Residue-Specific Contact Order and Contact Breadth in Single-Domain Proteins: Implications for Folding as a Function of Chain Elongation

ARTICLE *in* BIOTECHNOLOGY PROGRESS · JUNE 2008

Impact Factor: 2.15 · DOI: 10.1021/bp070475v · Source: PubMed

4 AUTHORS, INCLUDING:

Nese Kurt Yilmaz
University of Massachusetts Medical School
31 PUBLICATIONS **257** CITATIONS
SEE PROFILE

Bryan C Mounce
Institut Pasteur
13 PUBLICATIONS **61** CITATIONS
SEE PROFILE

Residue-Specific Contact Order and Contact Breadth in Single-Domain Proteins: Implications for Folding as a Function of Chain Elongation

Neşe Kurt, Bryan C. Mounce, Paul A. Ellison,† and Silvia Cavagnero*

Department of Chemistry, University of Wisconsin−Madison, 1101 University Avenue, Madison, Wisconsin 53706

Cotranslational protein misfolding and aggregation are often responsible for inclusion body formation during *in vivo* protein expression. This study addresses the relations between protein folding/misfolding and the distribution of intramolecular interactions across different regions of the polypeptide chain in soluble single-domain proteins. The sequence regions examined here include the C terminus, which is synthesized last in the cell. Emphasis is placed on two parameters reporting on short- and long-range interactions, i.e., residue-specific *contact order* (*RCO*) and a new descriptor of intramolecular protein interaction networks denoted as residue-specific *contact breadth* (*RCB*). *RCB* illustrates the average spread in sequence of the residues serving as interaction counterparts. We show that both *RCO* and *RCB* are maximized at the chain termini for a large fraction of single-domain soluble proteins. A direct implication of this result is that the C terminus of the polypeptide chain, which is synthesized last during ribosome-assisted translation, plays a key role in the generation of native-like structure by establishing long-range interactions and generating contacts with interaction counterparts widely distributed across the sequence. Comparison of our computational predictions with the experimental behavior of selected proteins shows that the presence and absence of large *RCO* and *RCB* at the chain termini correlates with the protein's ability to properly fold either after the C terminus has been synthesized or during chain elongation, respectively.

1. Introduction

A common problem in recombinant protein technology is the frequent inability to generate soluble native proteins. While the origins of inefficient expression, often associated with inclusion body formation, are not well understood, it is clear that protein misfolding and aggregation can take place both co- and post-translationally. In the former case, the absence of a given portion of primary structure may seriously affect the ability of a protein to properly fold, rendering it particularly prone to aggregation (*1*).

Protein folding to the bioactive native structure requires the establishment of a specific array of intramolecular interactions. During gene expression, proteins are synthesized vectorially from N to C terminus. The progression of intramolecular interactions as a function of chain elongation has been investigated experimentally by model systems consisting of purified N-terminal polypeptide fragments (reviewed in ref 2). In one of the earliest studies of this kind, it was reported that deletion of 13 amino acids from the C terminus of the 149-residue staphylococcal nuclease results in a compact unfolded protein lacking persistent secondary structure (*3*). Other *in vitro* studies highlighted the importance of the C-terminal portion of the sequence for correct folding by showing that deletion of just a few C-terminal amino acids results in the loss of native-like secondary and tertiary structure (*2*). Furthermore, C-terminal deletions extending to 20−25% of the sequence lead to incomplete burial of the hydrophobic core, assuming native-like interactions (*4*).

The above findings suggest that the polypeptide chain may not have the ability to fold in a native-like fashion in the cell until the C-terminal region has been synthesized. Before this happens, there is danger of protein aggregation during biosynthesis, and cellular defense mechanisms such as the chaperone systems play a vital role in maintaining homeostasis. One strategy proposed in recombinant protein technology has been to co-overexpress chaperones during protein production in *E. coli*. While this approach has sometimes proven successful to increase the yields of soluble protein (*5*), it is not of general applicability and it needs to be optimized for the specific system of interest.

Hence, a better understanding of the principles governing protein folding/misfolding in the cell is important to ultimately improve the quality of all the biotechnologically relevant processes that require correct *in vivo* folding.

One of the most poorly studied aspects of protein structure and folding is whether there are any peculiar trends in the distribution of native contacts across different regions of a protein sequence and whether any existing trends bear an impact on protein production in the cell. For instance, what is special about the contacts involving terminal residues in terms of their ability to promote folding cooperativity? How is polypeptide conformation affected by the lack of these contacts? Is the spatial and sequence distribution of native contacts, including those involving the C terminus, important to decipher how proteins fold and misfold cotranslationally?

