Geometric preferences of crosslinked protein-derived cofactors reveal a high propensity for near-sequence pairs

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · APRIL 2005

Impact Factor: 2.63 · DOI: 10.1002/prot.20403 · Source: PubMed

2 AUTHORS, INCLUDING:

David E Benson
Calvin College
28 PUBLICATIONS **1,674** CITATIONS

SEE PROFILE

Geometric Preferences of Crosslinked Protein-Derived Cofactors Reveal a High Propensity for Near-Sequence Pairs

Marla D. Swain and David E. Benson*
Department of Chemistry, Wayne State University, Detroit, Michigan

ABSTRACT **Protein-derived cofactors that are composed of covalently crosslinked amino acid side chains are of increasing importance in protein science. These crosslinked protein-derived cofactors (CPDC) are formed either through direct oxidation by metal/O_2-derived intermediates or through outer sphere oxidation by highly oxidizing cofactors. CPDCs that are formed by outer sphere oxidation do not require side-chain precursors to be coordinated by a metal center, and therefore are more difficult to identify than those formed by direct oxidation. To better understand the propensity for CPDC formation by outer sphere oxidation, the geometrical preferences of CPDCs were examined. The Dezymer algorithm has been used to identify all putative CPDC-forming mutations in 500 proteins. Geometrically, although chemically unrelated, these CPDCs were found to be similar to disulfide-bonded cysteine pairs. Additionally, the percentage of near-sequence pairs (i and i +1 to i and i + 5) increased as the average C_α–C_α distance between the amino acid pairs increased. This survey also examined the protein databank for proteins with pre-attack conformations for CPDCs, using non-bonded contacts reported by Procheck. A total of 323 unique proteins was identified, with 55 being near-sequence amino acid pairs. The high geometric propensity of near-sequence amino acid pairs for forming CPDCs is significant due to difficulties associated with detection by structural or mass spectrometric methods. Proteins 2005;59:64–71.** © 2005 Wiley-Liss, Inc.

Key words: cross-linking (including S—S bonds); computational analysis of protein structure; protein-derived cofactors; protein design; structural bioinformatics

INTRODUCTION

Protein-derived cofactors form a structurally novel biological cofactor class that has become increasingly important over the past decade.[1–7] Crosslinked protein derived cofactors (CPDCs) are a subset of protein-derived cofactors that contain post-translationally generated covalent bonds between amino acid side chains,[4,7,8] where the side chains are typically aromatic (Fig. 1). Many of these CPDCs perform one-electron redox chemistry in concert with associated mononuclear redox-active metal centers or amine dehydrogenation.[3,4,9] Biogenesis mechanisms for these cofactors have recently begun to be reported where initial side chain oxidation occurs by either a coordinated metal/O_2 adduct[10,11] or outer sphere electron transfer to a highly oxidizing cofactor.[12–15] To date, twelve types of CPDCs have been documented in approximately twenty proteins, primarily by high-resolution crystallography. CPDCs that are not coordinated to a metal center can be difficult to identify in the absence of well-resolved crystallographic and detailed mass spectrometric information. For example, mass spectrometric and high-resolution crystallography has recently shown not only a cysteine–tryptophylquinone (CTQ) CPDC but also two aspartate–cysteine (DC) CPDCs and a glutamate–cysteine (EC) CPDC in a bacterial amine dehydrogenase.[12,13,16] Mass spectrometry and peptide sequencing analysis of proteolytic digests for these proteins have assisted in confirming CPDC formation.[15–18] In the absence of tandem mass spectrometric (MS/MS) peptide sequencing experiments,[17] only far-sequence CPDCs can be identified by observation of a mass fragment comprised of two disparate proteolytic fragments.[15,17]

Mechanistic insights into CPDC biogenesis have begun to be uncovered. Tyrosine–cysteine (YC) cofactor biogenesis in galactose oxidase requires the presence of dioxygen, a copper ion,[10] and the formation of the Cu^I oxidation state.[11] A detailed examination of YC biogen-

Abbreviations: CPDC, crosslinked protein-derived cofactor; YC tyrosine–cysteine pair bonded between $C_{\epsilon 1/2(Tyr)}$–$S_{\gamma(Cys)}$; HC, histidine–cysteine pair bonded between $C_{\epsilon 1(His)}$–$S_{\gamma(Cys)}$; HY, histidine–tyrosine bonded between $N_{\epsilon 2/\delta 1(His)}$–$C_{\epsilon 1/2(Tyr)}$; EC, glutamate–cysteine bonded between $C_{\gamma(Glu)}$–$S_{\gamma(Cys)}$; DC, aspartate–cysteine bonded between $C_{\beta(Asp)}$–$S_{\gamma(Cys)}$; HY_β, histidine–tyrosine bonded between $N_{\epsilon 2/\delta 1(His)}$–$C_{\beta(Tyr)}$; CC, disulfide pair bonded between $S_{\gamma(Cys1)}$–$S_{\gamma(Cys2)}$; WYM, tryptophan–tyrosine–methionine trimer from two bonds between $C_{\eta 2(Trp)}$–$C_{\epsilon 1(Tyr)}$ and $S_{\delta(His)}$–$C_{\epsilon 2(Tyr)}$; LTQ, lysyltopaquinone bonded between $C_{\delta 1/2(Tpq)}$–$N_{\zeta(Lys)}$; CTQ, cysteine–tryptophylquinone bound between $C_{\epsilon 3(Trq)}$–$S_{\gamma(Cys)}$ and TTQ, tryptophan–tryptophylquinone $C_{\epsilon 3(Trq)}$–$C_{\delta 1(Trp)}$; MS/MS, tandem mass spectrometry; EPR, electron paramagnetic resonance

The Supplementary Materials referred to in this article can be found at http://www.interscience.wiley.com/jpages/0887-3585/suppmat

Grant sponsor: The Research Corporation, RI1131. Grant sponsor: National Institutes of Health; Grant number: 5R25GM058905-06 (M.G.S.). Grant sponsor: Wayne State University.

