://www.ncbs.res.in/~faculty/mini/dsdbase/suppl.htm

A. Vinayagam's current address is Department of Molecular Biophysics, German Cancer Research Centre (DKFZ), Im Neuenheimer Feld 280, 69120, Heidelberg, Germany

*Correspondence to: R. Sowdhamini, National Centre for Biological Sciences (TIFR), UAS-GKVK Campus, Bellary Road, Bangalore 560 065, India. E-mail: mini@ncbs.res.in

Received 22 July 2004; Accepted 16 September 2004

Published online 11 January 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20369

octapeptide SS loop in ribonuclease A. Stereochemically allowed, low energy conformations were assigned to the backbone of oxytocin to explore multiple backbone conformations that will permit facile disulfide bond closure.[19] Random, but stereochemically allowed backbone conformations were sampled for strainless disulfide bridge closure.[20] Molecular dynamics calculations have been attempted for enkephalins[21] and melanin-concentrating hormone.[22] However, for larger polypeptides such as toxins and growth factors, knowledge-based approaches have been attempted for structure prediction.[20,23]

Knowledge-based approaches for modeling disulfide-rich polypeptides are attractive since they are derived from protein structural databases and offer multiple conformations as possible solutions. The fragment-based structural modeling approach, originally attempted for refinement and model-building in crystallography,[24] could be employed to search for the query disulfide-bond connectivity pattern in the disulfide database. We have recently reported DSDBASE[25] where the size of the disulfide database could be enhanced substantially by including native disulfide bonds as well as those stereochemically feasible among pairs of residues in all protein structural entries. This has been previously employed, using a smaller dataset, in an attempt to predict the structure of enterotoxin ST1b[26] and the N-terminal domain of hepatocyte growth factor.[27] In this paper, we examine the quality of the disulfide bonds in the database and demonstrate its application to the knowledge-based modeling of disulfide-rich polypeptides. We have analysed the stereochemical features of MODIP-suggested disulfide crosslinks for all the annotated native disulfides in DSDBASE and discuss the cases of false positives and false negatives. We have tested the suitability of the database for structural templates by comparing substructures with similar loop size and disulfide connectivity. We then suggest an approach to the structure prediction of disulfide-rich polypeptides starting from amino acid sequence information and connectivity in the absence of structural homologs. The database could be successfully queried to propose models of several disulfide-rich polypeptides of known structure using non-homologous substructures. The procedure can be accessed from the World Wide Web (http://www.ncbs.res.in/~faculty/mini/dsdbase/dsdbase.html).

MATERIALS AND METHODS

Protein Database

Nineteen thousand six hundred and twelve (19,612) proteins (corresponding to the April 2003 release of Brookhaven Protein Data Bank[28,29]) have been examined by MODIP for substructures that can accommodate disulfide bonds giving rise to 2,385,617 substructures that can potentially accomodate disulfide bonds. Three forms of disulfide database resources are available: corresponding to full PDB database (*fulldb*) or the non-redundant PDB database (*nrdb*) or a subset of native disulfides of the nonredundant set of proteins(*nr-native*). *nrdb* is derived from a nonredundant list of 2849 protein chains where no two entries are related more than 25% sequence identity.[30]

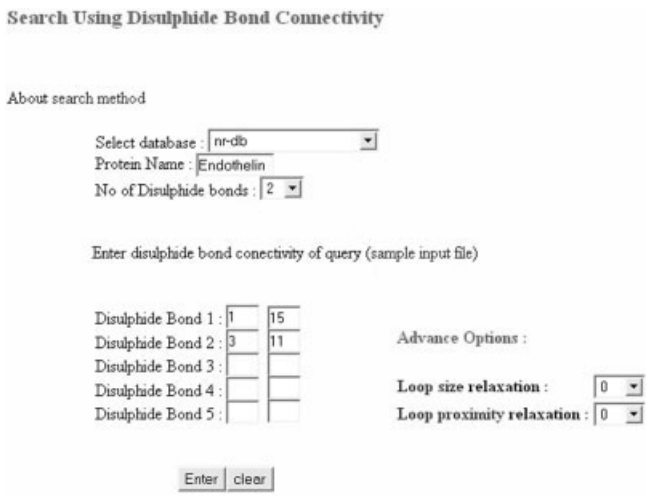

Fig. 1. Sample input parameters for the search procedure for query into the disulfide database. The primary structure and disulfide bond connectivity of the query disulfide-rich polypeptide are the required inputs. Option for user-defined relaxation of loop size and spacing between consecutive Cys residues is provided.

C^{α} proteins and other problematic cases were excluded and the rest employed for the creation of the disulfide database.[25]

Disulfide Bond Modeling

In the original MODIP program, residue pairs with C^{α}–C^{α} distances $\leq$ 6.5 Å and C^{β}–C^{β} distances $\leq$ 4.5 Å were chosen for geometric fixing of sulphur atoms and to assess the stereochemical quality of the modeled crosslink. For the disulfide database, however, we chose a broader range with a cut-off of 7 Å for C^{α}–C^{α} distance and 4.7 Å for C^{β}–C^{β} distance. For residue pairs of permissible distances, sulphur was fixed using the coordinates of N, C^{α} and C^{β} where C^{α}–C^{β}–S, C^{β}–S–S and C^{β}–S are preferred values of 114°, 104° and 1.87 Å, respectively.

Stereochemical Quality of Modeled Disulfides

The stereochemical quality of the disulfide bridge is assessed using criteria as described before[17,20] and guided by the analysis of protein disulfides.[13,14] Disulfide loop size is defined by the number of residues within the loop of the polypeptide chain including the two Cys residues.

Searching the Database for Desired Disulfide Connectivity

The disulfide database contains information about protein chains of all possible substructures that can successfully accommodate a disulfide bond by the MODIP method[17]: PDB code and chain identifier, loop size, end-to-end residues defining the loop (Fig. 1) and the stereochemical grade. This database can be queried over the Web using MULSS, a search procedure, for a particular disulfide loop or multiple disulfide bonds of known connectivity. The required inputs for searching the database are the total number of disulfide bonds and their connectivity. No structural information, other than SS connectivity was

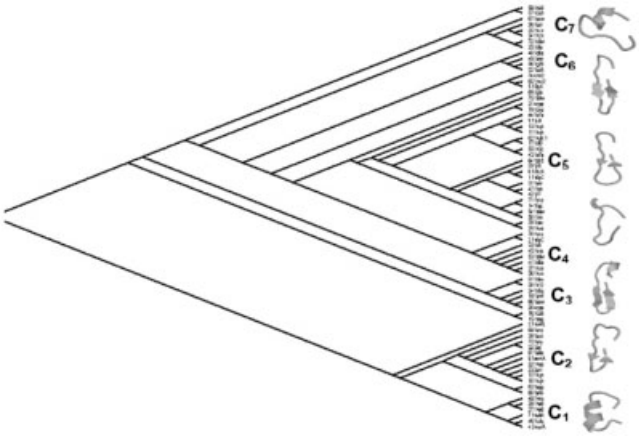

Fig. 2. Modelled disulfide loops: Size and conformational variations. Clustering of the substructures from various proteins which can accommodate a 17-residue disulfide loop on the basis of the backbone conformational variations. MNYFIT[32] has been used for rigid-body superposition of the loops. The major clusters have been marked (C_1 to C_7). Representative 17-residue disulfide loops after best-fit by rigid-body superposition shown next to the cluster. Backbone conformational variations are expected to be high for long disulfide loops, but a majority of them (C_2 to C_6), including ones in the most populated cluster (C_5), contain two extended strands in the backbone.

