Efficient Modeling of Disulfide Bonds with ROSETTA

Spencer Bliven

April 1, 2009

Abstract

1 Background

The ROSETTA computer program provides a framework for many types of macromolecular modeling problems. It is used to answer a wide variety of biochemical questions, from protein folding to protein design. We added functionality to efficiently compute the strength of disulfide bonds. These improved ROSETTA's ability to discriminate between native-like disulfides and decoys.

1.1 ROSETTA

The current version of ROSETTA [3] is composed of several layers of functionality. At the core lies ROSETTA's model of the molecules being studied. This is essentially a collection of atoms from which larger concepts such as residues and secondary structure elements are formed. The three dimensional structure of the protein is determined by specifying the torsion angles between atoms.

Associated with each atom are energies for the various forces the atom experiences: electrostatic repulsion, van der Waals interactions, hydrogen bonds, etc. Some energies are not dictated by understood physical processes, but rather reflect known properties for native proteins. Examples of this are the energy terms for the phi and psi angles of protein backbones. Although fully modeling the forces on the backbone atoms should theoretically determine the proper phi and psi angles, it is more computationally feasible to heuristically model these angles. The phi and psi energy functions measure how close the angles are to the accepted areas of the Ramachandran plot for a given amino acid. This results in low energy scores for structures which are more similar to native protein conformations.

Above the core model ROSETTA provides a layer of methods to manipulate the model. These could correspond to physical movement of the peptide chain, changes to model attributes such as secondary structure assignment, or to changes to the amino acid sequence itself for design purposes. Any change in the state of the protein triggers the protein's energy to be re-scored. Thus one can evaluate whether a particular movement decreases the total energy of the structure.

Although ROSETTA is used to solve a wide range of problems, conceptually all can be understood as the search for the minimum of some complex objective function. Any given state of the model can be represented as a point in a high-dimensional space defined by the values of the individual energy functions applied to that model. The objective function then maps this space to a single quantity, the minima of which will correspond to the best model for the current goal. For instance, for *ab initio* folding the objective function is simply the Gibbs free energy of the protein and the best conformation will be the most thermodynamically stable. For protein design more specific objective functions may be used, such as maximizing the number of ionic interactions.

Finding the global minimum is a difficult challenge due to the ruggedness of the objective function's landscape and the high number of degrees of freedom to be searched. To sample the full landscape ROSETTA uses a Monte-Carlo algorithm which combines large random perturbations with local refinements. To reduce the number of degrees of freedom, particularly computationally intensive problems use a simplified model of the protein. For each amino acid the atoms of the side chain are replaced with a single 'centroid' atom with similar volume and chemical properties to the original residue. This eliminates all rotamers of the amino acids from the calculation. After a fast pass to identify low-energy backbone conformations with the centroid amino acids, the full side chains can be reintroduced and a second minimization pass will refine the rough backbone structures and introduce the correct side chain rotamers.

One implication of having two types of amino acids is that separate scoring functions must be derived for the full atom and centroid models of proteins. The current version of ROSETTA has only full atom scoring functions for disulfide bonds, so disulfides can not be considered during the centroid pass. During the fast initial pass ROSETTA cannot identify the large decrease in energy associated with the formation of disulfide bonds or the increase in energy when they are broken. Thus the disulfide bonds are required to be manually specified and remain constant over the course of the search. By adding centroid scoring func-

tions for disulfide terms we expect future protocols in ROSETTA to handle disulfide bonds in a much more dynamic fashion.

1.2 Disulfide Bonds

The general structure of disulfide bonds has been characterized by Thornton [6] and Richardson [5]. Disulfide conformation can be specified by five dihedral angles around the bonds of the side chains (See figure 1a). The most energetically favorable conformation is a left-handed spiral with χ_2 and χ_3 of approximately -90 degrees [6], but the opposite right-handed spiral is observed as well.

The $S\gamma - S\gamma$ distance in disulfides in tightly constrained to the ideal distance of 2.02Å. This makes it relatively easy to determine the locations of disulfide bonds from a given structure. The energy wells around disulfides are deep and narrow, making it difficult to accidentally break a disulfide bond once it has been formed. For centroid mode a shallower scoring function was needed which could detect a disulfide bond even without perfect geometry. It had to be both fast and accurate without precise sulfur geometries.

2 Methods

Centroid scoring terms were originally designed by Bill Schief in 2002, but were never incorporated into the current version of ROSETTA. We first verified the continued relevance of the scores using current structures from the PDB, then proceeded to implement them in ROSETTA.