*Correspondence to: David Benson, Department of Chemistry, 5101 Cass Ave., Wayne State University, Detroit, MI 48202. E-mail: dbenson@chem.wayne.edu

Received 11 August 2004; Accepted 29 October 2004

Published online 4 February 2005 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.20403

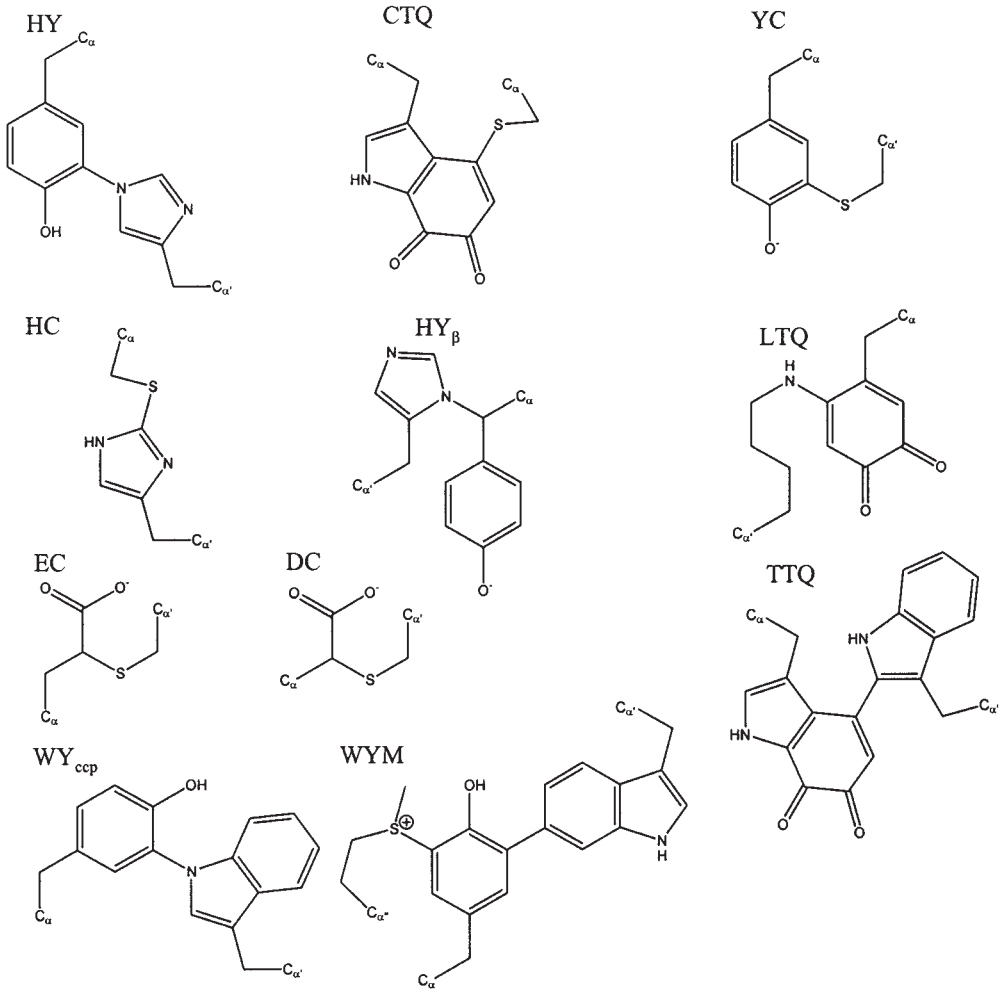
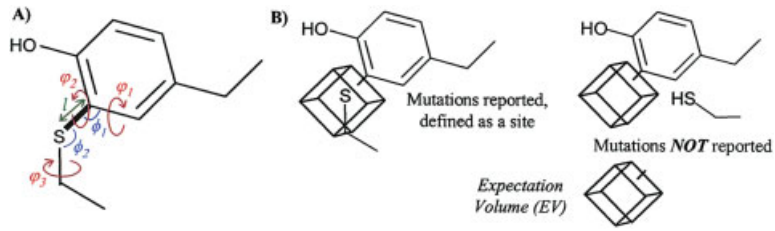


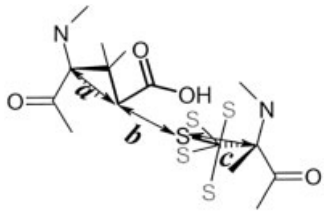
Fig. 1. Chemical structures of reported CPDCs. The crosslinked bond is highlighted: YC, tyrosine–cysteine pair bonded between $C_{\epsilon1/2(Tyr)}$–$S_{\gamma(Cys)}$;[1] HC, histidine–cysteine pair bonded between $C_{\epsilon1(His)}$–$S_{\gamma(Cys)}$;[28,29] HY, histidine–tyrosine bonded between $N_{\epsilon2\delta1(His)}$–$C_{\epsilon1/2(Tyr)}$;[30,31] EC, glutamate–cysteine bonded between $C_{\gamma(Glu)}$–$S_{\gamma(Cys)}$;[12,13] DC, aspartate–cysteine bonded between $C_{\beta(Asp)}$–$S_{\gamma(Cys)}$;[12,13] HY$_\beta$, histidine–tyrosine bonded between $N_{\epsilon2/\delta1(His)}$–$C_{\beta(Tyr)}$;[27] WYM, tryptophan–tyrosine–methionine trimer from two bonds between $C_{\eta2(Trp)}$–$C_{\epsilon1(Tyr)}$ and $S_{\delta(Met)}$–$C_{\epsilon2(Tyr)}$;[19,20] LTQ, lysyltopaquinone bonded between $C_{\delta1/2(Tpq)}$–$N_{\xi(Lys)}$;[63] CTQ, cysteine–tryptophylquinone bound between $C_{\epsilon3(Trq)}$–$S_{\delta(Cys)}$;[12,13] and TTQ, tryptophan–tryptophylquinone $C_{\epsilon3(Trq)}$–$C_{\delta1(Trp)}$.[32]