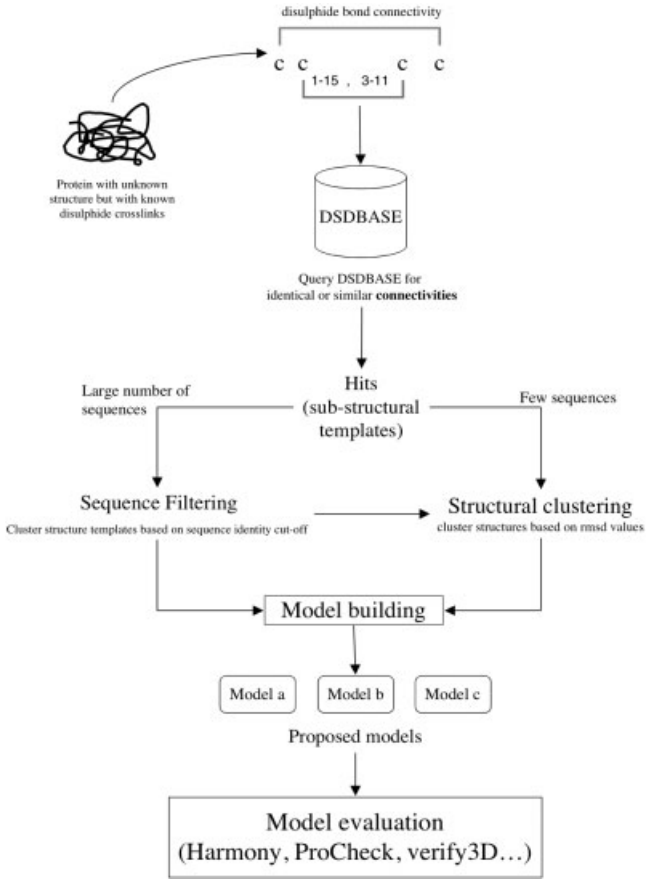

Fig. 3. Schematic representation of the procedure for modeling of disulfide-rich polypeptides by querying DSDBASE and subsequent analysis of templates and validation of models.

provided as inputs to the method. An user-defined relaxation is permitted for both loop size and the spacing to accommodate loop size variations and insertions/deletions at the polypeptide backbone (Fig. 1). Substructures from different proteins that satisfy the query disulfide bond connectivity are projected as outputs. The searches can be performed against any one of the three forms of DSDBASE: corresponding to full PDB database (*fulldb*) or the nonredundant PDB database (*nrdb*) or a subset of native disulfides of the nonredundant set of proteins (*nr-native*). It is possible to jack-knife and restrict the number of hits by obtaining a distinct set of substructural hits. The identity cut-off is user-defined and is especially useful while searching against the disulfide database corresponding to the full PDB release.

Choice of Disulfide-Rich Peptides From Protein Data Bank for the Benchmarking

Our dataset includes two, three, and four disulfide-bonded systems selected from small protein class of SCOP database.[31] The connectivity patterns of these benchmarking peptides are highly represented in the known protein structures. One hundred and nineteen (119) protein chains in the nonredundant database are small (less than 50 residues long) and 15 polypeptides in this nonredundant database contain two disulfide bonds and have been considered for benchmarking. No prior knowledge of the structure, such as the presence and position of the secondary structures, are considered during these searches. Searches were performed against the types of disulfide databases: native disulfides (*nr-native*), substructures identified from a set of nonredundant protein structural entries (*nr-db*), and all structural entries (*fulldb*). However in the current study, all the searches were performed against *fulldb* and the efficacy of using such a database is discussed elsewhere. Loop size relaxation and Cys spacing relaxations were provided only for the search against nr-db database. The loop size and Cys spacing relaxations were modified to obtain at least ten substructures as templates for modeling and these values were different for each query depending on the nature of the system. In the case of three- and four-disulfide-bonded systems, SCOP[31] representatives from each of the families were chosen for benchmarking leading to 17 and 18 peptides, respectively. For these higher order systems, loop size and proximity relaxations were provided in the searches.

Comparison of Templates

In case of searches against the full protein database (*fulldb*) for each of the peptide connectivity as a query, in order to avoid redundancies in the hits obtained, the substructures are compared at two levels: A sequence similarity filter is first placed at a 90% identity cut-off whereby no two proteins that are more than 90% identical in sequence are considered in the derived nonredundant dataset. This option is also available along with the search procedure over the Web for any query submission and is useful for searches against *fulldb*.

TABLE I. Comparison of the Stereochemical Grades of Native Disulfides Between MODIP Modeled Crosslink and that of Experimental Data

		Grades of modelled disulfides				
		A	B	C	Sulfur could not be fixed	Total
Grades of native disulfides	A	18124	1211	1317	604	21256
	B	454	1208	249	170	2081
	C	528	343	875	145	1891
	Total	19106	2762	2441	919	25228

In case of hits where the benchmarking peptides were provided as queries, an additional comparison was performed that allows substructures with similar folds to be clustered together: Pairwise rigid-body superimposition was performed without iterations using MNYFIT,[32] where the modelled cystine residues were provided as initial equivalencies. The root-mean-square deviations (RMSD) between the superposed coordinates at all C^{α} positions, within the substructures, were employed as descriptors of structural dissimilarity indices and to cluster the substructures using PHYLIP.[33] Representatives from individual and well-populated clusters can be considered for a detailed examination of their backbone conformation and for homology modeling. Models that deviate substantially from the crystal/NMR structure can be viewed as plausible alternate conformations of the polypeptide.

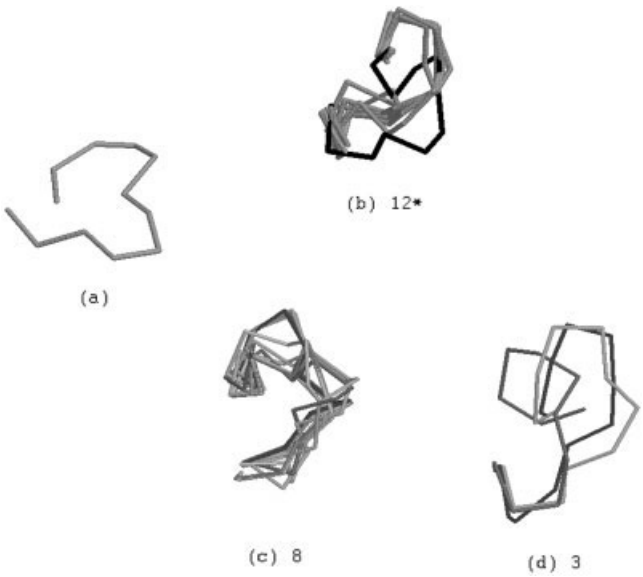

Fig. 4. Search for the disulfide bond connectivity of alpha-conotoxin GI and subsequent three-dimensional modeling. Only 23 nonredundant substructures from *fulldb* have been considered after a 90% sequence identity filter. **a:** C^{α}-trace of the experimental conformation of conotoxin (PDB code 1xga). **b–d:** C^{α}-traces of clusters of substructures obtained by searching the disulfide database. Clustering of the substructures was performed on the basis of structural dissimilarities measured as root-mean-square deviations (RMSD) using PHYLIP.[33] The number of members in the individual clusters has been denoted. The cluster that includes the experimental conformation of the query is shown in "*". The highest populated cluster (marked as "b") contains the unrelated substructure (ribosomal S8, PDB code 1hnz) that acquires the lowest RMSD with the experimental conformation of conotoxin. High similarity between this cluster and the crystal conformation is evident.

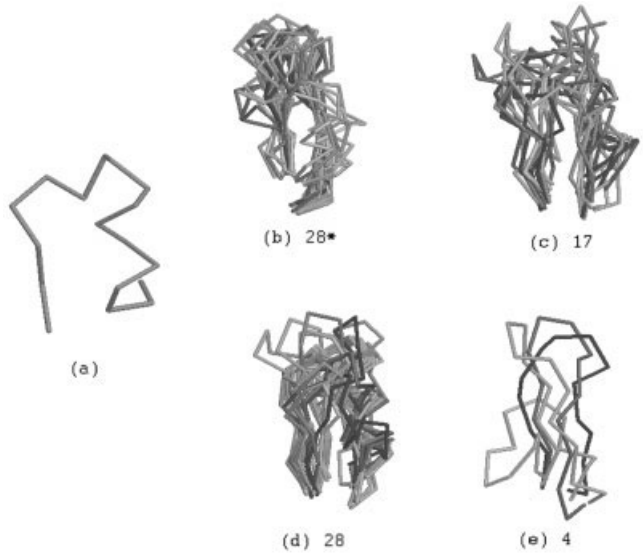

Fig. 5. Search in the disulfide database for possible conformations of endothelin (PDB code, 1edp). Seventy-seven nonredundant entries (see Table III) can be grouped into four major clusters (**b–e**). The cluster that includes the experimental conformation of the query is shown in "*". Each of the clusters have at least one member that is closely similar to the NMR conformation of endothelin (**a**) with an RMSD value less than 1 Å.

