Electric charge balance mechanism of extended soluble proteins

NOBUYUKI UCHIKOGA,¹ SHUN-YA TAKAHASHI,² RUNCONG KE,¹
MASASHI SONOYAMA,¹ AND SHIGEKI MITAKU¹

¹Nagoya University, Department of Applied Physics, Graduate School of Engineering, Nagoya, Aichi 464-8603, Japan

²Tokyo University of Agriculture and Technology, Department of Biotechnology, Tokyo 184-8588, Japan

(RECEIVED July 19, 2004; FINAL REVISION August 27, 2004; ACCEPTED August 27, 2004)

Abstract

Extended proteins such as calmodulin and troponin C have two globular terminal domains linked by a central region that is exposed to water and often acts as a function-regulating element. The mechanisms that stabilize the tertiary structure of extended proteins appear to differ greatly from those of globular proteins. Identifying such differences in physical properties of amino acid sequences between extended proteins and globular proteins can provide clues useful for identification of extended proteins from complete genomes including orphan sequences. In the present study, we examined the structure and amino acid sequence of extended proteins. We found that extended proteins have a large net electric charge, high charge density, and an even balance of charge between the terminal domains, indicating that electrostatic interaction is a dominant factor in stabilization of extended proteins. Additionally, the central domain exposed to water contained many amphiphilic residues. Extended proteins can be identified from these physical properties of the tertiary structure, which can be deduced from the amino acid sequence. Analysis of physical properties of amino acid sequences can provide clues to the mechanism of protein folding. Also, structural changes in extended proteins may be caused by formation of molecular complexes. Long-range effects of electrostatic interactions also appear to play important roles in structural changes of extended proteins.

Keywords: structural classification; extended protein; bioinformatics; structural genomics; mechanism of structural stabilization; physical properties of amino acid residues

Complete genomes include many orphan amino acid sequences, the functions and structures of which are unknown. Determination of the tertiary structure of the proteins corresponding to these sequences is important for elucidation of their function, because the structure of a protein is closely related to its function. However, some types of proteins with unknown structures, such as nonglobular proteins containing flexible extended segments, are difficult to crystallize. Nonglobular soluble proteins with flexible segments are often involved in regulatory and cell-signaling functions

(Wright and Dyson 1999; Ward et al. 2004). For example, calmodulin and troponin C appear to be nonglobular soluble extended proteins.

Such extended proteins provide very interesting problems involving structure and changes in structure. First, extended proteins lack some physical properties of globular proteins, and vice versa. The structure of single extended protein molecules, as exemplified by calmodulin and troponin C, consists of separate domains near each terminal linked by a central segment exposed to water (Babu et al. 1988; Houdusse et al. 1997; Chou et al. 2001). In contrast, globular proteins are stabilized by a hydrophobic core (Kauzmann 1959).

Second, extended proteins often contain a flexible segment, which allows changes in their structure to occur. For example, the central part of the region linking the terminal

Reprint requests to: Nobuyuki Uchikoga, Department of Applied Physics, Graduate School of Engineering, Nagoya University, Furocho, Chikusaku, Nagoya, Aichi 464-8603, Japan; e-mail: uchikoga@bp.nuap.nagoya-u.ac.jp; fax: +81-52-789-3708.

Article published online ahead of print. Article and publication date are at <http://www.proteinscience.org/cgi/doi/10.1110/ps.04984505>.

domains of calmodulin is flexible (Zhang et al. 1995; Chou et al. 2001). Because of this flexible segment, an extended protein such as calmodulin often has a globular form when it binds a short peptide molecule (Ikura et al. 1992; Meador et al. 1992). In contrast, when a zinc finger protein binds to double-stranded DNA molecules (Pavletich and Pabo 1993), it maintains its extended structure. In both cases, the extended proteins involved have the ability to undergo various structural changes. The mechanisms of stabilization and structural change of extended proteins involve interesting physical problems, because physical conditions of the extended structure must be changed to those of a globular structure. The two aspects of extended proteins discussed above can provide clues useful for classification of extended proteins. Also, they raise the following question: What is the most effective method for classification of extended proteins with a changeable structure?