esis kinetics and solvent isotope kinetics suggest that either thiyl radical or phenoxyl radical formation mediates bond formation.[11] Both of these mechanisms involve inner sphere electron transfer from tyrosine 272 or cysteine 228 to a $[CuO_2]^+$ moiety. Inner sphere oxidative initiation of CPDC biogenesis can only account for the CPDCs that are coordinated to Cu centers (YC, HY and HC) or to a heme iron (HY$_\beta$). The rest of the CPDCs shown in Figure 1 are not coordinated but adjacent to a metal center. Recent CPDC-containing peroxidase structures[19–21] have been reported in which both CPDCs contain the conserved distal tryptophan, which is well known to be oxidized by a compound I intermediate in cytochrome *c* peroxidase.[22] The involvement of outer-sphere electron transfer in the biogenesis has been proposed for CTQ, DC and EC by an accessory protein[12,13] with sequence homology to the S-adenosyl methionine-dependent radical-generating enzyme family.[23] This intermolecular electron transfer mechanism is further suggested by the finding that MauG, the accessory protein for tryptophan–tryptophylquinone (TTQ) biogenesis in methylamine dehydrogenase is homologous to diheme class cytochrome *c* peroxidases.[14,15] In both methylamine dehydrogenase and quinohemoprotein amine dehydrogenase, it is reasonable that the pre-attack conformation of the precursor residues (two tryptophans for TTQ, cysteine and tryptophan for CTQ, apartate and cystiene for DC, and glutamate and cysteine for EC) are at the end of an electron transfer relay, such as thiyl radical formation in the di-iron containing ribonucleotide reductase enzymes.[9,24] If it were not for the EPR signal, irreversible hydrazine derivative inhibition and quinone redox cycling of CTQ,[25] the DC and EC cofactors could remain uncharacterized.



Scheme 1. *SiteSearch* method for assessing geometric viability of mutations. (a) The crosslinked bond between an amino acid pair was described by a bond length (l), two bond angles ($\phi_1$and ϕ_2) and three dihedral angles (ϕ_1, ϕ_2 and ϕ_3). Constraints used for each product CPDC conformation are included in the supplementary materials. (b) If the atom that forms the crosslink in the second amino acid satisfies these constraints (represented by the expectation volume), then the pair of mutations is classified as a geometrically viable site (left); otherwise the mutation set is not recorded (right).

In this article, the geometric constraints of product and pre-attack conformations are reported. These calculations are relevant to future CPDCs formed by outer-sphere amino acid side chain oxidation. A high geometric propensity for near-sequence CPDCs is found as a result of amino acid pairs within the same secondary structure element. Identification of these near-sequence CPDCs requires tandem mass spectrometric analysis of proteolytic digests. This article also provides additional proteins for which CPDC biogenesis could potentially occur under oxidative stress conditions.



Scheme 2. Mean extension distance. The distance between the crosslinking atom and the backbone C_α, averaged over all sampled rotamers, were determined for both amino acids in the CPDC. These average backbone-crosslinking atom distances (*a* and *c*) were added to the bond length (*b*) of the crosslinking bond to provide a benchmark for CPDC extension.

MATERIALS AND METHODS

CPDC Product Conformation Calculation

The SiteSearch algorithm of the Dezymer package[26] was used to identify all putative CPDC-forming mutations. Geometric constraints for inter-side chain covalent bond formation in each CPDC were taken from known crystal structures[1,12,13,19–21,27–32] and model compounds[33,34] (Supplementary Table I). Six constraints were necessary to describe each bond: atomic separation, two angles and three dihedral angles (Scheme I). A standard rotamer library,[35] sampled every 10°, was used to describe potential side-chain conformations. Since backbone movement was not allowed, potential backbone–side chain clashes were evaluated but side chain–side chain clashes were not. For a particular protein, all putative CPDC-forming mutations were identified within 2–30 min on a Linux-loaded PC (800 MHz, AMD). Mean extension distances (Scheme II) for each CPDC (Table I) were calculated as the sum of the mean C_α-crosslinked atom distances for each residue comprising the CPDC and the CPDC crosslink bond length.

A library of geometrically viable CPDC-forming mutations was generated with 500 proteins. Protein structures selected for this study had high-resolution structures, no posttranslational modifications and divergent backbone conformations.[36] Structures were completely hydrogenated using the Reduce algorithm[37] before Dezymer calculations. This set of protein structures has been used previously for the analysis of side chain conformations.[35] These structures represent the major classes of protein structures, as defined by the structural classification of proteins.[38]

The geometric viability of this CPDC-forming mutant library was tested. One hundred sites were randomly selected for minimization to verify that these predicted mutants were geometrically reasonable. The NAMD program[39] was used for energy minimization with a CHARMM-based all-atom force field,[40] patched with CPDC-specific force fields. Mutation sets were introduced to an all-alanine (excluding proline or glycine residues) scaffolds representing the parental proteins to discount side chain–side chain interactions. After minimization, 99 of the 100 selected sites had structures similar to those of the parent proteins. Residue sequences, three-dimensional structures, and geometric parameters for each mutation were archived as text files. These files were parsed using awk (Linux) or Excel (Windows).

Geometric Formation Propensity Calculations

The number of predicted mutation sets in a given protein (sites per protein) was determined for each CPDC. Sites per protein were normalized to the number of residues in the protein (formation propensity) to standardize results between large and small proteins. Formation propensity means and root mean square deviations of these populations (Table I) are reported.

Local Topology Calculations

The Dezymer package was used to calculate $C_{\alpha1}$–$C_{\alpha2}$ atom distances of each CPDC mutation pair. Means and root mean square deviations of these distributions were tabulated (Table I). The number of sites for a particular C_α–C_α separation, binned at 0.1 Å resolution, are graphically presented [Fig. 2(a)].

TABLE I. Statistics of Virtual CPDC Library Analysis

	Sampling		Global		Distance	
CPDC	Possible conformations	Mean extended length	Total sites	Formation propensity	C_α–C_α distance	% i, i + (1–5)
LTQ	10587804	10.7 ± 1.2Å	48379	0.623 ± 0.005	7.1 ± 2.7Å	33.3%
HC	9425	8.9 ± 0.3Å	25094	0.328 ± 0.003	6.8 ± 2.6Å	41.6%
YC	8200	9.0 ± 0.4Å	30056	0.389 ± 0.003	6.3 ± 2.5Å	45.4%
HY	123656	10.3 ± 0.4Å	15246	0.201 ± 0.002	6.1 ± 2.5Å	55.8%
CTQ	8800	8.0 ± 0.4Å	10237	0.135 ± 0.002	5.5 ± 2.3Å	30.6%
CC	625	7.6 ± 0.1Å	8926	0.108 ± 0.002	5.6 ± 2.4Å	26.2%
EC	12800	7.2 ± 0.4Å	3343	0.052 ± 0.001	5.5 ± 2.3Å	20.1%
TTQ	123904	8.9 ± 1.1Å	2779	0.039 ± 0.001	5.7 ± 2.4Å	41.6%
HY_β	16211	6.9 ± 0.4Å	1633	0.028 ± 0.001	5.9 ± 2.4Å	24.6%
DC	575	6.1 ± 0.1Å	137	0.008 ± 0.001	4.8 ± 2.2Å	9.5%