Three-Dimensional Modeling and Examination of Models

Representative substructures from various protein folds that are compatible to the query disulfide bond connectiv-

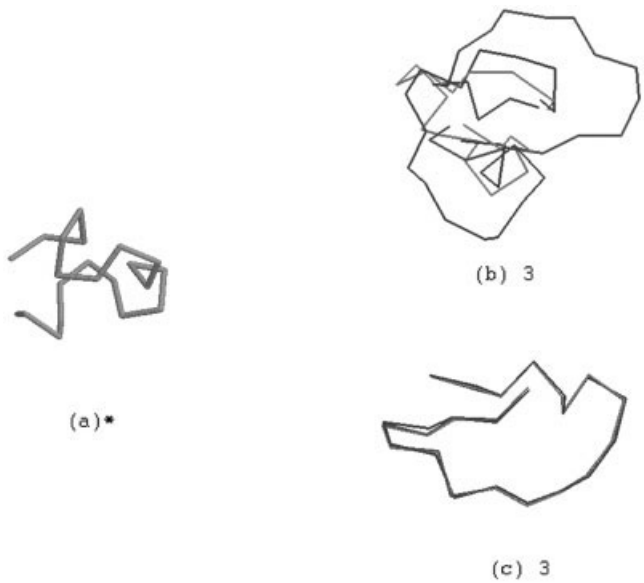

Fig. 6. Results obtained by searching the disulfide database for Shk toxin. NMR conformation (**a**) (PDB code 1c2u) compared to those of the four nonredundant substructures found in different proteins (**b–c**). None of the substructures have low RMSD to 1c2u but can be viewed as plausible templates for modeling alternate conformations.

TABLE II. Substructural Hits for 15 Two-disulfide Bonded Peptides Used for Benchmarking Against Three Different Databases

Protein and PDB code	Length	Disulfide-bonded residues		Number of hits		
				nr-native[1]	nr-db[2]	Fulldb[3]
α-Conotoxin GI, 1XGA	14	2–7	3–13	0	10	161
Charybdotoxin, 1BAH	37	7–28	17–35	2	7	21
Guanylin, 1GNA	13	88–96	91–99	0	10	17
Shk toxin, 1C2U	35	12–28	17–32	0	12	7
Enterotoxin (ST), 1ETL	13	6–14	9–17	0	10	17
C-Terminal domain of Cellobiohydrolase 1, 1CBH	36	8–25	19–35	1	11	42
RP 71955-Hiv replication Inhibitor, 1RPB	21	1–13	7–19	0	13	24
Uroguanylin, 1UYA	16	4–12	7–15	0	10	17
C-Terminal domain of midkine, 1MKC	43	62–94	72–104	1	9	27
Collagen-binding type II domain, 1PDC	45	1–27	15–42	1	10	46
Fibrin binding finger domain of tissue-type plasminogen activator, 1TPN	50	6–36	34–43	0	17	17
Tertiapin, 1TER	22	3–14	5–18	0	17	19
Endothelin, 1EDP	17	1–15	3–11	0	12	248
Immunodominant region of protein BRSV-G, 1BRV	32	173–186	176–182	1	20	561
Enterotoxin (STb), 1EHS	48	10–48	21–36	1	12	27

[1]In all cases, the search has been performed against the database of native disulfide bonds of the nonredundant set (*native-nrdb,* see Methods) of proteins from PDB.
[2]In all cases, the search has been performed against the database of derived from nonredundant set (*nrdb,* see Methods) of proteins from PDB. Note that loop size and proximity relaxation was applied, where needed, to get reasonable number of hits.
[3]In all cases, the search has been performed against the database derived from full protein database (*fulldb*).

ity are used as templates for obtaining three-dimensional models of the polypeptide. Modeling can be performed by simple side-chain replacement of the template structures as per the query sequence using SCRWL procedure.[34]

Model Validation

The best model is recognized by good PROCHECK values.[35] Further, the compatability of the ascribed fold to the entire sequence of the query is examined by scoring the amino acid preference of individual residues in the polypeptide to its local environment. This procedure, encoded as HARMONY, is similar in its approach to VERIFY3D.[36] Local environment of a residue is described by its backbone conformation (nine states), solvent accessibility patterns (three states) and hydrogen-bonding (two states) leading to 54 types. Amino acid frequencies in all possible local environment types are counted by examining a large number of nonredundant protein structures and normalized propensities are calculated from their raw occurrences as described in Wako and Blundell.[37] HARMONY score of a whole protein model is the sum of the scores at individual residues of the polypeptide. Individual residue scores are simply the normalized propensity of that particular residue to the observed local environment type. HARMONY scores are directly correlated to protein size. Misfolds are recognized by low HARMONY scores which when projected on an X-Y plot with scores in the Y-axis and protein length in the X-axis fall well below the fitted straight line. Such a straight-line fit passes through the origin and the slope of the fitted straight line (m(i)) after examining 4020 nonredundant protein structures is termed as the ideal normalized HARMONY scores.

RESULTS AND DISCUSSION

Modeling Disulfide Bonds: Benchmark Using Native Disulfides

A total of 25,620 annotated native disulfides (with a SSBOND record in PDB file) exist in the full protein dataset considered for the construction of DSDBASE.[25] Out of these, 25,239 annotated crosslinks (98.5%) could be identified using MODIP indicating that the algorithm has a high probability of accurately identifying substructures that can accommodate SS bonds. A majority of the annotated native SS bonds in the full dataset were modeled with good stereochemistry.[25]

Conformation Sampling of Substructures With Similar Loop Sizes: Example of 17-Residue Loops

Substructures with one disulfide of a particular loop size or which possess similar disulfide connectivity patterns can be structurally similar and the conformational variability at the polypeptide backbone is expected to increase with loop size. We tested our sampling procedure by selecting 17-residue loop substructures from DSDBASE. Seventy-five such native disulfide loops are present in the nonredundant disulfide database. For instance, one 17-residue loop is found in Shk toxin and forms the outer loop of the two-disulfide-bonded system. Rigid-body superposition and subsequent clustering of the 75 substructures (see Materials and Methods for details) reveal two prominent clusters: one consisting of substructures that contain two consecutive β-strands and another containing an α-helix (Fig. 2). Such a tight clustering in the backbone of such loops suggests that a simple rule of "majority-wins" could serve as a preliminary validation scheme in selecting

TABLE III. Summary of Benchmarking the Disulfide Loop Search Procedure for Two Disulphide Bonded Systems

Protein name	Number of hits on Fulldb[a]	Number of sub-structural clusters[b]	Population of clusters[c]	RMSD (Å) of homologs[d]	Template with lowest RMSD		
					PDB code and name[e]	RMSD[f] (Å)	Sequence Identity[g]
α-Conotoxin GI, 1XGA	26	4	12[h,i],10[i],3,1	0.54 (1)	1HNW (66T–77T) Ribosomal protein S18	0.50	0
Charybdotoxin, 1BAH	11	3	1,2,10[h,i]	0.33–1.09 (8)	1QUZ (3A–31A) Potassium channel toxin hstx1	0.75	37.9
Guanylin, 1GNA	6	4	1,2,1,2[h,i]	0.82–1.08 (2)	1DNP(384A–395A) C-terminal domain of DNA Phtolyase	1.28	0
Shk Toxin, 1C2U	6	2	3,3	0 (1)	1AOL (88–106) F-MuLV receptor-binding domain	3.32	0
Enterotoxin (ST), 1ETL	6	4	1,2,1,2[h,i]	0.82–1.21 (2)	1DNP (384A–395A) C-terminal domain of DNA Phtolyase	1.61	0
C-Terminal domain of cellobiohydrolase 1, 1CBH	13	4	1,1,7,3	0.71 (1)	1AEL (78–105) Intestinal fatty acid binding protein	6.18	3.6
RP 71955-Hiv replication inhibitor, 1RPB	10	3	6,1,3	0 (1)	1AS5 (4–22) Conotoxin Y-pIIIe	3.49	36.8
Uroguanylin, 1UYA	6	4	1,2,1,2[h,i]	1.08–1.21 (2)	1DNP (384A–395A) C-terminal domain of DNA Phtolyase	1.24	0
C-Terminal domain of midkine, 1 MKC	8	4	1,3,1,2[h]	0 (1)	1DBG (92A–134A) Chondroitinase B	1.69	2.3
Collagen-binding type II domain, 1PDC	18	5	1,5[h,i],3,2,7	0 (1)	1E88 (56A–97A) Fibronectin[j]	0.61	4.9
Fibrin binding finger domain of tissue-type plasminogen activator, 1TPN	8	4	1,2,1,4	0 (1)	1140 (74–111) Phage T4 lysozyme	1.44	5.3
Tertiapin, 1TER	6	3	3,1[h],2	0 (1)	1E6U (171A–186A) GDP-fucose synthetase	1.35	6
Endothelin, 1EDP	78	4	28[h,i],18[i],28[i],4[i]	1.36 (1)	1CBO (72A–86A) Cholesterol oxidase	0.30	0
Immunodominant region of protein BRSV-G, 1BRV	146	6	36,27[i],28[h,i],30,12,13	0 (1)	1CFA (10A–23A) Anaphylotoxin	0.69	0
Enterotoxin (STb), 1EHS	12	7	1,2,3[h],2,2,1,1	0 (1)	1A8J (154H–192H) Immunoglobulin	1.23	10.3