2.1 Scoring Functions

Five scoring terms are considered for centroid protein conformations (see figure 2). Each considers the geometry between a pair of residues (see figure 1b). Two functions involve the distance between residues: $C\beta$ distance and centroid distance. The other three are based on the angles between the two residues: the planar (θ) angle, χ_{alpha}, and χ_β. Additionally, the angular score functions are re-weighted based on the distance between residues, as discussed in section 2.2.

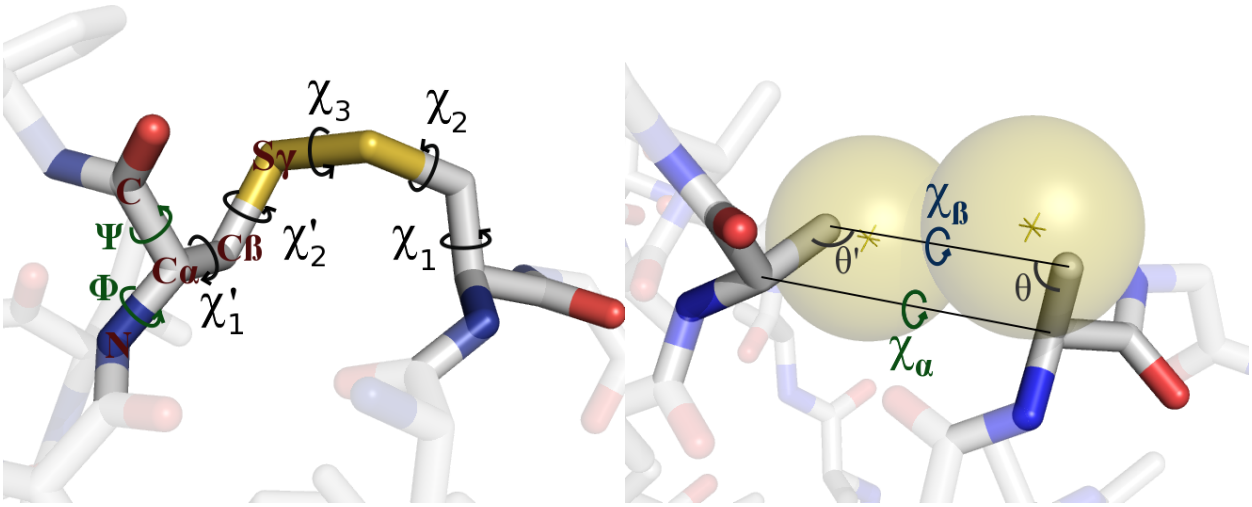
The scores were generated using structures from the ASTRAL SCOP 1.55 database [2, 4]. This aggregates protein structure information and removes domains more than 40% identical by sequence or structure to produce a high-quality, non-redundant database of protein domain structures. A set of 515 disulfide-containing domains was selected containing 2186 disulfide bonds. The geometry of these native disulfides was calculated and an initial distribution acquired from the histogram of each measurement. To control for any existing propensity of ROSETTA to create disulfide-like geometries between cysteines, a set of decoy protein structures was generated from ROSETTA. Thirty-eight of the proteins sequences were used as inputs to ROSETTA's *ab initio* folding protocol, generating 10,000 decoy structures. Each decoy represents a local minimum in the existing scoring functions in ROSETTA, but most are not very similar to the native structure for that sequence. Distributions where calculated for each of the five geometric measurements in the same manner as for the native disulfides. The score for a particular measurement was then given by the log ratio of natives to decoys:

$$f(x) = -log \frac{\text{fraction of native disulfides with } x}{\text{fraction of decoys with } x}.$$

These raw scores contained considerable noise, so smoothing techniques were applied to each. These included both automated methods using the IGOR PRO data analysis package [1] and manual massaging to create smooth scoring functions suitable for gradient descent algorithms. The $C\beta$ distance was fit using a sum of three Gaussian functions

$$\begin{aligned} f_{C\beta \text{ dist}}(x) = 5.0 &- 1.2575 \cdot \mathcal{N}\left(12.445,\, 1.1748^2\right) \\ &- 1.1084 \cdot \mathcal{N}\left(15.327,\, 2.1956^2\right) \\ &- 0.33851 \cdot \mathcal{N}\left(4.0,\, 0.35355^2\right). \end{aligned}$$