Here, we explore the distribution of intramolecular contacts established by the terminal and intermediate regions of the protein sequence in single-domain native structures. The analysis is carried out by the computational evaluation of parameters termed residue-specific contact order (*RCO*) (*6*) and residue-

* To whom correspondence should be addressed. Tel: 1-608-262-5430. Fax: 1-608-262-9918. Email: cavagnero@chem.wisc.edu.

† Present address: Department of Chemistry, University of California, Berkeley, CA 94720.

10.1021/bp070475v CCC: $40.75
Published on Web 05/10/2008

specific contact breadth (*RCB*). We find that the chain termini engage preferentially in long-range contacts and contacts with regions of the protein chain widely distributed across the sequence. Specific analysis of *RCO* and *RCB* for four experimentally studied proteins shows that the presence and absence of large *RCO* and *RCB* at the chain termini correlates with the protein's ability to properly fold either only after the C terminus has been synthesized or earlier on during chain elongation, respectively. Therefore, we propose that the *RCO* and *RCB* parameters are useful tools to estimate the propensity of proteins of known structure to exhibit chain-length-dependent folding. Finally, we anticipate that proteins characterized by a high *RCO* and *RCB* at the C terminus, together with a high nonpolar character (*4*), are particularly at risk of exhibiting aggregation during biosynthesis, especially in the absence of an enhanced chaperone support machinery.

2. Methods

2.1. Protein Structure Selection and Calculations. Protein structure analysis was performed on proteins whose structures are deposited in the protein data bank (PDB) (*7*) and selected according to the following criteria: (1) single-domain water-soluble proteins of at least 40 residues, (2) 3-dimensional structure solved by X-ray crystallography with $\leq$1.8 Å resolution and $\leq$0.3 *R*-factor, (3) no cocrystallized cofactors or other small molecules, (4) no amino acids beyond the 20 naturally occurring residues, (5) wild-type species, i.e., no mutants, (6) no complexes, fragments, or membrane proteins, (7) spatial coordinates reported for all heavy atoms, and (8) single-domain proteins with no multiple CATH codes (*8*). The PDB files containing protein structures conforming to the above criteria were automatically compiled via the PDB-REPRDB web server ((*9*), 06/10/05 update). The PDB-REPRDB analysis was followed by manual elimination of structures with more than one CATH identifier and multidomain proteins with 5-digit PDB files. The resulting set of 48 protein structures (which also includes the four proteins discussed in Section 3.6) span a wide range of topologies, corresponding to over 90% of the known protein structure topologies found in nature. The PDB codes of all the selected proteins are provided in Table 1.

The atomic coordinates were read from the PDB files via a computer program written in FORTRAN. Subsequent analyses involving calculating *RCO* and *RCB* values according to eqs 1 and 2 were performed using FORTRAN and MATLAB (The MathWorks, Inc., Natick, MA), respectively.

Two residues were considered to be in contact if they had any heavy atoms (i.e., all atoms other than H) closer than 6 Å in space, according to their PDB file coordinates. Neighboring contacts between amino acids closer than 4 residues along the sequence were not included in the calculations to eliminate local contacts due to helical secondary structure formation.

2.2. Definition of Chain Termini and Plot Types. We used three different definitions for the termini, namely, 10%, 20%, and 30%, indicating the respective percentages of the total chain length defined as either N or C terminus. The use of multiple cutoff percentages is aimed at testing whether the results depend on the definition of terminus. Average *RCO* and *RCB* values were calculated for the N-, C-terminal, and intermediate (i.e., midchain) regions of the protein sequence and plotted as three-point score plots for each member of the selected single-domain protein set (Figures 1 and 2).

The three-point plots were divided into four categories according to the relative score of the three chain regions. In case the values monotonically increase from N to C terminus, the plot is defined as *slope-up*; if they decrease, the plot is *slope-down*. When the midchain value is the maximum, the plot is defined as a *peak* plot. In contrast, the plots where the midchain value is the lowest across the protein chain are defined as *valley*-shaped.

Table 1. PDB Codes, Number of Residues and CATH Codes for Selected Single-Domain Proteins Used in the Analysis Presented[a]

PDB code	no. of residues	CATH code	PDB code	no. of residues	CATH code
153L	185	1.10	1SHG	57	2.30
1AGI	125	3.10	1TEN	90	2.60
1AKO	268	3.60	1TML	286	3.20
1BD8	156	1.25	1UDH	228	3.40
1BJ7	150	2.40	1UTG	70	1.10
1BNR	110	3.10	1VIE	60	2.30
1CEX	197	3.40	1VXF	153	1.10
1CHD	198	3.40	1WAB	212	3.40
1CRN	46	3.30	1WHI	122	2.40
1CTF	68	3.30	1XNB	185	2.60
1DBS	224	3.40	2CI2	65	3.30
1DHN	121	3.30	2END	137	1.10
1EY0	136	2.40	2ERL	40	1.20
1FNA	91	2.60	2GDM	153	1.10
1GVP	87	2.40	2HVM	273	3.20
1HKA	158	3.30	2PTH	193	3.40
1HYP	75	1.10	2RN2	155	3.30
1IFC	131	2.40	2RTA	121	2.40
1IGD	61	3.10	2SIL	381	2.120
1KNB	186	2.60	2TGI	112	2.10
1LIB	131	2.40	3LZM	164	1.10
1NAR	289	3.20	3PTE	347	3.40
1NPK	150	3.30	3VUB	101	2.30
1PPN	212	3.90	4LZT	129	1.10

[a] CATH codes denote protein secondary structure types (*8*).