[a]Number of hits after removing closely related substructures at 90% sequence identity.
[b]Number of substructural clusters formed based on the dissimilarity matrix derived from structural superimposition of the template structures (see Results and Discussion).
[c]The population of substructural clusters (number of template structures forming a cluster).
[d]RMSD of the homologous templates (belonging to same SCOP family) present in the hits. Number of such hits is mentioned in parantheses. Wherever query itself came as a hit, RMSD is metioned as zero.
[e]Nonhomologous template with the lowest root mean square deviation (RMSD) with the query protein from the total number of hits. Also provided is the start and end residue of the substructure along with the chain identifier, wherever applicable; in parentheses.
[f]RMSD of the nonhomologous template with the query protein when aligned with cysteine positions as equivalencies. Structural superimposition is performed using MNYFIT without update of the equivalencies.
[g]Sequence identity of the nonhomologous template with the query protein.
[h]The cluster which included query structure on coclustering with the template structures.
[i]These clusters include at least one substructural template with <1 Å RMSD with the query peptide.
[j]Templates that are related at the superfamily level to the query peptide.

preferred backbone conformations for these constrained systems.

Proposing Models of Disulfide-Rich Polypeptides

Benchmarking on two-disulfide-bonded systems

Here, we describe how the database could successfully be queried to propose models for disulfide-rich polypeptides based on their disulfide connectivity pattern in the absence of structural homologs and any structural information. A pattern search program, MULSS, was used to search against the three derived disulfide databases—*nrnative*, *nrdb* and *fulldb*, comprising of native disulfide bonds, substructures that were found to stereochemically allow disulfide bonds in a nonredundant database of protein structures and the entire PDB release, respectively (see Materials and Methods for details). This program is used to search for protein substructures encoded in DSDBASE using one or more disulfide bonds as a

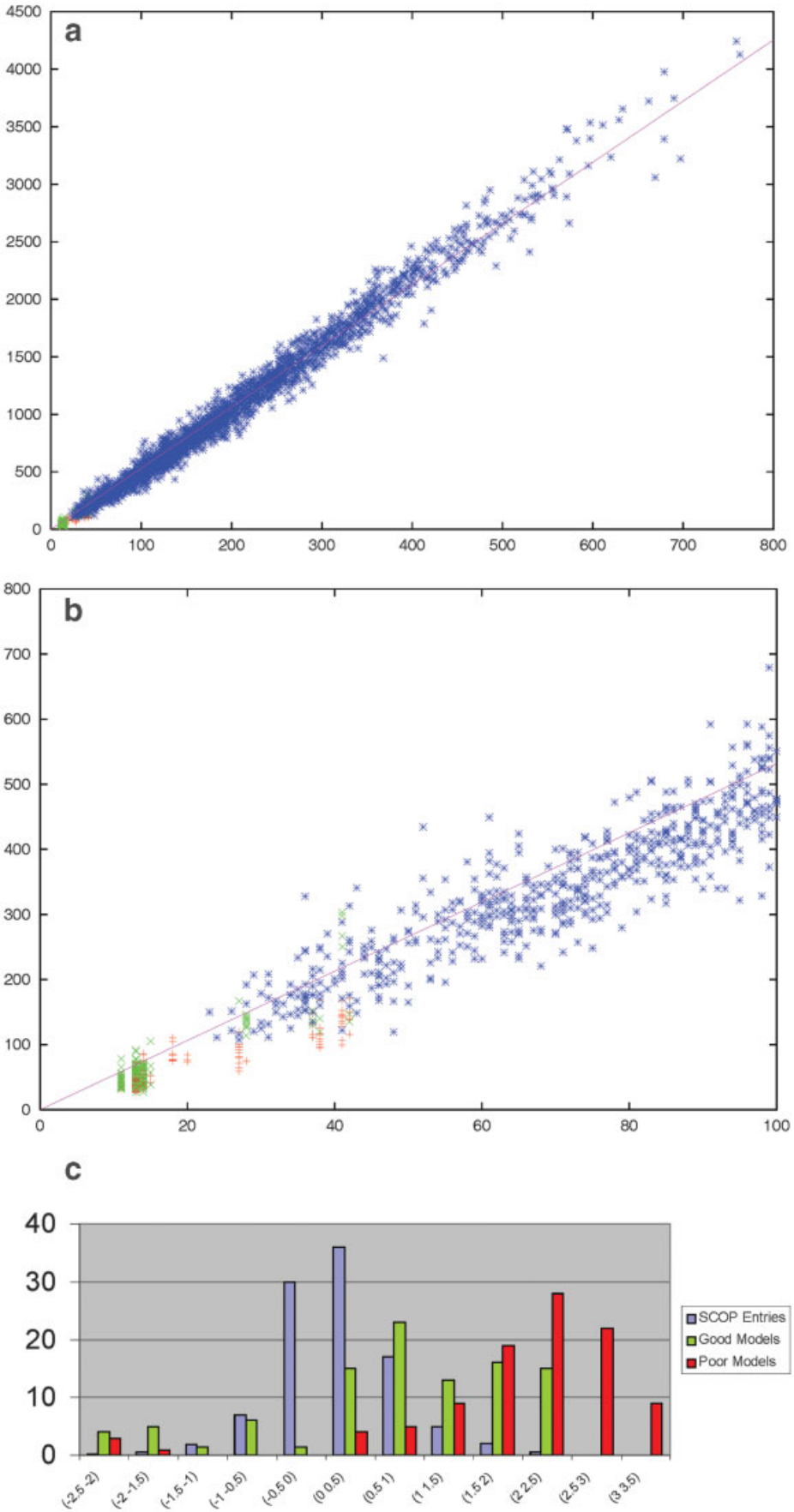
a
4500
4000
3500
3000
2500
2000
1500
1000
500
0
0
100
200
300
400
500
600
700
800
b
800
700
600
500
400
300
200
100
0
0
20
40
60
80
100
c
40
30
20
10
0
(-2.5,-2)
(-2,-1.5)
(-1.5,-1)
(-1,-0.5)
(-0.5,0)
(0,0.5)
(0.5,1)
(1,1.5)
(1.5,2)
(2,2.5)
(2.5,3)
(3,3.5)
SCOP Entries
Good Models
Poor Models

Figure 7.

pattern for recognition. The user can provide relaxation to the number of residues between two adjacent cysteines involved in separate crosslinks (loop proximity relaxation) or to relax the number of residues within a loop (loop size relaxation) to account for insertions and deletions in the alignment between query and the hits. Figure 3 provides a flow-chart of the approach starting from successful identification of substructures by probing into DSDBASE using MULSS.

Fifteen (15) nonredundant peptides, whose structural information is available and contain two disulfide bonds, were used for benchmarking (see Materials and Methods for details). In all the two-disulfide-bonded systems chosen for benchmarking, a stereochemically acceptable model could be obtained by this approach even without loop relaxation (please see below). In peptides where the disulfide bond connectivity follows a loop-within-loop pattern (a smaller disulfide loop within a larger loop), a large number of protein substructures are provided as plausible three-dimensional models. For example, in the case of disulfide bonds in immunodominant region of protein BRSV-G (PDB code 1brv) with a similar connectivity pattern (173–186; 176–182), a large number of hits were obtained when searched against the full database of DSDBASE (561 hits against *fulldb*).