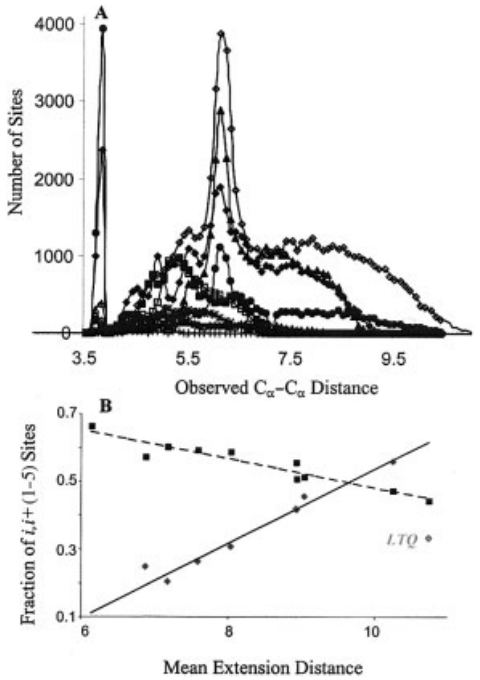


Fig. 2. Near-sequence CPDCs increase for more extended CPDCs. (a) A histogram of the observed extension distances: YC, closed diamonds; CTQ, closed squares; HC, closed triangles; HY, closed circles; LTQ, open diamonds; CC, open squares; TTQ, open triangles; EC, *; DC, +; HY_β, −. Two significant peaks are observed above these broad distributions at 3.7 and 6.2-6.3 Å. (b) The sum of sites with sequence separations of i, i + 1 to i, i + 5 as a function of mean extension distance (diamonds). A linear fit to these data, excluding LTQ, is presented as a solid line (R^2 = 0.96). These data are compared to the fraction of available sites (≤90% of the observed extension distances) that are i, i + 1 to i, i + 5, along with a linear fit (dashed line, R^2 = 0.90).

Pre-attack Conformation Search of Protein Data Bank

Non-bonded contact files from the Procheck algorithm,[41] which was run for every X-ray structure in the Protein Data Bank,[42] were parsed by awk programs. Non-bonded contacts between atoms that comprise each CPDC were selected and any contacts that were less than 87% of the sum of the covalent atomic radii[43] were identified. Each amino acid pair identified was examined against the rest of the dataset for uniqueness: no mutant, ligand or isozyme structures were accepted. Lengths, angles and dihedral angles for the putative crosslink were compared to energy-minimized structures of CPDC-containing proteins without constraints for the CPDC bond in the force field. Amino acid pairs that significantly deviated from the constraints estimated by the minimized CPDC protein structures were excluded.

RESULTS AND DISCUSSION

Two computational procedures were used in this study to address the geometric propensity of CPDC formation. The SiteSearch subroutine of the Dezymer algorithm[26] was used to search 500 proteins for amino acid mutations with the correct geometric potential for CPDC product formation. Analysis of this search will define what the dominant geometric constraints are for CPDC formation. Alternatively, the Protein Data Bank[42] was surveyed for amino acid pairs in reported X-ray resolved protein structures that are in pre-attack conformations for CPDC formation. These pre-attack conformations presume facile CPDC formation after outer-sphere oxidation of one or both of the amino acid side chains. Together these analyses will determine the geometric constraints of CPDC formation and the potential for proteins to form CPDCs by outer sphere oxidation.

Product Conformation Geometry Survey

Library Construction

Predicted mutations for this study were generated using the SiteSearch algorithm of the Dezymer package.[26] This algorithm uses a hard-sphere model with no input from electronic structure besides those introduced by geometrical preferences (Scheme I). Constraints for inter-side chain covalent bond formation in each CPDC were taken from known crystal structures (Supplementary Table I). Dihedral angles for aliphatic bonds were allowed complete rotational freedom but were constrained for aromatic bonds to maintain planarity of the crosslinked atom with the aromatic ring system. These predicted CPDC-forming mutation sets, termed sites, were analyzed for the frequency with which CPDC-forming mutations occurred in a given protein (formation propensity) and the distance between C_α atoms of each predicted CPDC. Due to the

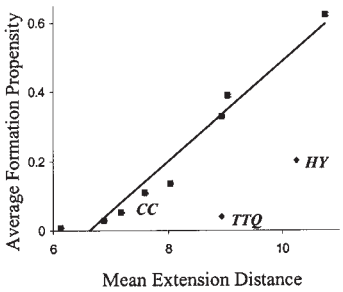


Fig. 3. Geometric formation propensity of product CPDC conformations. The mean formation propensity is plotted *versus* mean extension distance. A line has been drawn from a least squared fitting ($R^2 = 0.95$) of all points except TTQ and HY.

ubiquitous nature of the disulfide bonds, Cys–Cys (CC)* was included in our calculations as a benchmark. The predicted CC dataset is consistent with previous computational studies of disulfide bonds.[44–46]

Library Analysis

The formation propensities of five CPDCs (LTQ, YC, HC, HY and CTQ) were found to be higher than the formation propensity of the CC dataset (Fig. 3, Table I). Two factors contribute to formation propensity, the average backbone distance and the hybridization of the atoms forming the CPDC bond. As the distance between the backbone and the CPDC crosslink increases, higher frequencies of predicted sites are observed (Fig. 3). We used a linear estimate of the average backbone separation (mean extension distance, Scheme II) for these correlations. The number of predicted sites is dramatically suppressed for two CPDCs (TTQ and HY). Both of the atoms forming the crosslinking bond are sp^2 hybridized in TTQ and HY and are therefore conformationally limited by our geometric parameters that enforce planarity. A linear dependence was found for CPDCs, excluding TTQ and HY, between the mean extension distance and total predicted sites or formation propensity (Fig. 3, $R^2 = 0.95$). The comparison with the log of sampled conformations (Table I) was much less correlated ($R^2 = 0.28$), suggesting that rotameric space for these CPDCs was adequately sampled. Based only on geometric criteria, LTQ, YC, HC, HY and CTQ have a higher geometrical, but not chemical, propensity for crosslink formation relative to a disulfide bonds.