2.3. Calculation of *RCO* and *RCB* for a Random Contact Ensemble. As a reference, a hypothetical set composed of an ensemble of protein structures with random intramolecular contacts was constructed. Each residue pair in the set has the same chance to be involved in a contact. The average of such an ensemble corresponds to an effective contact map where all possible contacts are present. While this scenario is not physically attainable by individual polymer chains, it is a viable first-order approximation for an ensemble of distinct random chains. *RCO* and *RCB* values were computed for each residue according to eqs 1 and 2. An ensemble of random protein chains with an average length of 153 residues was considered. This size corresponds to the average chain length of the actual proteins in the selected single-domain protein database. Contacts between neighboring amino acids with less than 4 residue separation along the sequence were not included in the calculations. The *RCO* and *RCB* values for the residues belonging to the terminal and intermediate portions of the sequence were averaged and plotted as shown in Figure 3.

In order to further confirm the results, we have also performed similar *RCB* calculations on a series (1,000 and 10,000 to ensure convergence) of individual contact maps with completely random intramolecular contacts. Each map was imposed to have 10% of all possible contacts. This value corresponds to the average percent of intramolecular contact in the selected set of 48 single-domain protein structures. *RCB* values arising from the individual random maps were then averaged. As expected, identical *RCB* values were obtained by both methods.

3. Results and Discussion

3.1. Both N and C Termini Have High Relative Contact Order. The contact order (*CO*) is a widely used parameter whose value is a measure of the long-range character of

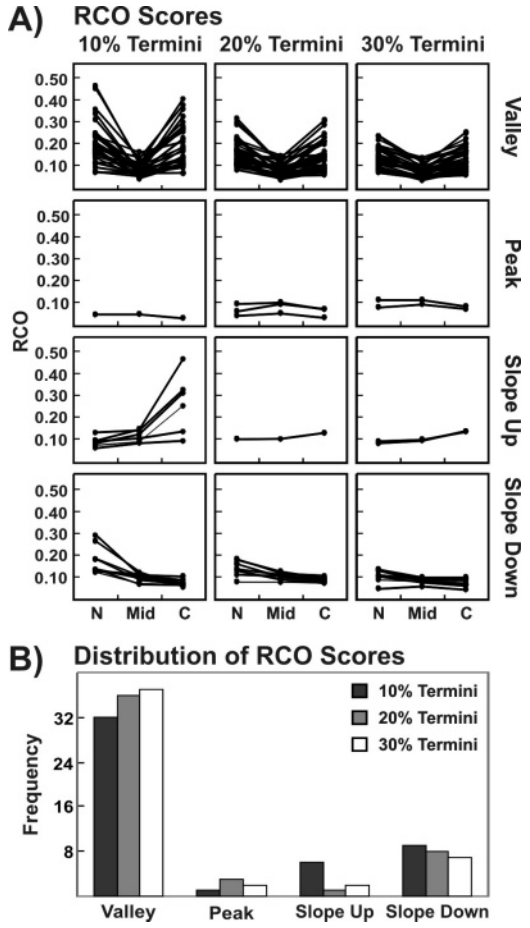

Figure 1. Contact order analysis of protein structures, analyzed by three-point score plots reporting the average residue-specific contact order, *RCO*, for three regions of the protein chain. The horizontal axis labels denote the N-terminal (N), midchain (Mid), and C-terminal (C) regions of the sequence. (A) Score plots are divided into four categories depending on the relative position of the three points. Results are provided for three different definitions of the chain termini. (B) Distribution of average *RCO* values into four different categories based on the score plot shapes. About 70% of the score plots are *valley*-shaped.