A substructure in ribosomal S8 (PDB code 1hnz) is a nonhomologous hit that is highly compatible to the disulfide bond connectivity of conotoxin acquiring the lowest RMSD with the experimental conformation of the peptide amongst the various hits. A total of 163 hits were obtained by using conotoxin as a query and 26 substructures are nonredundant at the 90% sequence identity cut-off [Supplementary material gives a complete list of hits obtained for conotoxin as query and the substructures belonging to the nonredundant list along with the major clusters as determined by structure comparisons (see Materials and Methods for details)]. Figure 4 compares the experimental conformation of conotoxin compared to the backbone conformation of 23 substructures grouped into three main clusters. The backbone conformation of the highest populated cluster (including 1hnz) is very similar to the experimental structure of conotoxin.

Fig. 7. Validation scores for models of two-disulfide-bonded polypeptides using HARMONY (Topham et al.[38]; Sowdhamini, R., Srinivasan, N., and Blundell, T.L., unpublished results). **a:** Actual HARMONY scores plotted as a function of protein residue length. Points marked in blue correspond to 4020 nonredundant globular domains as described in SCOP database.[31] Points marked in green and red correspond to "good" and "poor" models of two-disulfide polypeptides, respectively (see Materials and Methods for a definition of "good" and "poor" models). A best fit straight line passing through the origin including all the 4020 protein domains acquire a slope of 5.3. In general, points corresponding to "poor" models appear below the fitted straight line indicative of strained conformations or misfolds. **b:** Inset shows the closeup of the same plot but zoomed to show only values corresponding to small folds. **c:** Extent of deviation of HARMONY scores from ideal expected values. Δ-m is the difference between ideal value and observed HARMONY score after normalisation for the protein length. Percentage of SCOP protein domains, good and poor models that correspond to different bins of Δ-m. The color representation is same as in (a). Non-redundant SCOP protein domains deviate very little from ideality (+1 to −1). "Good" models undergo lesser deviations compared to "poor" models. High Δ-m are associated with strained or incorrect models.

This test procedure also demonstrates the efficacy of enhancing the database by including substructures where stereochemically facile disulfide bond modeling was possible. For example, in the case of endothelin, even after eliminating closely related templates by 90% sequence identity cut-off from an initial list comprising of 249 hits, the number of hits in the full dataset is as high as 77 (Supplementary material and Fig. 5). The comparison of structural similarity is useful in further reducing the number of hits that can help in identifying potential templates for modeling.

Seven hits were obtained by searching for the connectivity pattern as in Shk toxin where six substructures were distinct after the sequence similarity filter. These can be grouped into two clusters on the basis of structural similarity. The lowest RMSD with the experimental structure is as high as 3 Å (Supplementary material and Fig. 6). Interestingly, neither of the two clusters are structurally very close to the experimental conformation of shk toxin (PDB code, 1c2u). The presence of multiple conformations, even in constrained systems, is not uncommon and such structurally distinct templates could be viewed as alternate conformations.

Table II summarizes the results obtained by comparing the experimental conformations of the 15 peptides along with the proposed structures by the present approach and examined as clusters of structurally similar polypeptides (superposed substructures that correspond to the major clusters in comparison to the queries are provided as Supplementary material). After eliminating closely related hits and jack-knife tests to discard the near-self matches, 12 peptides could be associated with a DSDBASE nonhomologous substructure that has less than 2 Å RMSD to the experimentally determined structure (see Table III and Supplementary material). In several cases, the most populated cluster indeed includes a member that is structurally similar (<1 Å RMSD values) to the experimental conformation of the query.

Structure validation of models for two-disulfide-bonded systems

In order that the best-fit models can be identified in real-time structure prediction and modeling, we have examined the local structural environment of the individual residues to score the compatibility of the model to the sequence. This procedure is encoded as HARMONY[38] (R. Sowdhamini, N. Srinivasan, T.L. Blundell, unpublished results; also see Materials and Methods). HARMONY has been applied on a large nonredundant dataset of globular structural domains in SCOP[31] database derived from Protein Data Bank[28,29] for obtaining ideal normalized scores. In general, good models retain high HARMONY values and models with local or global errors acquire low HARMONY values; misfolds appear as points well below the straight-line fit in a X-Y plot correlating HARMONY score with protein residue number (see Materials and Methods for details). Models that are structurally similar (within 2 Å) to the query strucure often acquire higher values [Fig. 7(a)]. Most of the models structurally

TABLE IV. Summary of Benchmarking the Disulfide Loop Search Procedure for Three SS Bonded Systems

	Length of the protein segment	Connectivity	Total number of hits (relaxation)[a]	90% Filter[b]	RMSD of homologs[c]	Nonhomologs hits[d]	Nonhomologs template with lowest RMSD	
							PDB code and name[e]	RMSD[f] (Å)
1 clvl (g.3.2)	30	501–518	9 (1,0)	3	0 (1, self)	2	1kk4F	1.82 (15)
Alpha-amylase inhibitor		508–523					129–159	
		517–531					1m8nA	1.96 (15)
							9–39	
1kal (g.3.3)	24	5–22	45 (1,1)	17	1.58 (1)	16	1o7tE	1.8 (17)
Kalata B1		13–27					252–257	
		17–29					1bh4	1.62 (16)
							1–25	
1i2uA (g.3.7)	35	7–32	15 (1,0)	6	0 (1, self)	5	1dceA	2.22 (16)
Heliomicin		18–40					460–495	
		22–42					Not good but the see strs superimposed very well.	
1glol (g.4.1)	28	4–19	160 (1,0)	27	1.45–1.7 (4)	23	1k8kC	1.86 (23)
Alpha-chymotrypsin		14–32					110–138	
		17–27					1spiB	1.28 (21)
							197–225	
							1a18	1.51 (26)
							80–108	
1mknA(g.5.1)	37	15–39	11 (1,0)	8	0 (1, self)	7	1j5wA	1.91 (25)
N-terminal domain of midkine		23–48					107–144	
		30–52						
1kth(g.8.1)	50	5–55	73 (0,0)	6	0.13–1.48 (73)	—	1bik	1 (50)
Kunitz type domain[g]		14–38					26–76	
		30–51					1dtx	1.01 (50)
							7–57	
							1bunB	1.29 (50)
							7–57	
							1shp	1.08 (50)
							3–53	
1dfnA(g.9.1)	27	3–31	50 (2,1)	10	0.76–1.44 (3)	7	1hi7B	1.88 (17)
Defensin		5–20					6–44	
		10–30					1 an 11	1.5 (18)
							14–42	
1 ans(g.11.1)	19	3–17	2 (1,0)	2	0 (1,self)	1	1dk4A	2.13 (13)
Neurotoxin III		4–11					61–80	
		6–22						
1n7dA(g.12.1)	46	47–61	46 (2,0)	7	0 (1,self)	6	No good template	
Extracellular domain of LDL receptor		54–74						
		68–83						
1cxrA(g.13.1)	37	3–40	30 (1,1)	11	0–1.9 (5)	6	1dknA	1.75 (16)
Crambin		4–32					88–125	
		16–26					1 bn8A	1.87 (18)
							78–115	
1a0hA(g.14.1)	68	170–248	44 (2,0)	7	0.88–1.62 (7)	—	1nk1 A	1.13 (62)
Ppack-Meizothrombin[g]		191–231					128–206	
		219–243					1 pmkA	1.22 (60)
							2–80	
1tbqR(g.15.1)	42	6–31	16 (1,0)	6	1.31–1.38 (3)	3	1occB	2.18 (16)
Rhodnin		8–27					154–196	
		16–48					1cyx	2.02 (16)
							165–207	
1es7A(g.17.1) bone-morphogenic protein	99	14–79	9 (1,0)	3	1.15 (2)	1	1pir	1.56 (21)

TABLE IV. (Continued)

	Length of the protein segment	Connectivity	Total number of hits (relaxation)[a]	90% Filter[b]	RMSD of homologs[c]	Nonhomologs hits[d]	Nonhomologs template with lowest RMSD: PDB code and name[e]	Nonhomologs template with lowest RMSD: RMSD[f] (Å)
		43–111					11–110	
		47–113						
1bgk(g.19.1)	35	2–37	5 (1,0)	2	0 (1,self)	1	1a6yB	1.64 (21)
Sea Anemone toxin		11–30					125–160	
		20–34						
1icfI(g.28.1)	58	197–216	6 (1,0)	3	0 (1,self)	2	1bmg	1.72 (18)
Class 11 MHC		227–234					8–66	
		236–255					1 il8A	1.81 (18)
							111–169	
1egf(g.3.11)	36	6–20	50 (1,1)	15	1.40–1.61 (4)	11	1hp7A	1.63 (20)
Epidermal growth factor		14–31					228–264	
		33–42					2achA	1.51 (18)
							228–264	
							1jajA	1.42 (20)
							76–113	
							1bebA	1.97 (18)
							80–117	
1hrt(g.3.15)	33	6–14	24	8	0 (1,self)	7	1opbB	1.66 (21)
Hirudin		16–25					97–130	
		22–39						