In order to assess any local topological effects on CPDC formation, the distances between the two C_α atoms of these CPDCs were analyzed (Fig. 2). The observed mean C_α atom distances were shorter than the calculated mean extension distances (Table I). A moderate compression of observed mean extension distances to the calculated mean extension distance was observed at longer distances. Distributions of observed extension distances for each CPDC [Fig. 2(a)] revealed broad distributions with variable amounts of two distinctive peaks at 3.8 and 6.2 Å. Three CPDCs (HY, TTQ and YC) had more than 5% of their sites with an observed extension distance of 3.8 Å, which appears as a distinct peak in Figure 2(a). Each of these sites arose from adjacent amino acids (i, $i + 1$ positions). The secondary structure elements, defined by Molscript,[47] of these sites were either intra-helical (HY and YC) or intra-coil (TTQ). Three CPDCs (HC, LTQ and HY) had over 5% of their sites with an observed extension distance of 6.2 to 6.3 Å, which also form a distinct peak in Figure 2(a). Two-thirds of these sites had intra-helical residues at the i, $i + 4$ positions and the remaining sites had intra-strand residues at the i, $i + 2$ positions. Overall, a high percentage of near-sequence CPDCs were observed.

The percentage of near-sequence sites increased as the mean extension distance increased [Fig. 2(b)], except for LTQ. Within the i, $i + (1–5)$ sites, populations for different sequence separations varied, which reflected the different secondary structure preferences (Supplementary Table II). The high fraction of near-sequence, defined for our purposes as i, $i + (1–5)$, CPDC-forming mutations is statistically expected. The encounter frequencies of near-sequence positions, without the Dezymer constraints, were estimated by determining the percentage of near-sequence amino acid combinations with C_α–C_α separations smaller than the C_α–C_α separation observed for 90% of each Dezymer predicted CPDC population. These values were high but decreased as the C_α–C_α separation increased [66% to 44% n, squares in Fig. 2(b)]. The Dezymer predicted near-sequence CPDC-forming population increased as the C_α–C_α separation increased [10% to 55%, diamonds in Fig. 2(b)]. This comparison shows that the predicted increase in geometric propensity for near-sequence CPDCs is not due to an increased availability of near-sequence sites.

Pre-attack Conformation Geometric Analysis

Due to the geometric propensity for product CPDCs to adopt conformations similar to disulfide bonds, all of the structures in the Protein Data Bank[42] were analyzed for amino acid pairs in pre-attack conformations. The Procheck algorithm[41] was used to search all non-bonded contacts in all of the protein structures in the protein databank. The Procheck-based search is much more rapid than the Dezymer-based search, since Dezymer explores mutations at all positions and all allowed side-chain dihedral angles. Non-bonded contact files were searched by awk programs for specific atomic contacts (*e.g.* Tyr–$C_{\epsilon 1/2}$ and Cys–S_γ for YC) with less than or equal to 2.8-3.2 Å separation. Cut-off atomic separation values were 87% of the combined covalent radii[43] of the crosslinking atoms of each CPDC. A total of 323 unique proteins with these contacts are reported in Supplementary Table III and presented with respect to sequence separation in Figure 4. Over fifty of these CPDCs identified by non-bonded contacts were near-sequence pairs.

The propensity for CPDC forming mutations at near-sequence positions increased as the mean extension distance increased. The number of near-sequence positions available for selection is high, but decreases slightly as the mean extension length increases. The geometric propensity for near-sequence CPDCs, however, increases dramatically as the mean extension length increases, thus suggesting that this dependence is significant. The non-bonded

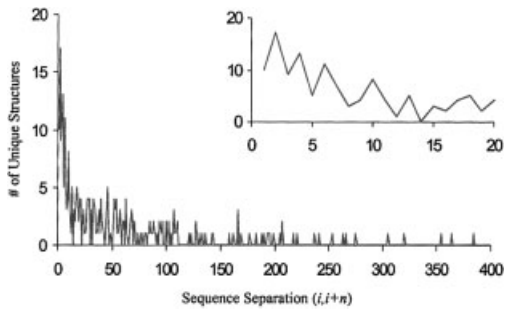


Fig. 4. Sequence position of pre-attack CPDC conformations identified in the Protein Data Bank. Inset is an expansion of the *i*, *i* + 1 to *i*, *i* + 20 region.

contact survey of Cys–Cys interactions was consistent with the low geometric propensity for near-sequence DC, EC and HYβ CPDCs: only 4% of disulfide-bonded cysteines were at near-sequence positions (1.9-2.1 Å filter) as opposed to 27% near-sequence reduced cysteine pairs (2.9-3.2 Å filter). Geometrically, near-sequence disulfide-bonded cysteine pairs are disfavored due to steric overlap between the side chain of one residue and the backbone of the adjacent residue. This geometric preference of near-sequence disulfide bonds is manifested chemically, in that a $^{+}$100 to $^{+}$150 mV reduction potential shift is observed for near-sequence position disulfides.[48] As the mean extension distance for an amino acid pair increases, steric overlaps with the backbone of the adjoining protein can be alleviated due to the increasing conformational degrees of freedom. Decreased side chain–backbone clashes allow for increased geometric propensity for near-sequence CPDC product conformations. As the side-chain conformational degrees of freedom increase further (e.g. LTQ), more conformations will be able to bridge between secondary structural units, decreasing the geometric propensity for near sequence CPDCs. Another aspect of this conclusion (since disulfide bonds decrease the entropy of the unfolded state to stabilize folded protein conformations)[49] is that only the DC and EC CPDCs should be expected to significantly stabilize protein structure. The high geometric propensity for near-sequence CPDCs that are more extended comes from decreased steric clashes with backbone atoms.