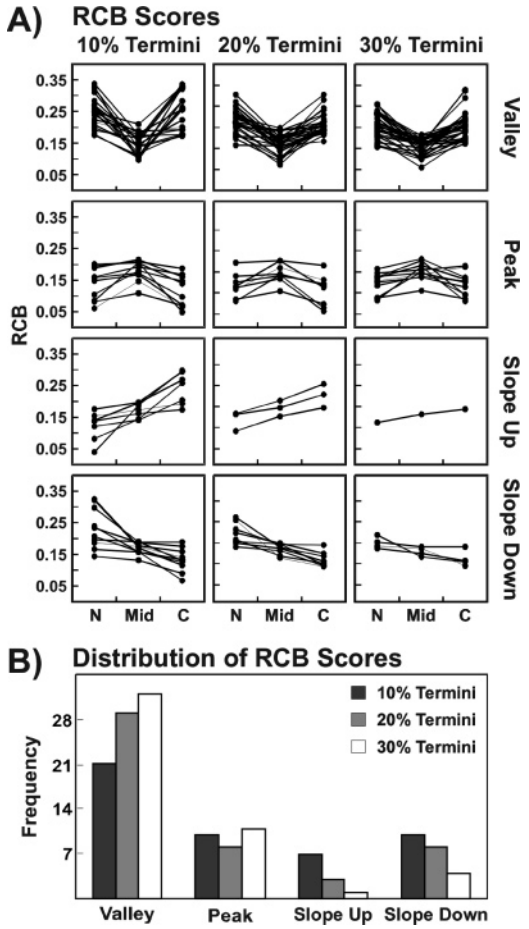

Figure 2. Contact breadth analysis of protein structures, displayed through three-point score plots illustrating the average residue contact breadth *RCB* of the three regions of the protein chain. Horizontal axis labels indicate N-terminal (N), midchain (Mid), and C-terminal (C) regions of the sequence. (A) Score plots are divided into four categories according to the relative position of the three points. Results are provided for three different definitions of the chain termini. (B) Block diagram illustrating the score plot distribution for each shape category. About 48% of the *RCB* score plots are *valley*-shaped (averaged across all definitions of termini).

intramolecular interactions in specific structures, weighted by the number of contacts between residue pairs (*6*). The *CO* has been primarily employed in the context of protein folding, both in its original (*6*) and modified (*10*, *11*) formulations on a per residue basis. We use here a modified version of *CO*, i.e., the residue-specific contact order *RCO*. This parameter is an extension of the original *CO*, except that it is evaluated for each residue *i* instead of the whole protein. The residue-specific contact order for the i^{th} residue is defined as

$$RCO_i = \frac{1}{L \times N_i} \sum_{j=1}^{L} N_{ij} \Delta S_{ij} \delta_{ij}$$

$$\text{where } \delta_{ij} = \begin{cases} 1 & \text{if } i \neq j,\ i \text{ and } j \text{ are in contact} \\ 0 & \text{otherwise} \end{cases} \quad (1)$$

where ΔS_{ij} is the residue separation between contacting residues *i* and *j* along the sequence, *L* is the number of amino acids in the protein, N_{ij} is the number of contacts between residues *i* and *j*, and N_i is the total number of contacts for residue *i*.

The average *RCO* value for N-terminal, midchain, and C-terminal portions of the protein sequences are displayed as three-point score plots in Figure 1A. Remarkably, most proteins in the database have *valley*-shaped *RCO* profiles, reflecting the fact that both C and N termini have high relative contact order. A strong dominance of *valley*-shaped score plots is found for all termini definitions used. *Valley* plots comprise 60%, 73%, and 77% of the proteins for 10%, 20%, and 30% termini, respectively.

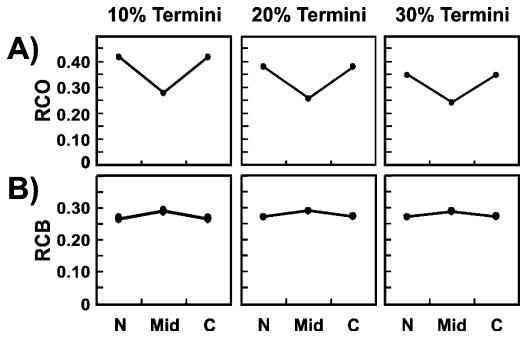

Figure 3. Computed (A) contact order and (B) contact breadth profiles for a hypothetical ensemble of protein structures with completely random contacts.

The predominance of *valley*-shaped score plots shows that the intramolecular interactions involving both termini are of long-range nature for the majority of proteins. There is a possibility that the high contact order observed for the termini

is related to the fact that most single-domain proteins have spatially close N and C termini in the folded state (*12*). Contacts between the two termini are expected to increase *RCO* considerably, as the sequence separation (ΔS) along the chain is large. However, *valley*-shaped *RCO* plots are not exclusively found in proteins with close N and C termini (data not shown).

Hence, even though the N- and C-terminal regions do not establish a larger number of contacts than the rest of the chain, they are engaged in more interactions with long-range character. As testified by experimental results on truncated N-terminal protein fragments (*2*), such contacts must be important to enable distal parts of the protein chain to participate in the formation of a stable structure.