The component Gaussians can be seen in figure 2a. The remaining functions were fit using polynomial splines. Since no native disulfides were found with planar angles less than 60 degrees, the raw score was



(a) (b)

Figure 1: Angle nomenclature. (a) Full atom angles. Phi (Φ) is the dihedral between the $C-N$ and $C\alpha-C$ bonds and psi (Ψ) is between the $N-C\alpha$ and $C-N$ bonds. χ_1 measures between the $N-C\alpha$ and $C\beta-S\gamma$ bonds. This figure shows the 6-48 dihedral of IGF-1 (pdb:1IMX). Bond angles are $\chi_1 = -58.92$, $\chi_2 = -81.43$, $\chi_3 = 96.69$, $\chi'_2 = 61.16$, and $\chi'_1 = 164.40$. (b) The centroid model of this bond. χ_α refers to the dihedral between the $N-C\alpha$ of the lower numbered residue (6, to the right) and the $C\alpha-C$ bond of the higher numbered residue (48, on the left). Angles are $\chi_\alpha = 106.15$, $\chi_\beta = 95.78$, $\theta = 79.97$, and $\theta' = 111.81$.

infinitely high over this range. The smooth version of the planar angle score was set arbitrarily to +100.

2.2 Weighting of angular scores

A problem with globally applying all the score functions is that residues which are far apart cannot influence each other enough to order their relative angles. Residues which randomly have the correct relative orientation for a disulfide bond should not receive low scores if they are too far apart. To account for this the angular terms are gradually weighted less and less as the distance between residues increases so that non-interacting pairs will not unduly influence each other. The initial score, calculated according to the functions in figure 2, are then weighted by a scoring factor $\omega(d)$ which depends on the $C\beta$ distance.

$$\omega(d) = \begin{cases} 0.0 & 0 < d < 3.16 \\ 1.0 & 3.16 < d < 4.69 \\ \frac{1}{20-4.69}(20-d) & 4.69 < d < 20 \\ 0.0 & 20 < d \end{cases}$$

2.3 Reference set

To verify the accuracy of the scoring functions we applied them to a more recent set of protein structures. Wang and Dunbrack publish lists of PDBs culled by their PISCES server for redundancy and resolution [7, 8]. A cutoff of 2.0Å resolution was chosen based on the need for precise knowledge of the disulfide geometries, and a threshold of 50% identity was selected due to indications that disulfides are very strongly conserved among homologous proteins