The propensity for near-sequence CPDCs can be seen in existing structures, from the product conformation search and from the non-bonded contact search. The survey of known CPDC-containing proteins shows that more than 20% are near-sequence CPDCs: HY, *i*, *i* + 4 intrahelix;[30,31] CTQ, *i*, *i* + 5 helix-coil;[12,13] and HC, *i*,*i* + 2 intracoil in hemocyanin[29,50] and *i*, *i* + 2 in tyrosinase.[50] A recent report of advantageous CPDC formation between Trp51–Tyr52 in H52Y cytochrome *c* peroxidase[21] continues this trend. The geometric survey of product CPDC conformations demonstrated that the propensity for near-sequence CPDCs increases as the covalently crosslinked atoms are farther from the protein backbone. For the HY cofactor, 55% of the product conformation predictions were at near-sequence positions. Finally, the search for pre-attack CPDC conformations in reported structures from the Protein Data Bank showed 55 near sequence amino acid pairs out of 323 unique amino acid pairs (17%). These analyses suggest that near sequence CPDCs will continue to constitute a significant portion of existing CPDCs.

Obviously, chemical constraints on CPDC biogenesis must also be considered, which are more demanding than disulfide bond formation. The biogenesis of the TTQ, CTQ, DC and EC cofactors suggests that side-chain oxidation occurs through outer sphere electron transfer, due to the lack of any coordinated metal centers.[12,13,15,32] The amino acid pairs identified by the non-bonded contact search, or mutations identified in the library of 500 protein structures, would require outer sphere electron transfer to a highly oxidizing intermediate of a transition metal center within 15 Å. The formation of such oxidized amino acid radicals has been well-documented in heme[51–55] and di-iron[24] enzymes. Surface tyrosyl radicals are typically formed in the presence of excess hydrogen peroxide, which leads to dimerization *in vivo*,[52,56] or potentially to stabilization of interprotein complexes *in vitro*.[57] Surface tyrosyl radicals are also well known products of reactive oxygen species,[58] which are precursors to protein degradation[58] or amyloidogenesis.[59] Cyclooxygenase-2 (COX-2) has been implicated in peroxide-dependent amyloidogenesis,[60] and three COX-2 structures (PDBIDs: 1CX2, 4COX and 5COX)[61] were observed in the non-bonded contact search. The S58-inhibited structure showed a non-bonded contact between Tyr 55 $C_{\epsilon 1}$–Cys 57 S_{γ} (3.1-3.2 Å), and two other structures (one with indomethachin bound) showed a non-bonded contact between Tyr 495 $C_{\epsilon 1}$–Met 501 S_{δ} (3.0-3.1 Å). COX-2 surface tyrosyl radicals have been recently shown to mediate polymerization of human amyloid-β and to form irreversibly associated COX-2–amyloid-β complexes.[60] If outer sphere oxidation of Tyr 55 or Tyr 495 in COX-2 occurs, then these putative crosslinks could serve as antioxidants to dissipate surface tyrosyl radicals. The long distance between these residues and the heme (26 Å, YM; 38 Å, YC) or the catalytically implicated Tyr 385 (29 Å, YC) is recognized; however, Tyr 495 is 13 Å from Tyr 504, which has been shown to be an alternate site for substantial unpaired electron density.[62] The case of putative YM and YC cofactors in COX-2 serves as an example of how, under excess levels of hydrogen peroxide (*i.e.* oxidative stress conditions), the amino acid pairs identified by non-bonded contacts may lead to more CPDC-containing proteins.

Considering both the propensity for outer-sphere oxidation-mediated CPDC biogenesis and the geometric propensity for near sequence CPDC formation, there could be many CPDC-containing proteins that are still unidentified. Additionally, near-sequence positions of CTQ,[16] HY[17] and HC[50] have complicated identification by protein sequencing. Here we have demonstrated that near-sequence CPDCs are likely to persist as more CPDCs are discovered. Due to the bias of one-dimensional mass spectrometric analysis against near-sequence CPDCs, the use of two-dimensional mass spectrometric analysis is necessary to further identify CPDCs. Here, we present an initial survey of known protein structures (Supplementary Table III)

that could potentially form CPDCs by an outer-sphere oxidation mechanism under oxidative stress conditions.