3.2. Spread of Contact Partners across Polypeptide Sequence: Contact Breadth. As seen in the previous section, contact order is a useful reporter of the long-range character of intramolecular contacts and an important signature of single-domain protein structure. On the other hand, this parameter does not provide any information on another important characteristic of protein architecture, i.e., the degree to which contact partners for each residue spread across the polypeptide sequence. An important consideration in investigating the quality of a residue's contacts is whether the contacts are made with only a certain part of the chain, or with many different regions. A residue with diverse contact partners is central to the structure, as it links different parts of the chain upon folding.

To explore whether portions of the polypeptide chain, including the C terminus, are characterized by any characteristic trends involving the sequence-spreading of intramolecular contacts, we defined a new parameter denoted here as residue-specific contact breadth (*RCB*). *RCB* is designed to illustrate the degree to which each residue makes contacts with different parts of the polypeptide chain, and it is defined for residue *i* as

$$RCB_i = \frac{1}{L}\sqrt{\frac{\sum_{j=1}^{n}(\overline{R} - R_j)^2}{n-1}} \quad \text{where } \overline{R} = \sum_{j=1}^{n} R_j/n, \quad (2)$$

where R_j is the residue number for the amino acids contacting residue *i*, $\overline{R}$ is the average over all R_j values, and *n* is the number of contact partners for residue *i*. The above equation is equivalent to the standard deviation for the residue numbers of an amino acid's contact partners, normalized by the protein chain length. If a given amino acid contacts residues from many different parts of the protein chain, its *RCB* is high. On the contrary, if it interacts with only one portion of the chain, *RCB* is low.

The latter is valid even if the contacts are very long-range in nature, in which case *RCO* values are high. The terminal regions may have high *RCO* value simply because they contact each other, but they do not have high *RCB* unless they contact residues from different parts of the chain. Therefore, *RCO* and *RCB* are, in principle, totally independent parameters.

RCB values for terminal and intermediate chain regions were generated for all the proteins and collected in Figure 2A. Strikingly, about half of the plots are *valley*-shaped, similarly to the case of the *RCO* plots of Figure 1. The above demonstrates that the contacts involving the terminal regions are not only long-range but also involve diverse portions of the chain.

3.3. Comparison with an Ensemble of Random Contacts. In order to test whether the observed predominance of valley-shaped *RCO* and *RCB* plots are a simple consequence of generic collapsed polymer chains (*13*), we have generated three-point score plots for a hypothetical ensemble of protein chains with random intramolecular contacts. The results are shown in Figure 3.

Because of the inherent nature of linear polymer chains, the termini have the opportunity to establish contacts with residues most widely separated in sequence. As a result of this intrinsic bias, *RCO* plots for the random ensemble are *valley*-shaped. It is important to note, however, that the plots for the actual single-domain protein structures in our set (Figure 1A) deviate significantly from the random case. Some proteins display deeper *valley*-shaped plots and some proteins do not display *valley* shapes at all.

The *RCB* plot for the randomly contacting ensemble is nearly flat and slightly peak-shaped (Figure 3). This is in striking contrast with the dominance of significantly steeper *valley*-shaped plots found in the set of actual protein structures (Figure 2).

Comparison of the single-domain protein set with expected random chain behavior reveals two different scenarios for *RCO* and *RCB*. In the case of *RCO*, the preferential involvement of the chain termini in long-range interactions suggested by Figure 1 is partially due to the inherent properties of linear polymer chains, given that *valley*-shaped plots are expected even for an ensemble of randomly contacting collapsed polymers. On the other hand, the ability of the chain termini to establish intramolecular contacts with diverse regions of the protein chain described by *RCB* is a specific property of single-domain protein structures.

3.4. Relations with Protein Topology. An important question at this juncture is whether the trends identified in the previous sections bear any relationship with protein topology. The *RCO* and *RCB* parameters were taken as representative of the observed trends. We found that there is no apparent correlation between topology and tendency to fall into a particular *RCO* and *RCB* score plot category.

However, the α-β roll (CATH code 3.10), α-β barrel (CATH code 3.20), and 3-layer sandwich (CATH code 3.40) topologies are an exception in that they have a ca. 20% higher than average probability to generate *valley*-shaped *RCB* plots, especially in the case of the 30% definition for the termini. Proteins with the above three topologies constitute 27% of the structures analyzed here. All the α-β roll and barrel proteins analyzed in this work have closely spaced termini with several N-to-C intramolecular interactions. In contrast, proteins with 3-layer sandwich topology do not always follow this trend and sometimes have far apart termini. In general, termini in close contact are not a strict requirement for large *RCB* and *RCO* values at the chain termini, although the two properties are usually correlated.