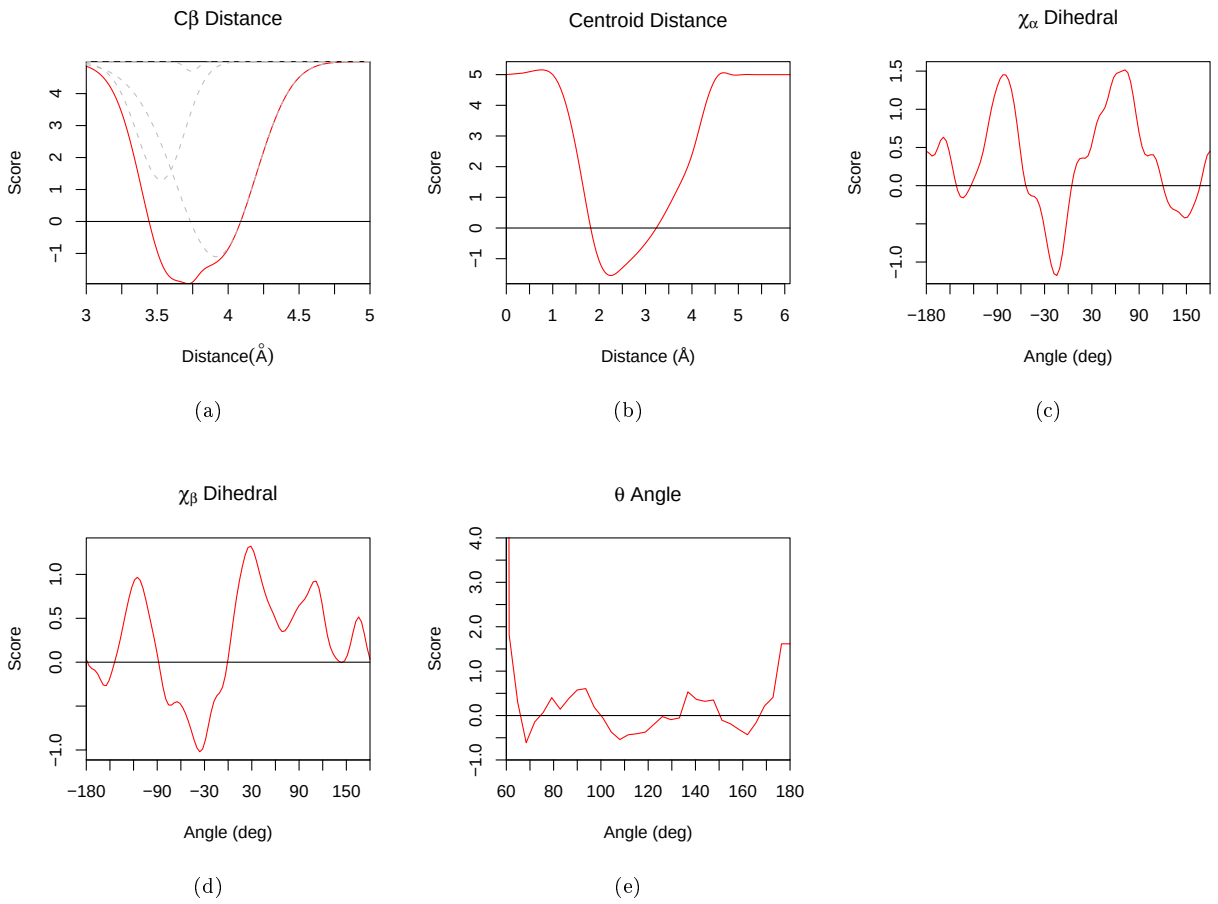


Figure 2: Centroid Scoring terms. (a) The $C\beta$ distance score function, and the three Gaussian functions of which it is composed. (b) Centroid distance score. (c) χ_α dihedral angles. (d) χ_β dihedral angles. (e) The θ angle score over the range $[60, 180]$ degrees. The score is defined as $+100$ in $[0, 60)$ degrees.

more closely related than that [6]. The PISCES list[1] contains 1134 proteins with disulfide bonds, giving 2839 native disulfides for our data set. The location of disulfide bonds was determined by ROSETTA's full-atom disulfide detection algorithm, which considers the proximity of $S\gamma$ and corresponds closely to PDB annotations. None of the structures used to generate the scores were used in the verification set.

In the creation of the scoring term a very specific decoy set was used since the goal was the improvement of ROSETTA'S *ab initio* folding. However to assess the utility of the scores a more general reference set was required. We used two such distributions. The first reference set consists of all permutations of cysteines not involved in disulfide bonds. This generated 11,616 pairs of cysteines. The second set contained 52,844 pairs chosen at random from among all 37.6 million possible pairs of residues in the PISCES structures.

