Participation of protein sequence termini in crystal contacts

Oliviero Carugo*

Department of Chemistry, Pavia University, Pavia, Italy

Received 12 April 2011; Revised 27 June 2011; Accepted 30 June 2011

DOI: 10.1002/pro.690

Published online 7 July 2011 proteinscience.org

Abstract: The analysis of the crystal packing interactions, in a nonredundant set of high resolution and monomeric globular protein crystal structures, shows that the residues located at the N- and C-termini of the sequence tend to participate in packing interaction more often than expected and that often they interact with each other. Since the sequence termini are, in general, conformationally very flexible and since they host electrical charges of opposite sign, it can be hypothesized that they play a crucial role in the early formation of the nonphysiological contacts that bring to protein crystallization. It is thus not surprising that modest lengthening/shortening of the sequence termini have often a dramatic effect on protein crystallogenesis.

Keywords: C-terminus; crystallization; crystallogenesis; N-terminus; protein sequence; protein structure

Introduction

It is well known that it is extremely difficult to crystallize proteins¹ and it has been hypothesized that evolution worked against protein crystallization.² However, protein crystallization is of crucial importance as it is a mandatory prerequisite of crystal structure determinations. Protein crystallographers, who need to crystallize proteins, commonly adopt multiconstruct strategies, where many different constructs of the same protein are used by lengthening/shortening the amino acid sequence.^{3,4} Many crystallographers remember some anecdotic case in which well diffracting crystals were obtained by removing/adding just one or two residues at the sequence termini. These “miracles” can have various interpretations.

First, it is possible that minor modifications of the amino acid sequence may improve the folding free energy with the consequence that proteins are better folded and conformationally more homogeneous. However, the folding energy perturbations are likely to be quite modest, since both the N- and the C-terminus tend to be solvent exposed and it is well known that mutations/insertion/deletions at the protein surface do not affect considerably the protein folding thermodynamics, especially if they do not occur in helices and strands.^{5,6} It is also possible that subtle changes at the sequence termini might influence the overall solubility, with consequences in the crystallization process.⁷ And it is also possible, under the hypothesis of a cotranslational folding mechanism at the ribosome level, that the length of the N-terminal segment, which would be sequestered, directly or indirectly, by the ribosome, might have an importance in the expression of well-folded proteins.⁸

Successful crystallogenesis requires large amounts of chemically and conformationally pure protein.

Grant sponsor: GEN-AU initiative (through the BIN-III project)

*Correspondence to: Oliviero Carugo, Dipartimento di Chimica, Viale Taramelli 12, I-27100 Pavia, Italy.

E-mail: oliviero.carugo@univie.ac.at

Table I. Overall Propensity Values, Computed by Summing all the Variables that Define $P(TP)$ Over the 399 Structures Examined in This Article, and Percentage of Structures where the Individual Propensity $P(PT)$ is Larger than 1. Five Values of Nt and Ct (Collectively Named Tt in the table) were Considered

Tt	$P(TP)$	% Cases with $P(TP) > 1$
1	1.962	70.4
2	1.802	65.7
3	1.705	64.9
4	1.638	65.7
5	1.615	64.8

Flexible N- and C-termini, even if relatively short, increase conformational heterogeneity of the sample and can therefore hamper crystallization.

In the present communication, a different explanation of the critical relationship between protein crystallization and amino acid sequence termini is presented. It is based on the observation that N- and C-termini are involved in crystal packing contact more often than expected. They seem to be often crucial in stabilizing the protein solid state phase, and thus it is not surprising to observe that small lengthening/shortening of the sequence termini are so important in determining protein crystallization.

Results and Discussion

The propensity $P(TP)$ of the N- and C-terminal residues to be solvent accessible and to form crystal packing interactions is defined as

$$P(TP) = \frac{nTSP/nSP}{nTS/nS}$$

where nS is the number of solvent accessible residues, nTS is the number of solvent accessible residues that are at the sequence termini, nSP is the number of residues that are solvent accessible and form crystal packing contacts, and $nTSP$ is the number of sequence terminal residues that are solvent accessible and that form crystal packing contacts. Analogously, it is possible to define the propensity $P(CP)$ of C-terminal residues to be solvent accessible and to form crystal packing contacts and the propensity $P(NP)$ of N-terminal residues to be solvent exposed and to form crystal packing contacts.

By definition, propensity values higher than 1.0 indicate that terminal residues tend to be found at the protein surface and to form crystal packing contacts more frequently than expected.

The propensities of terminal residues to be involved in crystal packing interactions are shown in Table I. They were computed by summing the $nTSP$, nSP , nTS , and nS values over all the protein crystal structures and by considering the first Nt and the last Ct residues to be at the N- and at the C-terminus respectively ($Nt = 1,2,3,4$, and 5 ; $Ct =$

$1,2,3,4$, and 5). It appears that the values of propensity are considerably larger than 1.0. This clearly shows that the residues close to the sequence termini participate in crystal packing interactions more often than expected. This conclusion is reinforced by the observations that the propensities are larger than 1.0 in the majority of the individual crystal structures (Table I).