3.5. Role of Protein Chain Length. We specifically tested whether *RCO* and *RCB* depend in any way on protein size. Average *RCO* and *RCB* were calculated for each protein chain, thus obtaining one specific descriptor for each member of the database. As shown in Figure 4, average *contact orders* and *contact breadths* do not significantly correlate with protein chain length ($r^2 = 0.30$ and 0.20, respectively, and p values $= 0.57$ and 0.27, respectively, with a sample size of 48). Similar trends were detected for the average *RCO* and *RCB* evaluated solely across the N- and C-terminal regions. The weak protein size dependence of the average *contact order* is consistent with previous reports (*14*).

The negligible dependence of contact breadth on protein size reveals that the extent of sequence spreading of the network of

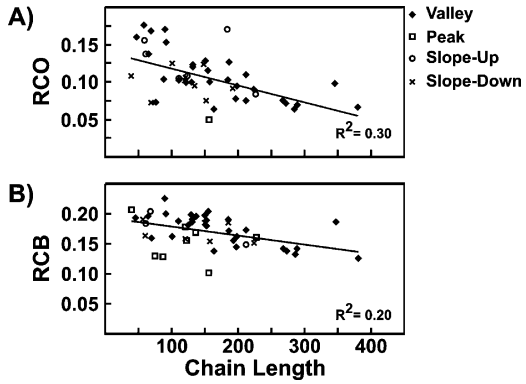

Figure 4. Dependence of average (A) contact order (*RCO*) and (B) contact breadth (*RCB*) on protein chain length (i.e., total the number of amino acids).

intramolecular interactions is essentially independent of protein chain length. Interestingly, all six longest proteins (270−370 residues) of our database have *valley*-shaped *RCB* and *RCO* plots, regardless of the termini definition. In general, longer proteins tend to have very complex topologies with intertwining helices and sheets. These specific topologies give rise to the termini spreading their contacts over diverse portions of the polypeptide chain, leading to higher *RCB* and *RCO* values in the terminal regions. Therefore, the chain termini are in contact-dense portions of the chain in longer proteins. The presence of long-range and sequence-widespread contacts at the termini may serve the purpose to facilitate holding a long protein chain folded in a single globular conformation.

3.6. Proteins Whose Chain Elongation Has Been Studied Experimentally. In addition to the selected set of proteins, we specifically examined the properties of the only four proteins whose *in vitro* folding has been systematically studied experimentally (*1*, *15*−*19*) and computationally (*2*, *4*) as a function of polypeptide chain elongation from the N to the C terminus. This set includes apomyoglobin (apoMb), barnase (bar), staphylococcal nuclease (SNase), and chymotrypsin inhibitor 2 (CI2). The *RCO* and *RCB* score plots for all four proteins are shown in Figure 5. The 20% definition of chain termini was chosen here for simplicity, as it represents the average behavior of the system (in between the 10% and 30% definitions). The *RCO* and *RCB* score plots (Figure 5) indicate that three out of four proteins (CI2, apoMb, and barnase), exhibit *valley*-shaped plots for *RCO* and *RCB*. In contrast, staphylococcal nuclease (SNase) displays *peak*-type plots. (*12*)

The above analysis implies that SNase has a different type of intramolecular contact network relative to the other proteins. Interestingly, SNase is the only protein in the set able to progressively develop local native-like secondary structure as its polypeptide chain elongates (*18*, *19*). Both barnase and CI2 develop secondary and tertiary structure only during the final stages of chain elongation (*15*−*17*), corresponding to incorporation of the last ∼30% of the amino acid sequence. ApoMb generates non-native self-associated β-strand structure as its chain elongates, due to its high nonpolar amino acid content (*4*).

Given that incomplete N-terminal protein chains lack the C-terminal residues, the properties of the native contacts involving this region are expected to be relevant for the chain length dependence of folding. We therefore speculate that the C-terminal regions of barnase and CI2 may be able to develop compact native structure near chain termination due to their high *RCO* and *RCB* values at the C terminus. Similarly, we propose that apoMb acquires the ability to become native-like and

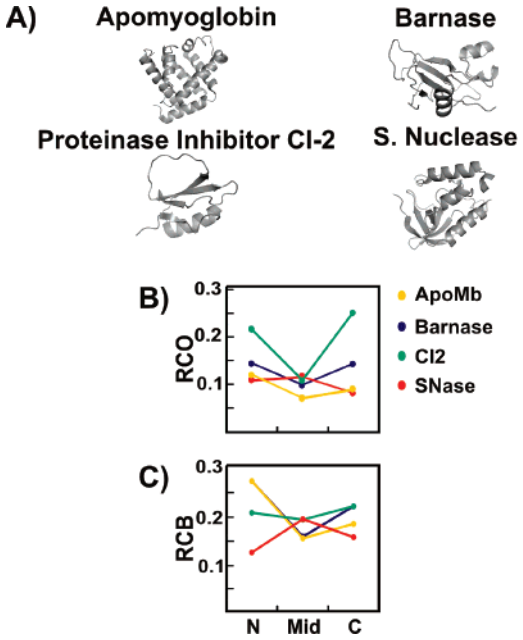

Figure 5. (A) Three-dimensional structure of the four proteins whose structural variations as a function of chain elongation have been studied experimentally. (B) *RCO* and (C) *RCB* score plots for the proteins in panel A.

monomeric upon addition of the last amino acids due to its high *RCO* and *RCB* values for the C terminus.