[1] Generated November 14, 2008

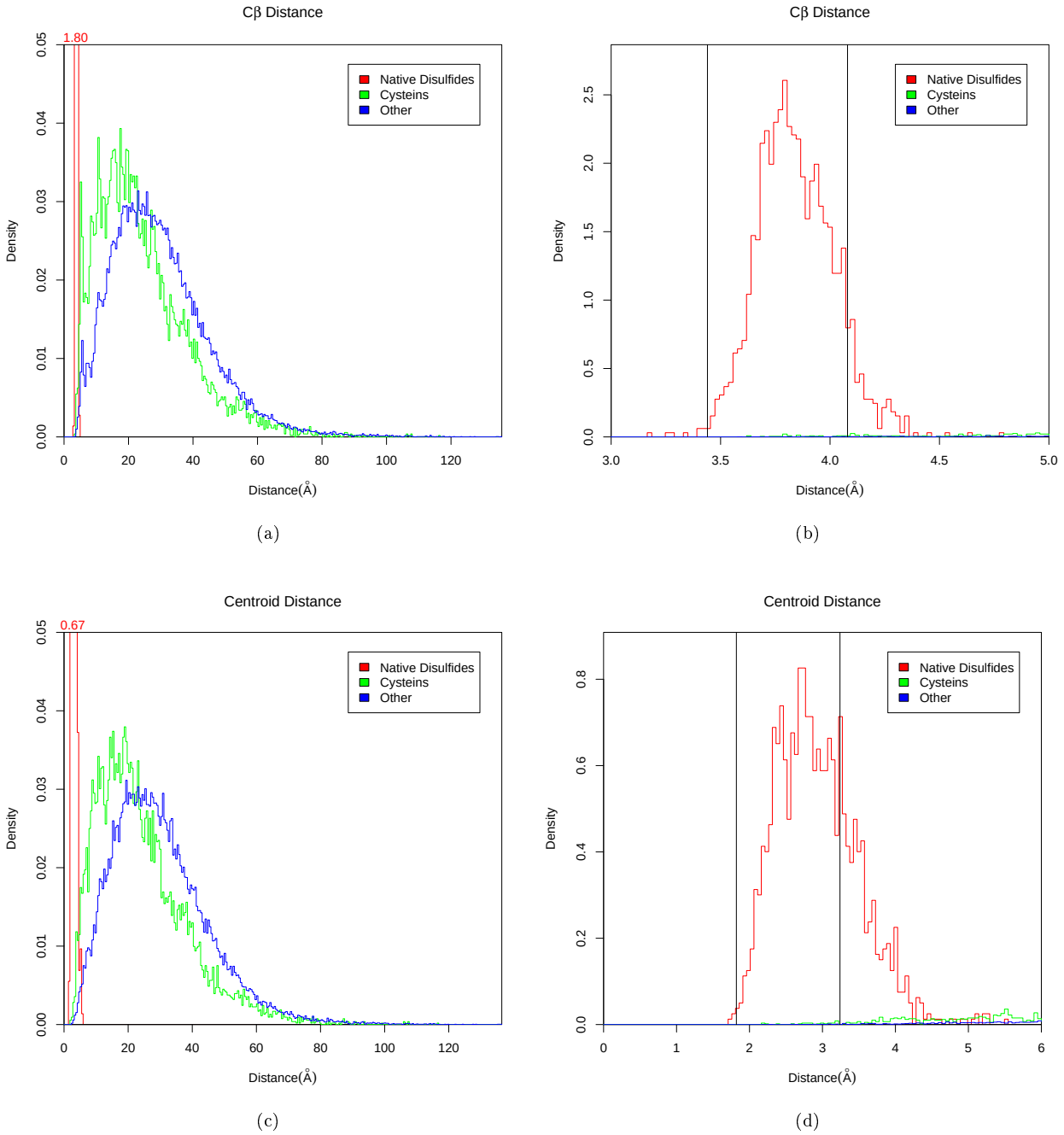
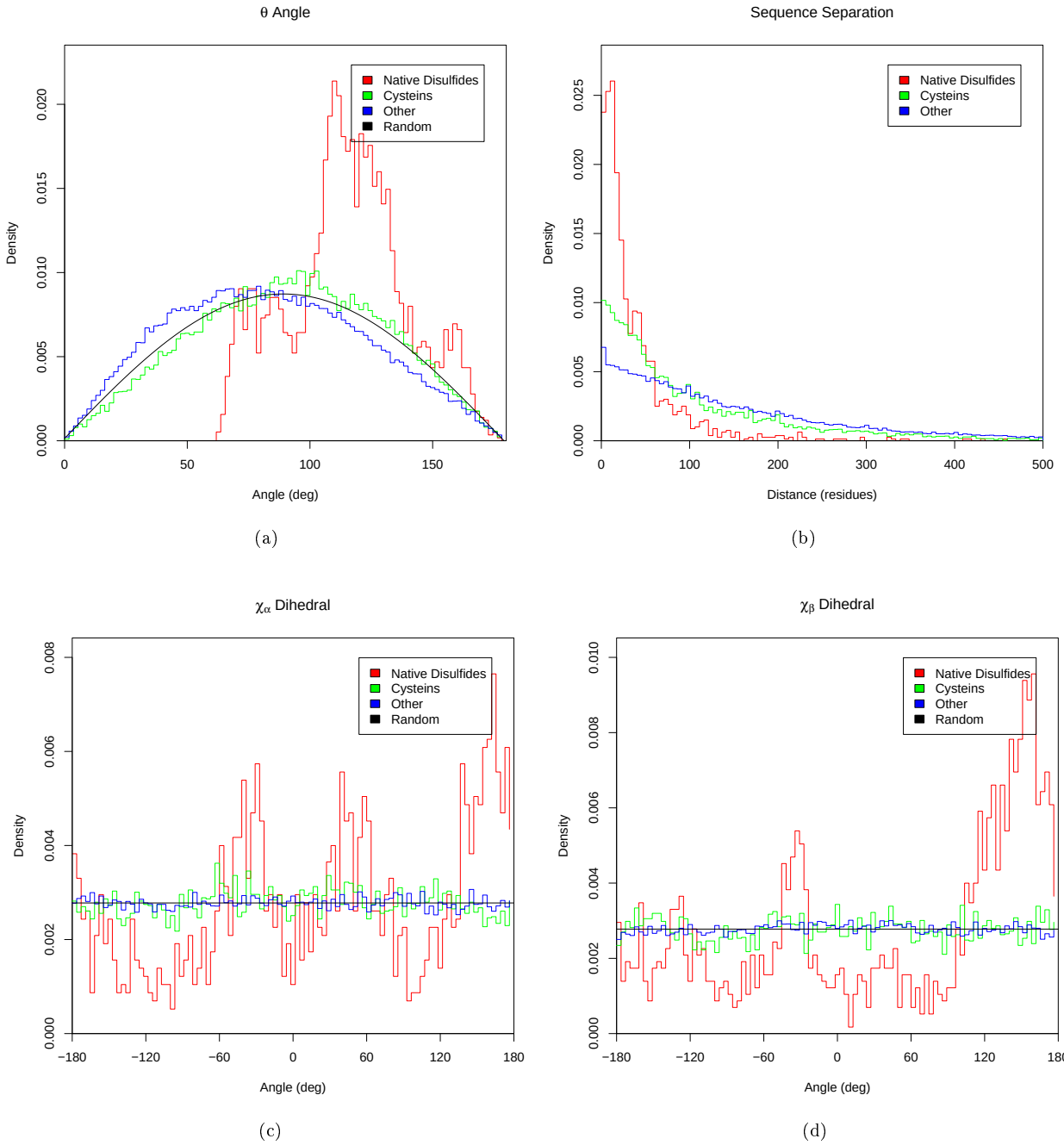