The most widely used approaches to structural classification relating the structures and functions of proteins are based on comparative methods. Prediction of disordered or unstructured regions is performed using genome-scale comparative analysis methods; e.g., the prediction tool DISOPRED (Jones and Ward 2003; Ward et al. 2004). However, completely orphan sequences cannot be analyzed by comparative methods. *Ab initio* prediction is potentially applicable for genome-scale analysis of extended proteins, but that method appears to require further development for analysis of large amino acid sequences (Hardin et al. 2002). Because structure can be partly inferred from function, classification of extended structures at low resolution can be effective in complete genome analysis. If physical properties that distinguish extended proteins from globular proteins are indicated by examination of amino acid sequences, this can provide clues to the physical principles involved in protein folding and structural change. Membrane proteins have the following physical properties: at least one very hydrophobic segment that spans the hydrophobic region of the membrane, and clusters of amphiphilic residues at the membrane-water interface. On the basis of those physical characteristics, a highly accurate computational method for classifying membrane proteins has been developed (Hirokawa et al. 1998). This indicates that investigation of physical properties of amino acid sequences is useful for classification of proteins.

In the present study, we examined structures and amino acid sequences in the Protein Data Bank (PDB; Berman et al. 2000) to elucidate stabilization mechanisms of extended proteins. We found that the most important factor in discrimination of extended proteins from globular proteins was long-range electrostatic repulsion between separated charges within an extended molecule. The second most important factor was the enhanced propensities of amphiphilic amino acids in the central region exposed to water. In this paper, we also discuss the structural changes that occur

when extended proteins bind other molecules. Furthermore, we developed a software tool (SOSUdumbbell) for predicting extended proteins, based on the two factors described above. This software tool can be applied to genome-scale data sets.

Results

Long-range effects and charge density

Physical properties of amino acid sequences of typical extended proteins were investigated. Typical extended proteins were selected from a data set of single proteins in PDB release 96. The scheme for selection of extended proteins based on 3D structure (Fig. 1A) is presented in detail in the Material and Methods section. We first examined net charges of proteins. In PDB release 96, there were 26,075 amino acid sequences with bell-shaped net charge distribution (Fig. 2A). We designated charges of amino acid residues as follows: Arg, Lys, and His are positive; Asp and Glu are negative; and other amino acids are zero. Figure 2A shows that most proteins have a very small net charge. However, typical extended proteins, such as calmodulin and troponin C, have a significant negative charge. The net charges of calmodulin and troponin C are -23 and -28 , respectively. Only 1.05% ($n=274$) of the above 26,075 proteins had a total net charge more negative than -20 , and

Figure 1. (A) Schematic model of dumbbell-type (extended) protein; $W_{NN} = W_{CC} = 1$. (B) Dispersion diagram of W_{CC} vs. W_{NN} for selection of dumbbell-type proteins based on the criterion $W_{NN} = W_{CC} = 1$.

Figure 2. (A) Distribution of net charge, from single structural data in PDB. (B) Distribution of charge density. (C) Charge balance of N- and C-terminal halves.

0.61% ($n=159$) had a total net charge more positive than +20. Thus, calmodulin and troponin C are very unique with respect to their net charges. This suggests that net charge is the most important factor for discrimination of extended proteins.

If a protein is large and has more than two domains, the interaction between domains is weaker than for extended proteins with only two interacting domains. Therefore, we also used charge density (dividing net charge by sequence length) of total amino acid sequence as a parameter for prediction of extended proteins. Calmodulin and troponin C, which are typical extended proteins, both have a charge density (D_Q) >0.14 (Fig. 2B).

Charge balance

The net charges of extended proteins were >20 or <-20 . If electrostatic repulsion is a dominant factor in stabilization

of the extended structure, the amino acid sequences of these proteins should all show the same pattern of distribution of electric charges.