It is also interesting to observe that the crystal packing contacts between N- and C-terminal residues are rather frequent. For example, for $Nt = Ct = 1$, only 2.9% (± 0.1) of the solvent accessible residue belong to the sequence termini. However, 6.3% (± 1.0) of the crystal packing contacts link N- and C-terminal residues. Similar results are observed for other values of Nt and Ct . On the contrary, contacts between sequence termini and other regions of the proteins do not show any tendency to involve particular chemical groups rather than others.

Interestingly, the crystal symmetry is not correlated with the presence of packing interactions that involve the protein termini. The distributions of the space groups are very similar in crystals with contacts involving N- and/or C-termini and in crystals lacking these interactions. They are also closely similar to the space group histograms distributed by the protein data bank (PDB) (<http://www.pdb.org/pdb/statistics/histogram.do?mdcat=symmetry&mditem=symmetry&name=Space%20Group>). Analogously, the presence of screw axes is not associated with packing contacts that involve the protein termini.

It must also be observed that no difference was observed for between residues at the N-terminus and the residues at the C-terminus. Both termini have the same propensity to be involved in crystal packing contacts in the structures examined here.

Propensities were also calculated for different levels of relative solvent accessibility (Table II) and

Table II. Numerical Values Necessary to Compute The Propensities $P(TP)$ for Different Ranges of Relative Solvent Accessibility (SAA) and for Two Nt and Ct (Collectively Named Tt in the Table)

Relative SAA	$nTSP$	nSP	nTS	nS	$P(TP)$
$Tt = 1$					
< 20%	2	2592	33	30836	0.721
20–40%	26	5006	54	16832	1.619
40–60%	40	6607	119	15189	0.773
60–80%	79	5186	148	9663	0.995
> 80%	308	3069	447	5062	1.136
All	455	22460	801	77582	1.962
$Tt = 5$					
< 20%	124	2592	826	30836	1.786
20–40%	272	5006	702	16832	1.303
40–60%	398	6607	809	15189	1.131
60–80%	426	5186	710	9663	1.118
> 80%	591	3069	827	5062	1.179
All	1811	22460	3874	77582	1.615

no clear trends appeared. For example, for $Nt = Ct = 1$, the highest propensity of terminal residues to be involved in crystal packing interactions is observed for amino acids that have a relative solvent accessibility in the 20–40% range. On the contrary, for $Nt = Ct = 5$, the propensity tends to decrease if the accessibility to the solvent increases.

The high frequency of packing interactions involving and between N- and C-termini suggests that they frequently clash in solution, especially at the high concentrations that precede nucleation and crystallization. This might depend on the opposite charges of the N-terminal ammonium and of the C-terminal carboxylate, which attract each other. However, attractive Coulombic interactions cannot be limited only to the sequence termini but might involve also the numerous cationic or anionic side-chains that are usually disseminated on the globular protein surface. Therefore, it seems more alike that the importance of the electric charges at the sequence termini is due to the spatial delocalization. The C- and N-termini tend to protrude from the globular core and to fluctuate in the solvent surrounding the protein. Because of this conformational freedom, the electric charges at the sequence termini can explore several positions with a consequent high probability of interacting with other groups. This feature is typical of the sequence termini.

Modest lengthening/shortening of the N- and C-terminal moieties can thus have severe effects on the formation of nascent, intermolecular interactions that lead to the crystallization. Obviously, the relationship between the length of the termini and the crystallization tendency depends not only on the termini but also on the overall shape, size, and plasticity of the globular protein. It is therefore impossible to design a detailed strategy for exploiting this relationship in practical cases with the consequence that that systematic experimental approaches are used to increase the success rate of crystallization.^{9,10} It must however be kept in mind that an important factor in determining the production of high quality protein crystal is likely to be the length of the sequence termini, which seem to work like the fingers that recognize the aircrafts in the airports and allow the boarding of the passengers.

Material and Methods

The attention was focused on monomeric proteins (experimental annotations from the UniProt database¹¹). This was necessary in order to make the dataset as homogeneous as possible: in a dimer, for example, the N-terminus of a protomer interacts with three other sequence termini, which are the C-terminus of the same protein chain and the two termini of the other molecule. Such a situation is completely different from the case of the monomeric proteins, where there are only two sequence termini of

opposite charge. Membrane proteins and proteins containing coiled-coils were also disregarded. Only X-ray crystal structures were retained from the PDB^{12,13} and those with crystallographic resolution lower than 2.0 Å were disregarded. Sequence redundancy was reduced to 90% of identity to reject multiple copies of the same structure, mutants, isoforms, and complexes with different small molecules. A more severe redundancy cutoff was not necessary, since crystallogenes and sequence homology are likely to be unrelated. Structures with electron density gaps were also ignored. They were detected by looking at the following lines of the PDB files: The REMARK 465 lines, which list the residues that lack completely the positional coordinates; the REMARK 470 lines, which list the nonhydrogen atoms of the amino acids that do not have positional coordinates; and the REMARK 475 lines, which enumerate the residues modeled with zero occupancy. Since most of these missing atoms are at the protein surface, structures that contain them must be eliminated when the crystal packing contacts are examined. Also structures with residues conformationally disordered were eliminated, since different solvent accessibilities can be measured for different conformations not only for the disordered residue but also for the amino acids that surround it. Eventually, structures containing nonstandard amino acids were disregard, since the solvent accessible surface of these residues cannot be estimated by the program NACCESS. 399 protein crystal structures were eventually retained.