Figure 3: Distributions of distances for native disulfides, non-bonded cysteines, and other pairs of residues. (a) and (b) The $C\beta$-$C\beta$ distance distributions, with different scales. Black vertical bars represent distances corresponding to zero energy (See figure 2a). (c) (d) Similar histograms for centroid-centroid distances.



(a) (b)

(c) (d)

Figure 4: (a) Distribution of θ angles, along with distribution of angles expected for random orientations. (b) The sequence separation between residues. The (c)χ_α and (d) χ_β dihedral angles, along with the expected distribution of dihedral angles between random vectors.

3 Results

3.1 Distribution of geometries

Inspecting the distributions for the five centroid geometry terms (figures 3 and 4) shows the separation between native disulfides and the reference residue pairs. Disulfides fall into a narrow window of distance apart, with nearly all disulfide $C\beta$ atoms within 3.5 to 4.25Å apart.

The planar angle shows a clear difference between the distributions for disulfides and non-disulfides. Non-bonded residues approximate the expected distribution of angles between a fixed vector and random point in three dimensional space. Binning the angles into bins of width ϵ, we expect the bin starting at angle θ to be given by

$$\frac{1}{2\epsilon}\left(cos(\theta) - cos(\theta + \epsilon)\right).$$

Non-disulfide bonded pairs correspond very closely to this expected distribution (figure4a). In contrast, in native disulfides this angle is restricted to between 60 and 180 degrees with a strong peak around 120 degrees.

The dihedral angles also show a deviation from the random distribution for native disulfides. Disulfides seem to prefer backbone orientations of approximately -30, 50, and 160 degrees. Interestingly, these peaks do not correspond to the minima of the χ_α score, which are located at -130, -30, and 150. The χ_β distribution also does not correspond well to the score, showing a preference for +150 and a smaller peak at -40, in comparison with the scoring minima at -30, 150, and to a lesser extent -150.

3.2 Classification as disulfides

A key property of a scoring term is its ability to discriminate between disulfides and non-disulfides. A good score should maximize the separation between the scores for disulfides and for non-disulfides. Figure 5 plots this separation.

The scores for $C\beta$ and centroid distance show a strong separation between native disulfides and other residue pairs (figures 5a and 5b). In both cases nearly all the non-disulfide pairs of residues are scored close to 5.0 based solely on the distance between them.

Score	Sensitivity	Specificity
$C\beta$ distance	0.91	0.99
Centroid distance	0.71	0.99
θ angle	0.68	0.90
χ_α dihedral	0.42	0.91
χ_β dihedral	0.26	0.90
Total	0.75	.99

Table 1: Sensitivity and specificity of the scores. A pair of residues is considered as a disulfide if it scores below zero by more than a slight amount. Sensitivity is the fraction of native disulfides with negative scores. Specificity is the fraction of all non-disulfide residues with a positive score.