[a]Number of hits obtained on searching against the database derived from full protein database (PDB-April 2001 release).
[b]Number of hits after removing closely related substructures at 90% sequence identity.
[c]RMSD of the homologous templates (belonging to same SCOP family) present in the hits. Number of such hits is mentioned parantheses. Wherever query itself came as a hit, RMSD is metioned as zero.
[d]Number of nonhomologous templates.
[e]Nonhomologous template with the lowest root mean square deviation (RMSD) with the query protein from the total number of hits. Also provided is the start and end residue of the substructure along with the chain identifier.
[f]RMSD of the nonhomologous template with the query protein. Structural superimposition is performed using Multiprot (Shatsky et al. 2002). Number of aligned residues is mentioned in parentheses.
[g]All the hits for this query are structural homologs (same SCOP family), however the sequence identity dips below 40%.

distant from the query (more than 2 Å) acquire lower scores and such points fall well below the straight-line fit indicative of misfolds. Further, the scores obtained for 4020 nonredundant domains in SCOP database allowed us to obtain HARMONY scores that are normalized for their length (m(i)). This value is also the slope of the best straight line fit after considering the scores of 4020 protein domains plotted against protein size. Normalized HARMONY score:

$$m(o) = \frac{\text{observed HARMONY score}}{\text{Total number of residues in the protein domain}}$$

$$\Delta m = m(i) - m(o)$$

Good models, characterized by low structural deviations from the query experimental conformation, have smaller deviations from ideal normalized HARMONY score (Δm) whereas poor models have larger deviations in normalized HARMONY score from ideality [Fig. 7(c)]. This result is especially useful since such validation schemes can be applied to identify potential structural templates for modeling disulfide-rich polypeptides of unknown structure. Models with normalized HARMONY score close to an ideal value of 5.3 are ideal targets for further structure prediction. Models with good m-values reflected either as Δm close to zero or positive could be viewed as alternate conformations if very different from the query structure.

Modeling larger systems: three- and four-disulfide-bonded polypeptides

In general, search for suitable templates and possible structural homologs for disulfide-rich systems is much more effective in DSDBASE. For instance, kunitz-type domain from human type VI collagen alpha3 (IV) (PDB code-1kth) is listed under the superfamily "kunitz-like" in SCOP database[31] and relates to 77 protein entries at the family level. We first searched using PDB-Blast[39] and were able to identify only 42 hits, whereas a search in DSDBASE has resulted in 75 hits (97% coverage) since the cysteine positions and disulfide bonding pattern are highly conserved. The disulfide connectivity as observed in representative PDB entries for all the three- and four-disulfide-bonded systems yield nontrivial structural templates for several cases (20 out of 25 peptides [80%]). The results for the three- and four-disulfide-bonded systems are summarized in Tables IV and V,

TABLE V. Summary of Benchmarking the Disulfhide Loop Search Procedure for Four SS Bonded Systems

Length of the protein segment	Connectivity	Total number of hits (relaxation)[a]	90% Filter[b]	RMSD of homologs[c]	Nonhomologs hits[d]	Nonhomologs template with lowest RMSD	
						PDB code and name[e]	RMSD[f] (Å)
41	12–52 16–48 23–41 26–37	170 (1,0)	10	0 (1,self)	9	1qu7A 372–412 1rcc 23–63 1fha 27–67 1nkd 12–52	1.68 (35) 1.95 (28) 1.85 (27) 1.72 (25)
42	1–15 8–20 14–31 16–42	7 (3,1)	3	1.53 (2)	1	1bykA 1–44	1.82 (30)
44	7–25 10–27 29–48 36–50	72 (2,1)	10	0 (1,self)	9	1skz 37–81 C-terminal region superimposed very well	1.45 (23)
50	5–40 12–41 14–37 23–54	34 (3,1)	10	0 (1,self)	9	1adz 13–61 C-terminal region superimposed very well	1.83 (24)
55	3–22 17–37 39–51 52–57	5 (2,1)	4	0 (1,self)	3	No good template	—
105	7–16 15–78 44–109 48–111	255 (2,1)	8	1;10–1.44 (3)	5	No good template	—
36	1–10 6–29 7–34 19–36	6 (2,1)	3	1.85 (2)	2	No good template	—
37	3–39 4–31 12–29 16–25	91 (2,1)	20	0.61–0.92 (2)	18	No good template	—

[a]Number of hits obtained on searching against the database derived from full protein database (PDB-April 2001 release).
[b]Number of hits after removing closely related substructures at 90% sequence identity.
[c]RMSD of the homologous templates (belonging to same SCOP family) present in the hits. Number of such hits is mentioned in parentheses. Wherever query itself came as a hit, RMSD is metioned as zero.
[d]Number of nonhomologous templates.
[e]Nonhomologous template with the lowest root-mean-square deviation (RMSD) with the query protein from the total number of hits. Also provided is the start and end residue of the substructure along with the chain identifier.
[f]RMSD of the nonhomologous template with the query protein. Structural superimposition is performed using Multiprot (Shatsky et al. 2002). Number of aligned residues is mentioned in parentheses.

respectively. In the cases of α-chymotrypsin, kunitz-type domain, crambin, meizothrombin and epidermal growth factor, nonhomologous substructures are quite similar to the query; indeed, their RMSD values are lesser than observed between query and their respective homologs. Figure 8 provides illustrative examples of two-, three- and four-SS-bonded systems where non-trivial nonhomologous substructures could be identified by the present approach.

SUMMARY

Several extracellular regions of receptors and many genes involved in signal transduction contain disulfide-rich domains. There are also large number of bioactive molecules like toxins and inhibitors that are rich in disulfides. Similarities between disulfide-rich domains and convincing alignments are usually hard to establish by simple sequence procedures due to their high disulfide

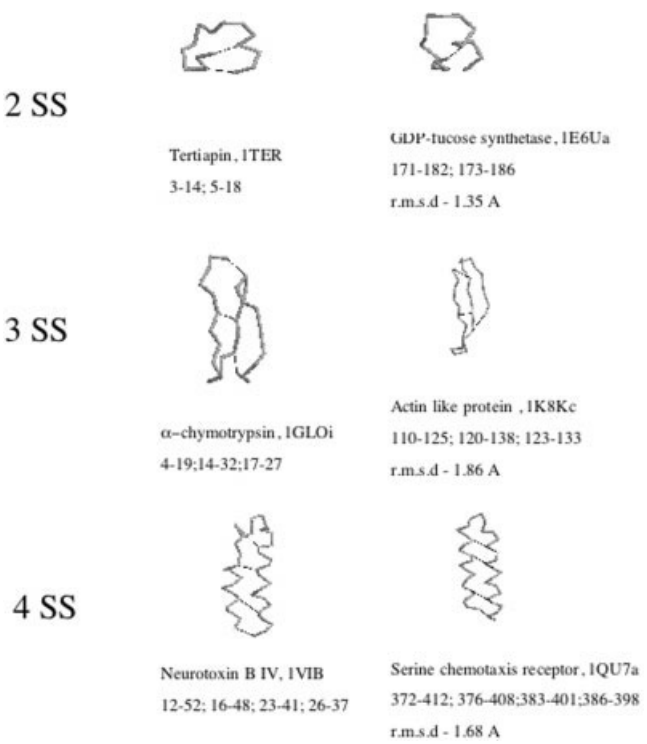

Fig. 8. Illustrative examples of two-, three- and four-disulfide-bonded systems employed for benchmarking the modeling procedure. In each case, the C^{α}-trace and the positions of the crosslinks are shown. The best template is a substructure from an unrelated protein entry whose name and PDB code are mentioned. MODIP-(17)-modeled disulfide bonds are also projected on the templates. As shown by the RMSD between query and the template, in every case a structurally similar nonhomologous template could be identified from a protein lacking similar disulfide bond patterns.

content and spurious high percentage identities. In the absence of a comprehensive database, they are usually recognized by a manual examination and subjective knowledge-based techniques. Chuang and coworkers[40] report several interesting distant relationships between proteins using disulfide bonding connectivity and native disulfides. Recently, Vlijmen and coworkers[41] employed disulfide pattern information or annotation to draw relation between homologous proteins and also to identify distantly related proteins by searching in sequence databases. Ranganathan and coworkers[42] have reported the availability of a web resource for homology modeling of disulfide-rich systems starting from a database that consists of proteins less than 100 amino acids and containing two disulfide bonds for the selection of appropriate templates. However, the other important limitations in these approaches to structure prediction are the restriction to native disulfides and/or the limited number of structures available for disulfide-rich systems. DSDBASE includes modeled disulfides and substructures that have a stereochemical compatability to accept disulfides. Indeed, only 2% of the database relates to native disulfides[25] and therefore provides higher number of protein substructures as potential templates. We propose an approach for generating meaningful models of disulfide-rich polypeptides in the absence of structural homologs or even similar disulfide crosslinks in templates.