REFERENCES

1. Ito N, Phillips SEV, Stevens C, Ogel ZB, Mcpherson MJ, Keen JN, Yadav KDS, Knowles PF. Novel thioether bond revealed by a 1.7-Å crystal-structure of galactose-oxidase. Nature 1991;350:87–90.
2. McIntire WS, Wemmer DE, Chistoserdov A, Lidstrom ME. A new cofactor in a prokaryotic enzyme: tryptophan tryptophylquinone as the redox prosthetic group in methylamine dehydrogenase. Science 1991;252:817–824.
3. Okeley NM, van der Donk WA. Novel cofactors via post-translational modifications of enzyme active sites. Chem Biol 2000;7:R159–171.
4. Rogers MS, Dooley DM. Posttranslationally modified tyrosines from galactose oxidase and cytochrome c oxidase. Adv Protein Chem 2001;58:387–436.
5. Whittaker JW. Galactose oxidase. Adv Protein Chem 2002;60:1–49.
6. Klinman J. Commentary - how many ways to craft a cofactor? Proc Natl Acad Sci 2001;98:14766–14768.
7. Davidson VL. Pyrroloquinoline quinone (PQQ) from methanol dehydrogenase and tryptophan tryptophylquinone (TTQ) from methylamine dehydrogenase. Adv Protein Chem 2001;58:95–140.
8. Duine JA, Jongejan JA. Quinoproteins, enzymes with pyrroloquinoline quinone as cofactor. Annu Rev Biochem 1989;58:403–426.
9. Stubbe J, van der Donk WA. Protein radicals in enzyme catalysis. Chem Rev 1998;98:705–762.
10. Rogers M, Baron A, McPherson M, Knowles P, Dooley D. Galactose oxidase pro-sequence cleavage and cofactor assembly are self-processing reactions. J Am Chem Soc 2000;122:990-991.
11. Whittaker MM, Whittaker JW. Cu(I)-dependent biogenesis of the galactose oxidase redox cofactor. J Biol Chem 2003;278:22090–22101.
12. Datta S, Mori Y, Takagi K, Kawaguchi K, Chen Z, Okajima T, Kuroda S, Ikeda T, Kano K, Tanizawa K, Mathews F. Structure of a quinohemoprotein amine dehydrogenase with an uncommon redox cofactor and highly unusual crosslinking. Proc Natl Acad Sci 2001;98:14268–14273.
13. Satoh A, Kim JK, Miyahara I, Devreese B, Vandenberghe I, Hacisalihoglu A, Okajima T, Kuroda S, Adachi O, Duine JA, Van Beeumen J, Tanizawa K, Hirotsu K. Crystal structure of quinohemoprotein amine dehydrogenase from *Pseudomonas putida*. Identification of a novel quinone cofactor encaged by multiple thioether cross-bridges. J Biol Chem 2002;277:2830–2834.
14. Wang Y, Graichen ME, Liu A, Pearson AR, Wilmot CM, Davidson VL. MauG, a novel diheme protein required for tryptophan tryptophylquinone biogenesis. Biochemistry 2003;42:7318–7325.
15. Pearson AR, De la Mora-Rey T, Graichen ME, Wang Y, Jones LH, Marimanikkupam S, Agger SA, Grimsrud PA, Davidson VL, Wilmot CM. Further insights into quinone cofactor biogenesis: Probing the role of mauG in methylamine dehydrogenase tryptophan tryptophylquinone formation. Biochemistry 2004;43:5494–5502.
16. Vandenberghe I, Kim JK, Devreese B, Hacisalihoglu A, Iwabuki H, Okajima T, Kuroda S, Adachi O, Jongejan JA, Duine JA, Tanizawa K, Van Beeumen J. The covalent structure of the small subunit from *Pseudomonas putida* amine dehydrogenase reveals the presence of three novel types of internal cross-linkages, all involving cysteine in a thioether bond. J Biol Chem 2001;276:42923–42931.
17. Buse G, Soulimane T, Dewor M, Meyer HE, Bluggel M. Evidence for a copper-coordinated histidine-tyrosine cross-link in the active site of cytochrome oxidase. Protein Sci 1999;8:985–990.
18. Pearson AR, Jones LH, Higgins L, Ashcroft AE, Wilmot CM, Davidson VL. Understanding quinone cofactor biogenesis in methylamine dehydrogenase through novel cofactor generation. Biochemistry 2003;42:3224–3230.
19. Carpena X, Loprasert S, Mongkolsuk S, Switala J, Loewen PC, Fita I. Catalase-peroxidase KatG of *Burkholderia pseudomallei* at 1.7 Å resolution. J Mol Biol 2003;327:475–489.
20. Yamada Y, Fujiwara T, Sato T, Igarashi N, Tanaka N. The 2.0 Å crystal structure of catalase-peroxidase from *Haloarcula marismortui*. Nature Struct Biol 2002;9:691–695.
21. Bhaskar B, Immoos CE, Shimizu H, Sulc F, Farmer PJ, Poulos TL. A novel heme and peroxide-dependent tryptophan-tyrosine cross-link in a mutant of cytochrome c peroxidase. J Mol Biol 2003;328:157–166.
22. Sivaraja M, Goodin DB, Smith M, Hoffman BM. Identification by ENDOR of Trp191 as the free-radical site in cytochrome c peroxidase compound ES. Science 1989;245:738–740.
23. Sofia HJ, G. C, Hetzler BG, Reyes-Spindola JF, Miller NE. Radical SAM, a novel protein superfamily linking unresolved steps in familiar biosynthetic pathways with radical mechanisms: functional characterization using new analysis and information. Nucleic Acids Res 2001;29:1097–1106.
24. Stubbe J, Nocera DG, Yee CS, Chang MCY. Radical initiation in the class I ribonucleotide reductase: Long-range proton-coupled electron transfer? Chem Rev 2003;103:2167–2201.
25. Takagi K, Torimura M, Kawaguchi K, Kano K, Ikeda T. Biochemical and electrochemical characterization of quinohemoprotein amine dehydrogenase from *Paracoccus denitrificans*. Biochemistry 1999;38:6935–6942.
26. Hellinga HW, Richards FM. Construction of new ligand-binding sites in proteins of known structure .1. Computer-aided modeling of sites with predefined geometry. J Mol Biol 1991;222:763–785.
27. Bravo J, Verdaguer N, Tormo J, Betzel C, Switala J, Loewen PC, Fita I. Crystal structure of catalase HPII from *Escherichia coli*. Structure 1995;3:491–502.
28. Klabunde T, Eicken C, Sacchettini JC, Krebs B. Crystal structure of a plant catechol oxidase containing a dicopper center. Nature Struct Biol 1998;5:1084–1090.
29. Cuff ME, Miller KI, van Holde KE, Hendrickson WA. Crystal structure of a functional unit from Octopus hemocyanin. J Mol Biol 1998;278:855–870.
30. Ostermeier C, Harrenga A, Ermler U, Michel H. Structure at 2.7 angstrom resolution of the *Paracoccus denitrificans* two-subunit cytochrome c oxidase complexed with an antibody F-V fragment. Proc Natl Acad Sci 1997;94:10547–10553.
31. Yoshikawa S, Shinzawa-Itoh K, Nakashima R, Yaono R, Yamashita E, Inoue N, Yao M, Fei MJ, Libeu CP, Mizushima T, Yamaguchi H, Tomizaki T, Tsukihara T. Redox-coupled crystal structural changes in bovine heart cytochrome c oxidase. Science 1998;280:1723–1729.
32. Chen L, Doi M, Durley RC, Chistoserdov AY, Lidstrom ME, Davidson VL, Mathews FS. Refined crystal structure of methylamine dehydrogenase from *Paracoccus denitrificans* at 1.75 Å resolution. J Mol Biol 1998;276:131–149.
33. Baumann C, Brockelmann M, Fugmann B, Steffan B, Steglich W, Sheldrick WS. Pigments of fungi. 62. Haematopodin, an unusual pyrroloquinoline derivative isolated from the fungus *Mycenahaehaematopus agaricales*. Angew Chemie Int Ed 1993;32:1087–1089.
34. Schonenberger B, Jacobson AE, Brossi A, Streaty R, Klee WA, Flippen-Anderson JL, Gilardi R. Comparison of (−)-eseroline with (+)-eseroline and dihydroseco analogs in antinociceptive assays: confirmation of rubreserine structure by x-ray analysis. J Med Chem 1986;29:2268–2273.
35. Lovell SC, Word JM, Richardson JS, Richardson DC. The penultimate rotamer library. Proteins Struct Funct Gen Bioinfor 2000;40:389–408.
36. Richardson DC, Richardson JS. 500 Filtered Structures. http://kinemage.biochem.duke.edu/databases/top500.php; 2000.
37. Word JM, Lovell SC, LaBean TH, Taylor HC, Zalis ME, Presley BK, Richardson JS, Richardson DC. Visualizing and quantifying molecular goodness-of-fit: small-probe contact dots with explicit hydrogen atoms. J Mol Biol 1999;285:1711–1733.
38. 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 1995;247:536–540.
39. Kalé L, Skeel R, Bhandarkar M, Brunner R, Gursoy A, Krawetz N, Phillips J, Shinozaki A, Varadarajan K, Schulten K. NAMD2: Greater scalability for parallel molecular dynamics. J Comp Phys 1999;151:283–312.
40. MacKerell AD, Bashford D, Bellott M, Dunbrack RL, Evanseck JD, Field MJ, Fischer S, Gao J, Guo H, Ha S, Joseph-McCarthy D, Kuchnir L, Kuczera K, Lau FTK, Mattos C, Michnick S, Ngo T, Nguyen DT, Prodhom B, Reiher WE, Roux B, Schlenkrich M, Smith JC, Stote R, Straub J, Watanabe M, Wiorkiewicz-Kuczera J, Yin D, Karplus M. All-atom empirical potential for molecular modeling and dynamics studies of proteins. J Phys Chem B 1998;102:3586–3616.
41. Laskowski RA, Macarthur MW, Moss DS, Thornton JM. Procheck