In summary, the data in Figure 5 are consistent with the fact that large *RCO* and *RCB* at the chain termini correlate with the ability of the C-terminal region of certain proteins to establish the contacts necessary to generate the native structure as chain elongation (from N to C terminus) gets completed.

The above results suggest that evaluating *RCO* and *RCB* profiles for any given protein is a useful aid in the prediction of protein folding as a function of chain elongation. In case the C-terminal region has relatively small *RCO* and *RCB* values (relative to the other regions), it is likely less essential for the overall structure formation. Hence, the polypeptide chain may have the ability to develop native-like interactions as it elongates. As seen (Figure 3A), high C-terminal *RCO* values are expected even for random polymer chains. On the other hand, relatively high *RCB* values for the C terminus are more diagnostic and may indicate that the chain does not have the ability to start folding before biosynthesis and release from the ribosome are complete. Therefore we propose computing *RCB* to assist in evaluating the requirement for the C terminus to establish the majority of intramolecular contacts necessary for native-like structure formation.

Despite the intriguing correlations between the calculations of Figure 5 and the experimental behavior of the four proteins, it is extremely important to note that a more detailed comparative analysis between computations and experiments is highly desirable before the above criteria can be regarded as having an absolute predictive power. This task will be facilitated in the future, as the chain length-dependent folding behavior of additional proteins becomes available.

As we previously showed, the overall nonpolar character of the N-terminal incomplete protein chains plays an important role in self-association (*4*) as it can drive aggregate formation due to hydrophobic interactions. Proteins with a high nonpolar content are expected to greatly benefit from enhanced support of the cellular machinery such as cotranslationally active chaperones. Therefore, the combined evaluation of *RCO*, *RCB*,

and nonpolar content (as in ref 4) as a function of chain elongation may also be useful to evaluate the balance between intramolecular folding and protein aggregation as a function of chain elongation. Existing algorithms to predict aggregation propensities of full-length proteins (*20*−*24*) have already proven useful to evaluate the likelihood of post-translational self-association.

4. Conclusions

This work introduces *RCO* and *RCB* as useful parameters to evaluate the nature of protein intramolecular contacts. *RCO* and *RCB* are often, but not exclusively, maximized at the chain termini. This result implies that the C terminus, which is generated during the final stages of ribosome-assisted polypeptide chain elongation, is important for the generation of native-like structure. The C terminus contributes to both establishing long-range interactions involving the chain termini and to generating contacts with interaction counterparts widely distributed across the sequence. We propose *RCO* and *RCB* are useful to evaluate the role of the C terminus in protein folding. These parameters are of general applicability and easy to calculate for any specific protein of interest.

Acknowledgment

This research was supported by the National Science Foundation (grants 0544182 and 0215368), the Milwaukee Foundation (Shaw Scientist Award to S.C.), and the Research Corporation (Research Innovation Award to S.C.). B.C.M. and P.A.E. were supported by Research Experience for Undergraduates (REU) funds from the National Science Foundation.

Notation

L	number of amino acids in the protein
N_i	total number of contacts for residue i
n	number of contact partners for a given residue
R_j	residue number for the contacting amino acid j
RCB	residue-specific contact breadth
RCO	residue-specific contact order
ΔS_{ij}	residue separation between contacting residues i and j along the sequence

References and Notes

(1) Chow, C.; Chow, C.; Rhagunathan, V.; Huppert, T.; Kimball, E.; Cavagnero, S. The chain length dependence of apomyoglobin folding: structural evolution from misfolded sheets to native helices. *Biochemistry* **2003**, *42* (23), 7090−7099.

(2) Cavagnero, S.; Kurt, N. Folding and misfolding as a function of polypeptide chain elongation: conformational trends and implications for intracellular events. In *Misbehaving Proteins: Protein (Mis)-folding, Aggregation and Stability*; Murphy, R., Tsai, A., Eds.; Springer: New York, 2006; pp 217−246.

(3) Flanagan, J. M.; Kataoka, M.; Shortle, D.; Engelman, D. M. Truncated staphylococcal nuclease is compact but disordered. *Proc. Natl. Acad. Sci. U.S.A.* **1992**, *89* (2), 748−752.

(4) Kurt, N.; Cavagnero, S. The burial of solvent-accessible surface area is a predictor of polypeptide folding and misfolding as a function of chain elongation. *J. Am. Chem. Soc.* **2005**, *127*, 15690−15691.