The angle score distributions are not so clearly separated (figures 5c, 5d and 5e). Since distant residues are weighted toward zero, much of the non-disulfide bonded pairs have scores around zero. The θ angle score varies between -.6 and .6 within the valid range for disulfides, but increases to 100 for orientations where the second $C\beta$ is aligned behind the first in a position that would clash with the backbone atoms. These extremely high scores lead to a long tail on the reference distribution. The native disulfides for all three angular scores are spread across the range of possible scores rather than clustering below zero. This calls into question the ability of the angular terms to discriminate disulfide bonds.

A clearer indicator of the strength of the scores as discriminator functions is given by the sensitivity and specificity of the scores (table 1). A score of below zero was considered a positive indication that the two residues are disulfide bonded. A perfect discriminator would give sensitivity of 1.0, indicating that every native disulfide was correctly assigned a negative score, and a specificity of 1.0, indicating that all the pairs not natively bonded had scores of zero or more.

All the tests were quite specific, correctly rejecting over 90% of the non-bonded pairs. This is partly due to the abundance of non-bonded pairs in our validation set, leading to the immediate rejection of clear

cut cases. The $C\beta$ distance is a very effective discriminator, with both high sensitivity and specificity. The angular terms are not as successful. The dihedral angles even have sensitivities below 0.5, indicating that they have more false negative identifications than true positives. The total term can be viewed as a hyperplane discriminator in a five-dimensional space. The slope of this hyperplane is determined by the weighting of the terms when they are summed. In this case all scores were weighted equally, so the plane lies orthogonal to the vector $(1, 1, 1, 1, 1)$. This choice yields a scoring function that has good specificity and reasonable sensitivity.

3.3 Non-disulfide bonded cysteines

Although the differentiation of disulfides from non-disulfides was the primary purpose of the analysis, the availability of data comparing non-disulfide bonded cysteines to other residue pairs shows several intriguing results. For the majority of the measurements the distribution of cysteine pairs align very well with the distribution of other residue pairs. However, the distance between residues tends to be shorter (figure 3a) and the sequence separation tends to be smaller (figure 4b). The two may be correlated, since the expected distance between the ends of an idealized polymer chain under Brownian motion is proportionate to the number of monomers in that chain.

4 Discussion

The scoring terms presented here seem to fit the data reasonably well, but the original training data seems to have held slightly different biases from the more recent PISCES set. It may not have held as many large proteins as this set. This would explain the false negatives in the distance scores, especially the centroid distance. The number of proteins in the PDB has nearly tripled since 2002, so it is not surprising that some bias existed in the set available to train the scores.

The efficiency of centroid mode is needed during the coarse search for minima. Due to the large movements across the energy landscape, the chances of making a disulfide bond with optimal geometries is slim. It is more important to have a broad energy decrease such that the potential disulfide bond will be detected in the coarse search. Thus a liberal scoring function may even be desirable in situations where disulfide bonds are desirable for their stabilizing properties.

While a liberal scoring function may be acceptable, it should not be biased. Any bias away from native disulfides could make ROSETTA predict disulfides where none would exist in nature.

5 Conclusion

6 Acknowledgments

Jacob Corn provided erudite advice as my postdoc mentor. David Baker was a tip-top PI. Bill Schief designed the original scoring functions. Thanks to Dominik Gront for advice on disulfide bond conformations, and to Justin Ashworth, Andrew Leaver-Fay, Oliver Lange, and Rob Vernon for their extensive programming advice.

References

[1] IGOR Pro.

[2] S E Brenner, P Koehl, and M Levitt. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res*, 28(1):254–6, Jan 2000.

[3] Rhiju Das and David Baker. Macromolecular modeling with Rosetta. *Annual review of biochemistry*, 77:363–82, Jan 2008.

[4] A G Murzin, S E Brenner, T Hubbard, and C Chothia. SCOP: a structural classification of proteins database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247(4):536–40, Apr 1995.

[5] J S Richardson. The anatomy and taxonomy of protein structure. *Adv Protein Chem*, 34:167–339, Jan 1981.

[6] J M Thornton. Disulphide bridges in globular proteins. *Journal of Molecular Biology*, 151(2):261–87, Sep 1981.

[7] Chu Wang, Philip Bradley, and David Baker. Protein-protein docking with backbone flexibility. *Journal of Molecular Biology*, 373(2):503–19, Oct 2007.