The search program, MULSS, is available over the Web along with the database and queries with user-defined relaxations in the connectivity patterns can be submitted to provide immediate results. We demonstrate that three-dimensional models can be obtained reliably by querying DSDBASE for an observed SS bond connectivity and no other experimental inputs. Forty polypeptides were chosen from the PDB for benchmarking studies and include two-, three-, and four-disulfide-bonded polypeptides ranging from 13–99 residues in length. In all 40 polypeptides, substructures could be identified in DSDBASE that satisfy the crosslink connectivities (see Tables II, III, and IV). The current approach could identify templates from nonhomologous protein substructures, that are structurally similar to the query, for 35 out of 40 cases studied. The five polypeptides where we failed to obtain satisfactory templates are extracellular domain of LDL receptor (PDB code: 1n7d, 3 SS system), dendroaspin (1drs, 4 SS system), TGF-β3 (1tgj, 4 SS system), obtustatin (1mpz, 4 SS system), and β-purothionin (1bhp, 4 SS system). Perhaps modeling of larger systems can be guided by the predicted locations of secondary structures or in stages leading to a rule-based assembly of fragments.

In 69% of two-disulfide-bonded systems, cases studied using known structures, the most populated cluster is indeed closer to the experimental conformation by means of coclustering the query. Therefore, the correct conformation of a peptide can be identified, in the absence of experimental data or sensitive validation techniques, with 70% reliability. Twelve (12) out of 15 known examples (80%) of two-disulfide-bonded systems could be predicted within 2 Å RMSD from the query without any relaxation of loop parameters. Models within 2 Å RMSD from the query could be arrived for 29 out of 40 polypeptides (72%) chosen for benchmarking starting from a nonhomologous structural template. These results clearly emphasize the importance of examining the local stereochemistry of known protein structures as a knowledge-base for predicting the structure of small polypeptides which generally do not follow the principles that underly the folding of large globular proteins. This is further strengthened by the observation of Kiharo and Skolnick[43] that all small proteins have significant structure alignments to other proteins in different secondary structures that can cross fold and structural class boundaries.

A similar approach has also been successfully applied by us earlier to predict the structure of N-terminal preactivation peptide (PAP) domain of hepatocyte growth factor[27] and the importance of such a database is emphasized.[25] Several studies revealed the importance of disulfide bonding pattern in determining protein structure.[25,29,31,44–47] In this work, we demonstrate that modeled disulfide bonding patterns in known protein substructures can be effectively employed to predict the structure of disulfide-rich polypeptide systems. Our approach is not limited by the absence of structural homologs or native disulfide bonds in the database. A higher number of possible substructures could be obtained for the right disulfide bond connectivity using this search procedure in

the case of endothelin. This approach, after being tested for a number of other examples, may find valuable application in predicting the disulfide-bond connectivity of polypeptides rich in Cys residues. The determination of the correct connectivity of disulfide-rich peptides is often time consuming and sometimes tedious.

SS-rich polypeptides are quite peculiar folds with poor secondary structures and higher numbers of solvent-exposed hydrophobic residues. A recent feature in MODIP has been the measurement of energetically unfavorable nonbonded contacts with the modeled disulfide[48] that can help to discriminate unrealistic models. Independently, validation methods such as HARMONY are useful since high scores reflect better compatability of query sequence to the model. The calculation of HARMONY scores for a large number of crystal structures of globular proteins and plotting the scores as a measure of protein length (residue number) is a reliable approach to identifying potential misfolds. Three-quarters of the models that are less than 2 Å to the query structure do not fall into the most populated cluster. Therefore, HARMONY scores would be helpful to include additional models that do not belong to the most populated cluster. It is worthwhile to note that for ab initio structure prediction of small folds, models that are 6.5 Å from the query are still considered as acceptable and near-native models.[49] In our analysis, we find that the distribution of structural dissimilarity with the query leads to an RMSD cutoff of 2 Å (data not shown).

Globular proteins undergo deviations in normalized HARMONY scores (Δm) within $+/-1$ from ideality. Out of 318 models of two-disulfide-bonded systems examined by us, 81% of the "poor" models (more than 2 Å RMSD from the query) have high extents of deviation from ideal normalized scores (Δm 1.5 to 3.0) indicating that simple HARMONY scores are sufficient to identify "poor" or strained models. For instance, by placing an upper threshold of 2.5 for Δm, around 65% of false positives or poor models can be eliminated without compromising on the number of true positives or good models. It is rewarding to note that a simple validation scheme, by attributing scores to the local environment of particular residues, can effectively allow us to identify potential false positives. Another interesting observation that we report in this paper is in the case of higher order systems such as three- and four-disulfide-bonded systems employed for benchmarking, models derived from unrelated/nonhomologous protein substructures acquire smaller RMSD values from the query than even the RMSD of the query from some of its structural homologs. Additional inputs/constraints such as the position of secondary structures or a cis peptide bond or extra distance restraints will also be included in the search algorithms. As reported earlier, favored combinations are observed for the side chain rotamers in disulfide bridges[13,50,51] and preferred topologies[52] that can be employed to scan the huge conformational space or to restrain the internal parameters.

ACKNOWLEDGMENTS

R.S. is awarded the International Senior Research Fellowship by the Wellcome Trust, UK. A.V., T.R.R., A.M. and G.P. are supported by The Wellcome Trust. T.R.R. is currently supported by a PhD grant from the Conseil Regional de La Reunion. We thank Prof. Balaram for initiating this idea, Prof. Tom Blundell, Dr. N. Srinivasan, and Dr. Alexandre de Breven for useful discussions. We thank Dr. Axel Innis for critical reading of the manuscript. The original version of the MODIP procedure was developed in Prof. P. Balaram's laboratory in collaboration with Prof. C. Ramakrishnan. Contributions from Prof. C. Ramakrishnan in the development of MODIP program are gratefully acknowledged.