- a program to check the stereochemical quality of protein structures. J Appl Cryst 1993;26:283–291.
42. 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.
43. Shrake A, Rupley JA. Environment and exposure to solvent of protein atoms. Lysozyme and insulin. J Mol Biol 1973;79:351–371.
44. Pabo CO, Suchanek EG. Computer-aided model-building strategies for protein design. Biochemistry 1986;25:5987–5991.
45. Petersen MT, Jonson PH, Petersen SB. Amino acid neighbours and detailed conformational analysis of cysteines in proteins. Protein Eng 1999;12:535–548.
46. Mas JM, Aloy P, Marti-Renom MA, Oliva B, de Llorens R, Aviles FX, Querol E. Classification of protein disulphide-bridge topologies. J Comp Aided Design 2001;15:477–487.
47. Kraulis PJ. MOLSCRIPT: a program to produce both detailed and schematic plots of protein structures. J Appl Cryst 1991;24:946–950.
48. Woycechowsky KJ, Raines RT. The CXC motif: a functional mimic of protein disulfide isomerase. Biochemistry 2003;42:5387–5394.
49. Goldenberg DP, Frieden RW, Haack JA, Morrison TB. Mutational analysis of a protein-folding pathway. Nature 1989;338:127–132.
50. Lerch K. Primary structure of tyrosinase from *Neurospora crassa*. II. Complete amino acid sequence and chemical structure of a tripeptide containing an unusual thioether. J Biol Chem 1982;257:6414–6419.
51. Fishel LA, Farnum MF, Mauro JM, Miller MA, Kraut J, Liu Y, Tan XL, Scholes CP. Compound I radical in site-directed mutants of cytochrome c peroxidase as probed by electron paramagnetic resonance and electron-nuclear double resonance. Biochemistry 1991;30:1986–1996.
52. Zhang H, He S, Mauk AG. Radical formation at Tyr39 and Tyr153 following reaction of yeast cytochrome c peroxidase with hydrogen peroxide. Biochemistry 2002;41:13507–13513.
53. Gibson JF, Ingram DJE. Free radical produced in the reaction of methemoglobin with hydrogen peroxide. Nature 1958;181:1398-1399.
54. Svistunenko DA, Dunne J, Fryer M, Nicholls P, Reeder BJ, Wilson MT, Bigotti MG, Cutruzzola F, Cooper CE. Comparative study of tyrosine radicals in hemoglobin and myoglobins treated with hydrogen peroxide. Biophys J 2002;83:2845–2855.
55. Fenwick CW, English AM. Trapping and LC-MS identification of protein radicals formed in the horse heart metmyoglobin-H_2O_2 reaction. J Am Chem Soc 1996;118:12236-12237.
56. Minetti M, Pietraforte D, Carbone V, Salzano AM, Scorza G, Marino G. Scavenging of peroxynitrite by oxyhemoglobin and identification of modified globin residues. Biochemistry 2000;39:6689–6697.
57. Amini F, Denison C, Lin HJ, Kuo L, Kodadek T. Using oxidative crosslinking and proximity labeling to quantitatively characterize protein-protein and protein-peptide complexes. Chem Biol 2003;11:1115–1127.
58. Hawkins CL, Davies MJ. Generation and propagation of radical reactions on proteins. Biochim Biophys Acta 2001;1504:196–219.
59. Atwood CS, Perry G, Zeng H, Kato Y, Jones WD, Ling KQ, Huang X, Moir RD, Wang D, Sayre LM, Smith MA, Chen SG, Bush AI. Copper mediates dityrosine cross-linking of Alzheimer's amyloid-β. Biochemistry 2004;43:560–568.
60. Nagano S, Huang X, Moir RD, Payton SM, Tanzi RE, Bush AI. Peroxidase activity of cyclooxygenase-2 (COX-2) cross-links β-amyloid (Ab) and benerates Aβ-COX-2 hetero-oligomers that are increased in Alzheimer's disease. J Biol Chem 2004;279:14673–14678.
61. Kurumbail RG, Stevens AM, Gierse JK, McDonald JJ, Stegeman RA, Pak JY, Gildehaus D, Miyashiro JM, Penning TD, Seibert K, Isakson PC, Stallings WC. Structural basis for selective inhibition of cyclooxygenase-2 by anti-inflammatory agents. Nature 1996;384:644–648.
62. Rogge CE, Liu W, Wu G, Wang LH, Kulmacz RJ, Tsai AL. Identification of Tyr504 as an alternative tyrosyl radical site in human prostaglandin H synthase-2. Biochemistry 2004;43:1560–1568.