[8] Guoli Wang and Roland L Dunbrack. PISCES: recent improvements to a PDB sequence culling server. *Nucleic Acids Res*, 33(Web Server issue):W94–8, Jul 2005.

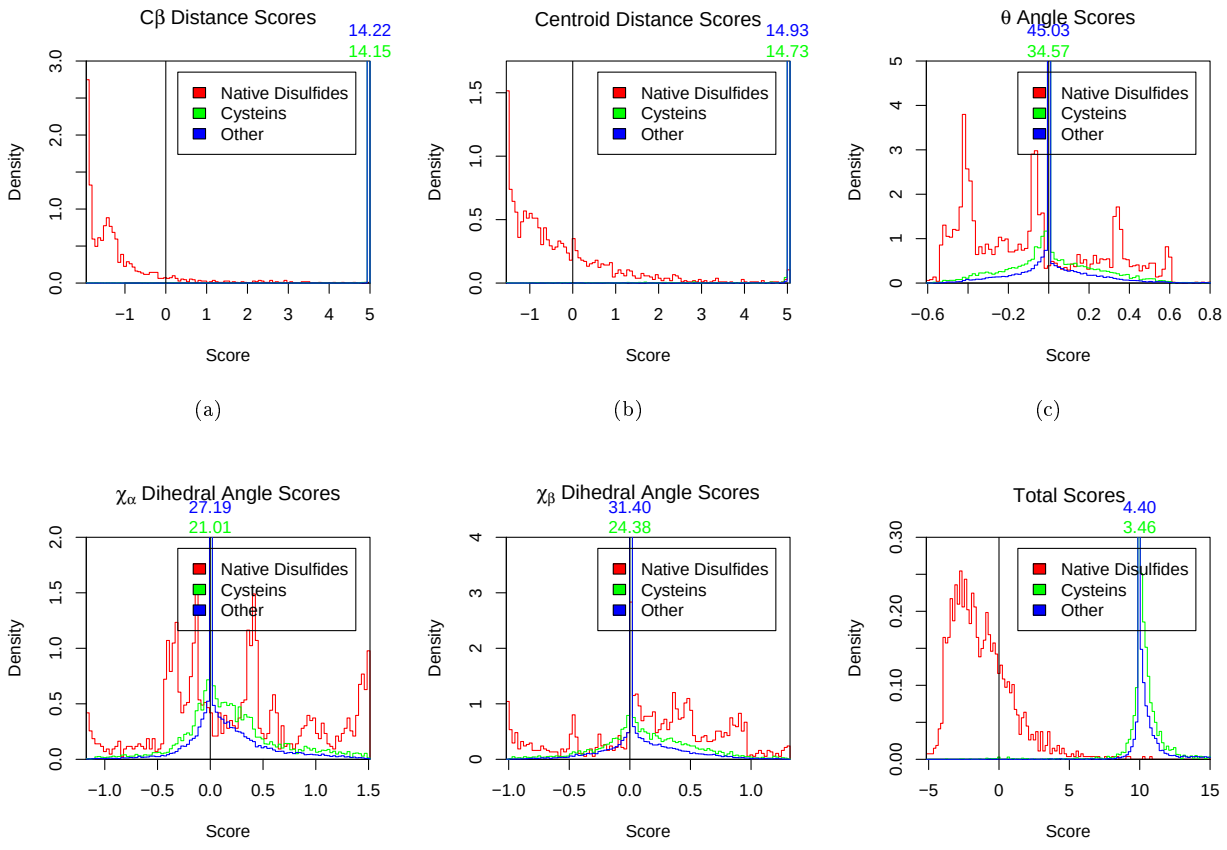


Figure 5: Distribution of scores. The vertical axis has been scaled such that the distribution of native disulfides is clear. Densities which fall outside this range are labeled above the graph margin. The (a) $C\beta$ and (b) centroid distance scores. Most non-disulfide residue pairs are far apart and so receive a score near 5. (c) Planar angle scores. The sharp clustering of non-disulfide pairs towards zero comes from the large number of these residues with weights near zero. A long tail reaches +100 from pairs where one residue lies directly behind the other, but has too little density to be visible. The dihedral angles in (d) and (e) show the same clustering of non-disulfide pairs around zero scores as the θ angle. (f) A sum of all scores. Scores are unweighted except for the θ angle, which is halved to compensate for its separate evaluation for both residues.