REFERENCES

1. Poland DC, Scheraga HA. Statistical mechanics of noncovalent bonds in polyamino acids. VIII. Covalent loops in proteins. Biopolymers 1965;3:379–399.
2. Pace CN, Grimsley GR, Thomson JA, Barnett BJ. Conformational stability and activity of ribonuclease T1 with zero, one, and two intact disulfide bonds. J Biol Chem 1988;263:11820–11825.
3. Wetzel R, Perry LJ, Baase WA, Becktel WJ. Disulfide bonds and thermal stability in T4 lysozyme. Proc Natl Acad Sci USA 1988;85:401–405.
4. Matsumura M, Matthews BW. Control of enzymatic activity by an engineered disulfide bond. Science 1989;243:792–794.
5. Gokhale RS, Agarwalla S, Francis VS, Santi DV, Balaram P. Thermal stabilization of thymidylate synthase by engineering two disulfide bridges across the dimer interface. J Mol Biol 1994;235:89–94.
6. Gale AJ, Xu X, Pellequer J-L, Getzoff ED, Griffin JH. Interdomain engineered disulfide bond permitting elucidation of mechanisms of inactivation of coagulation factor Va by activated protein C. Protein Sci 2002;11:2091–2101.
7. Ivens A, Mayans O, Szadkowski H, Jurgens C, Wilmanns M, Kirschner K. Stabilization of a ({beta}{alpha})8 -barrel protein by an engineered disulfide bridge. Eur J Biochem 2002;269:1145–1153.
8. Pikkemaat MG, Linssen ABM, Berendsen HJC, Janssen DB. Molecular dynamics simulations as a tool for improving protein stability. Protein Eng 2002;15:185–192.
9. Farzan M, Choe H, Desjardins E, Sun Y, Kuhn J, Cao D, Archambault Kolchinsky P, Koch M, Wyatt R, Sodroski J. Stabilization of human immunodeficiency virus type 1 envelope glycoprotein trimers by disulfide bonds introduced into the gp41 glycoprotein ectodomain. J Virol 1998;72:7620–7625.
10. Velanker SS, Gokhale RS, Ray SS, Gopal B, Parthasarathy S, Santi DV, Balaram P, Murthy MR. Disulfide engineering at the dimer interface of Lactobacillus casei thymidylate synthase: crystal structure of the T155C/E188C/C244T mutant. Protein Sci 1999;8:930–933.
11. Hinck AP, Truckses DM, Markley JL. Engineered disulfide bonds in staphylococcal nuclease: effects on the stability and conformation of the folded protein. Biochemistry 1996;35:10328–10338.
12. Richardson JS. The anatomy and taxonomy of protein structures. Adv Prot Chem 1981;34:167–339.
13. Thornton JM. Disulfide bridges in globular proteins. J Mol Biol 1981;151:261–287.
14. Srinivasan N, Sowdhamini R, Ramakrishnan C, Balaram P. Conformations of disulfide bridges in proteins. Int J Pept Protein Res 1990;35:147–155.
15. Pabo CO, Suchanek EG. Computer-aided model-building strategies for protein design. Biochemistry 1986;25:5987–5991.
16. Hazes B, Dijkstra BW. Model-building of disulfide bonds in proteins with known three-dimensional structure. Protein Eng 1988;2:119–125.
17. Sowdhamini R, Srinivasan N, Shoichet B, Santi DV, Ramakrishnan C, Balaram P. Stereochemical modelling of disulfide bridges: Criteria for introduction into proteins by site-directed mutagenesis. Protein Eng 1989;3:95–103.

18. Hruby VJ, Kretenasky JL, Cody WL. Ann Rep Med Chem 1984;19:303–312.
19. DeCoen JL, Ralston E. In Bergmann, ED and Pullman B, editors. The Jeruselam symposia on quantum chemistry and biochemistry. Jeruselam: Jeruselam Academic Press; 1973;5:41.
20. Sowdhamini R, Ramakrishnan C, Balaram P. Modelling multiple disulfide loop containing polypeptides by random conformation generation. The test cases of alpha-conotoxin GI and endothelin I. Protein Eng 1993;6:873–882.
21. Froimowitz M. Conformational search in enkephalin analogues containing a disulfide bond. Biopolymers 1990;30:1011–1025.
22. Paul PK, Dauber-Osguthorpe P, Campbell MM., Brown DW, Kinsman RG, Moss C, Osguthorpe DJ. Accessible conformations of melanin-concentrating hormone: a molecular dynamics approach. Biopolymers 1990;29:623–637.
23. Blundell TL, Bedarkar S, Rinderknecht E, Humbel RE. Insulin-like growth factor: a model for tertiary structure accounting for immunoreactivity and receptor binding. Proc Natl Acad Sci USA 1978;75:180–184.
24. Jones TA, Thirup S. Using known structures in protein model building and crystallography. EMBO J 1986;5:819–822.
25. Vinayagam A, Pugalenthi G, Rajesh T, Sowdhamini R. DSD-BASE: A consortium of native and modelled disulfide bonds in proteins. Nucleic Acid Res 2004;32:D200–202.
26. Sowdhamini R. Ph.D. Motifs in proteins: Disulfide constraints and their application to protein engineering and peptide modelling [dissertation]. Bangalore, India: the Indian Institute of Science; 1992.
27. Donate LE, Gherardi E, Srinivasan N, Sowdhamini R, Aparicio S, Blundell TL. Molecular evolution and domain structure of plasminogen-related growth factors (HGF/SF and HGFl/MSP). Protein Sci 1994;3:2378–2394.
28. Bernstein FC, Koetzle TF, Williams GJ, Meyer EE, Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M. The Protein Data Bank: a computer-based archival file for macromolecular structures. J Mol Biol 1977;112:535–542.
29. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. The Protein Data Bank. Nucleic Acids Res 2000;28:235–242.
30. Hobohm U, Scharf M, Schneider R, Sander C. Selection of representative protein data sets. Protein Sci 1992;1:409–417.
31. Murzin AG, Brenner SE, Hubbard T, Chothia C. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 1994;247:536–540.
32. Sutcliffe MJ, Haneef I, Carney D, Blundell TL. Knowledge based modelling of homologous proteins, Part I: Three-dimensional frameworks derived from the simultaneous superposition of multiple structures. Protein Eng 1987;1:377–384.
33. Felsenstein J. An alternating least squares approach to inferring phylogenies from pairwise distances. Syst Biol 1987;46:101–111.
34. Bower MJ, Cohen FE, Dunbrack RL Jr. Prediction of protein side-chain rotamers from a backbone-dependent rotamer library: a new homology modeling tool. J Mol Biol 1997;267:1268–1282.
35. Morris AL, MacArthur MW, Hutchinson EG, Thornton JM. Stereochemical quality of protein structure coordinates. Proteins 1992; 12:345–364.
36. Bowie JU, Luthy R, Eisenberg D. A method to identify protein sequences that fold into a known three-dimensional structure. Science 1991;253:164–170.
37. Wako H, Blundell TL. Use of amino acid environment-dependent substitution tables and conformational propensities in structure prediction from aligned sequences of homologous proteins. I. Solvent accessibility classes. J Mol Biol 1994;238:682–692.
38. Topham CM, Srinivasan N, Thorpe CJ, Overington JP, Kalsheker NA. Comparative modelling of major house dust mite allergen Der p I: structure validation using an extended environmental amino acid propensity table. Protein Eng 1994;7:869–894.
39. Li W, Jaroszewski L, Godzik A. Sequence clustering strategies improve remote homology recognitions while reducing search times. Protein Eng 2002;15:643–649.
40. Chuang CC, Chen CY, Yang JM, Lyu PC, Hwang JK. Relationship between protein structures and disulfide-bonding patterns. Proteins 2003;53:1–5.
41. Van Vlijmen HW, Gupta A, Narasimhan LS, Singh J.A novel database of disulfide patterns and its application to the discovery of distantly related homologs. J Mol Biol 2004;335:1083–1092.
42. Kong L, Lee BT, Tong JC, Tan TW, Ranganathan S. SDPMOD: an automated comparative modelling server for small disulfide-bonded proteins. Nucleic Acids Res 2004;32:W356–359.
43. D. Kihara, J. Skolnick. The PDB is a covering set of small protein structures. J Mol Biol 2003;334:793–802.
44. Cohen FE, Kuntz I. Tertiary structure prediction. In: Fasman GD, editor. Prediction of protein structure and the principles of protein conformation. New York: Plenum Press, 1989. p 647–707.
45. Rufino SD, Blundell TL. Structure-based identification and clustering of protein families and superfamilies. J Comput Aided Mol Des 1994;8:5–27.
46. Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. CATH—a hierarchic classification of protein domain structures. Structure 1997;5:1093–1098.
47. Mas JM, Aloy P, Marti-Renom MA, Oliva B, Blanco-Aparicio C, Molina MA, de Llorens R, Quero IE, Aviles FX. Protein similarities beyond disulfide bridge topology. J Mol Biol 1998;284:541–548.
48. Dani VS, Ramakrishnan C, Varadarajan R. MODIP revisited: re-evaluation and refinement of an automated procedure for modeling of disulfide bonds in proteins. Protein Eng 2003;16:187–193.
49. Huang ES, Samudrala R, Ponder JW. Ab initio fold prediction of small helical proteins using distance geometry and knowledge-based scoring functions. J Mol Biol 1999;290:267–281.
50. Engh RA, Huber R. Accurate bond and angle parameters for X-ray protein structure refinement. Acta Crystallogr A 1991;47:392–400.
51. Lovell SC, Michael Word J, Richardson JS, Richardson DC. The penultimate rotamer library. Proteins 2000;40:389–408.
52. Harrison PM, Sternberg MJ. Analysis and classification of disulfide connectivity in proteins. The entropic effect of cross-linkage. J Mol Biol 1994;244:448–463.