Crystal packing contacts were determined with the program *cpc.c*, which is based on a cutoff interatomic distance of 4.5 Å and was successfully used in previous studies.^{14,15} Solvent accessibilities were determined with the program NACCESS (with default parameter values; www.bioinf.manchester.ac.uk/naccess/) that provides relative solvent accessible surface areas for each residues (these are the ratios between the solvent accessible area of the residue X in the protein and the solvent accessibility of the residue X in the tripeptide AXA in an extended conformation). A residue was considered to be accessible to the solvent if its relative accessibility was larger than zero. Residues were considered to be at the N-terminus if they were within the first Nt sequence positions ($Nt = 1, 2, 3, 4$, and 5). Analogously, the C-terminal residues were identified within the last Ct positions ($Ct = 1, 2, 3, 4$, and 5). A variety of Nt and Ct values was used to reinforce the results.

Acknowledgment

Helpful discussions with Kristina Djinovic-Carugo (Department of Structural and Computational Biology, MFPL-Vienna) are gratefully acknowledged.

References

1. McPherson A (1999) Crystallization of biological macromolecules. Cold Spring Harbor: Cold Spring Harbor Laboratory Press.
2. Doye JPK, Louis AA, Vendruscolo M (2004) Inhibition of protein crystallization by evolutionary negative design. *Phys Biol* 1:9–13.
3. Dong A, Xu X, Edwards AM; Midwest Center for Structural Genomics; Structural Genomics Consortium, Chang C, Chruszcz M, Cuff M, Cymborowski M, Di Leo R, Egorova O, Evdokimova E, Filippova E, Gu J, Guthrie J, Ignatchenko A, Joachimiak A, Klostermann N, Kim Y, Korniyenko Y, Minor W, Que Q, Savchenko A, Skarina T, Tan K, Yakunin A, Yee A, Yim V, Zhang R, Zheng H, Akutsu M, Arrowsmith C, Avvakumov GV, Bochkarev A, Dahlgren LG, Dhe-Paganon S, Dimov S, Dombrowski L, Finerty P Jr, Flodin S, Flores A, Gräslund S, Hammerström M, Herman MD, Hong BS, Hui R, Johansson I, Liu Y, Nilsson M, Nedyalkova L, Nordlund P, Nyman T, Min J, Ouyang H, Park HW, Qi C, Rabeh W, Shen L, Shen Y, Sukumard D, Tempel W, Tong Y, Tresagues L, Vedadi M, Walker JR, Weigelt J, Welin M, Wu H, Xiao T, Zeng H, Zhu H (2007) In situ proteolysis for protein crystallization and structure determination. *Nat Methods* 4:1019–1021.
4. Gräslund S, Sagemark J, Berglund H, Dahlgren LG, Flores A, Hammarström M, Johansson I, Kotenyova T, Nilsson M, Nordlund P, Weigelt J (2008) The use of systematic N- and C-terminal deletions to promote production and structural studies of recombinant proteins. *Protein Expr Purif* 58:210–221.
5. Huang LT, Gromiha MM (2010) First insight into the prediction of protein folding rate change upon point mutation. *Bioinformatics* 26:2121–2127.
6. Gromiha MM (2007) Prediction of protein stability upon point mutation. *Biochem Soc Trans* 35:1569–1573.
7. Smialowski P, Martin-Galiano AJ, Cox J, Frishman D (2007) Predicting experimental properties of proteins from sequence by machine learning techniques. *Curr Protein Pept Sci* 8:121–133.
8. Hardesty B, Tsalkova T, Kramer G (1999) Co-translational folding. *Curr Opin Struct Biol* 9:111–114.
9. Ericsson UB, Hallberg BM, Detitta GT, Dekker N, Nordlund P (2006) Thermofluor-based high-throughput stability optimization of proteins for structural studies. *Anal Biochem* 357:289–298.
10. Knaust RK, Nordlund P (2001) Screening for soluble expression of recombinant proteins in a 96-well format. *Anal Biochem* 297:79–85.
11. AAVV (2007) The universal protein resource (UniProt). *Nucl Acids Res* 35:D193–D197.
12. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE (2000) The protein data bank. *Nucl Acids Res* 28:235–242.
13. Bernstein FC, Koetzle TF, Williams GJ, Meyer EF Jr, Brice MD, Rodgers JR, Kennard O, Shimanouchi T, Tasumi M (1977) The Protein data bank: a computer-based archival file for macromolecular structures. *J Mol Biol* 112:535–542.
14. Sikic K, Tomic S, Carugo O (2010) Systematic comparison of crystal and NMR protein structures deposited in the protein data bank. *Op Biochem J* 4:83–95.
15. Weiss MS (2007) On the interrelationship between atomic displacement parameters (ADPs) and coordinates in protein structures. *Acta Crystallogr D* 53:1235–1242.