(5) de Marco, A.; Deuerling, E.; Mogk, A.; Tomoyasu, T.; Bukau, B. Chaperone-based procedure to increase yields of soluble recombinant proteins produced in E. coli. *BMC Biotechnology* **2007**, *7*.

(6) Plaxco, K. W.; Simons, K. T.; Baker, D. Contact order, transition state placement and the refolding rates of single domain proteins. *J. Mol. Biol.* **1998**, *277* (4), 985−994.

(7) Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E. The Protein Data Bank. *Nucleic Acids Res.* **2000**, *28* (1), 235−242.

(8) Orengo, C. A.; Michie, A. D.; Jones, S.; Jones, D. T.; Swindells, M. B.; Thornton, J. M. CATH - a hierarchic classification of protein domain structures. *Structure* **1997**, *5* (8), 1093−1108.

(9) Noguchi, T.; Akiyama, Y. PDB-REPRDB: a database of representative protein chains from the Protein Data Bank (PDB) in 2003. *Nucleic Acids Res.* **2003**, *31* (1), 492−493.

(10) Gromiha, M. M.; Selvaraj, S. Comparison between long-range interactions and contact order in determining the folding rate of two-state proteins: Applications of long-range order to folding rate prediction. *J. Mol. Biol.* **2001**, *310* (1), 27−32.

(11) Kihara, D. The effect of long-range interactions on the secondary structure formation of proteins. *Protein Sci.* **2005**, *14* (8), 1955−1963.

(12) Krishna, M.; Englander, S. The N-terminal to C-terminal motif in protein folding and function. *Proc. Natl. Acad. Sci. U.S.A.* **2005**, *102* (4), 1053−1058.

(13) Flory, P. J. *Statistical Mechanics of Chain Molecules*; Wiley: New York, 1969.

(14) Ivankov, D. N.; Garbuzynskiy, S. O.; Alm, E.; Plaxco, K. W.; Baker, D.; Finkelstein, A. V. Contact order revisited: Influence of protein size on the folding rate. *Protein Sci.* **2003**, *12* (9), 2057−2062.

(15) de Prat Gay, G.; Ruiz-Sanz, J.; Neira, J. L.; Corrales, F. J.; Otzen, D. E.; Ladurner, A. G.; Fersht, A. R. Conformational pathway of the polypeptide chain of chymotrypsin inhibitor-2 growing from its N terminus *in vitro*. Parallels with the protein folding pathway. *J. Mol. Biol.* **1995**, *254*, 968−979.

(16) Itzhaki, L. S.; Neira, J. L.; Ruizsanz, J.; Gay, G. D.; Fersht, A. R. Search for nucleation sites in smaller fragments of chymotrypsin inhibitor-2. *J. Mol. Biol.* **1995**, *254* (2), 289−304.

(17) Neira, J. L.; Fersht, A. R. Exploring the folding funnel of a polypeptide chain by biophysical studies on protein fragments. *J. Mol. Biol.* **1999**, *285*, 1309−1333.

(18) Shortle, D.; Meeker, A. K. Residual structure in large fragments of staphylococcal nuclease: Effects of amino acid substitution. *Biochemistry* **1989**, *28*, 936−944.

(19) Tian, K.; Zhou, B.; Geng, F.; Jing, G. Folding of SNase R begins early during synthesis: the conformational feature of two short N-terminal fragments of staphylococcal nuclease R. *Int. J. Biol. Macromol.* **1998**, *23*, 199−206.

(20) Caflisch, A. Computational models for the prediction of polypeptide aggregation propensity. *Curr. Opin. Chem. Biol.* **2006**, *10* (5), 437−444.

(21) Conchillo-Sole, O.; de Groot, N. S.; Aviles, F. X.; Vendrell, J.; Daura, X.; Ventura, S. AGGRESCAN: a server for the prediction and evaluation of "hot spots" of aggregation in polypeptides. *BMC Bioinformatics* **2007**, *8*.

(22) Dubay, K. F.; Pawar, A. P.; Chiti, F.; Zurdo, J.; Dobson, C. M.; Vendruscolo, M. Prediction of the absolute aggregation rates of amyloidogenic polypeptide chains. *J. Mol. Biol.* **2004**, *341* (5), 1317−1326.

(23) Fernandez-Escamilla, A. M.; Rousseau, F.; Schymkowitz, J.; Serrano, L. Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.* **2004**, *22* (10), 1302−1306.

(24) Trovato, A.; Seno, F.; Tosatto, S. C. E. The PASTA server for protein aggregation prediction. *Protein Eng. Design Selection* **2007**, *20* (10), 521−523.

Received November 30, 2007. Accepted April 1, 2008.

BP070475V