When the amino acid sequence of calmodulin is divided into N- and C-terminal globular domains and a linking helix (as per PDB annotation), these three parts have negative charges of -12 , -9 , and -2 , respectively. Troponin C shows a very similar distribution: -14 , -12 , and -2 , respectively. This special charge distribution suggests that a repulsion mechanism is responsible for the structural stability of these extended proteins; i.e., repulsive interaction between two globular domains can prevent collapse of the extended structure. Even if the central helix in these proteins is flexible (Zhang et al. 1995; Chou et al. 2001), moderate repulsive interaction between the end points can maintain stability. To elucidate the repulsion mechanism of extended proteins, net charges of the N- and C-terminal halves of various proteins were examined. The distribution of net charge was generally random around the origin of the dispersion diagram (Fig. 2C).

We then estimated the strength of the charge balance mechanism involved in structural stability of extended proteins. For two domains with a net charge of 10 each, at a distance of 2.5 nm, and with the dielectric constant of water set at 80, we obtained a repulsive interaction energy value of 1.2×10^{-19} joules. This value is about 30 times the thermal energy required to affect the conformation of a protein (4×10^{-21} joules). Therefore, proteins with sufficiently great repulsive energy between the two terminal domains to affect conformation are located in the region of the graph where $Q_N \times Q_C > 100$ (Fig. 2C). Solid circles in Figure 2C indicate proteins with a charge density >0.14 . Many proteins that would be plotted in the area of the graph where $Q_N \times Q_C > 100$ have a charge density (D_Q) <0.14 . Only 58 of the original 26,075 amino acid sequences satisfied the following three inequality conditions for electric charge distribution: $|Q_{\text{Total}}| > 20$, $Q_N \times Q_C > 100$ and $D_Q > 0.14$.

Central domain exposed to water

Some of the 58 proteins that satisfied the three inequality conditions of charge balance had a compact structure. This suggests that interaction between amino acid residues and water is also an important factor in stabilization of the extended structure. We examined the amino acid sequences of extended proteins to identify domains exposed to water.

We calculated relative propensity of amino acids in two parts of a domain exposed to water: center and termini. The center and termini of a central helix linking N- and C-terminal domains in typical extended proteins have complementary amino acid compositions (Fig. 3A). The center mainly contains amphiphilic amino acids (Lys, Arg, His, Glu, and Gln, in order of value) with a side chain containing a polar group and a hydrophobic stem (Hirokawa et al.

Figure 3. (A) Relative amino acid propensity of termini and central region of central helices of 11 extended proteins normalized by propensity in globular domains. If a value of relative propensity was greater than unity in a region, correspondent residues appeared frequently in that region. Helix termini were defined as the eight residues shown in Fig. 3B (two shaded regions). The remaining region of the helix was designated the central region. (B) Example of hydrophobicity and amphiphilicity plot of calmodulin (IOSA). The average value of every seven residues in each sequence was plotted as a function of sequence number. Solid line indicates hydrophobicity, using the K-D index (Kyte and Doolittle 1982). Dotted line indicates amphiphilicity, using the index developed by Mitaku et al. (2002). Amphiphilicity index: K, 3.67; R, 2.45; H, 1.45; E, 1.27; Q, 1.25.

1998; Mitaku et al. 2002). In contrast, hydrophobic residues were abundant at the termini of central helices (Fig. 3A).

The characteristics of a central helix of calmodulin are clearly shown in Figure 3B, in which the hydrophobicity scale of Kyte and Doolittle (1982) and a recently reported amphiphilicity index (Mitaku et al. 2002) are used. A sharp peak in amphiphilicity and a deep valley for hydrophobicity were observed at the center of the helix, and peaks of hydrophobicity at the helix termini sandwiched the amphiphilicity peak. The physical meaning of the amphiphilicity peak is not completely clear, but it may stabilize the helix via the mechanism involved in helix formation by detergents such as SDS. These properties can be used to identify a domain predicted to be exposed to water in an amino acid sequence, and the thresholds are detailed in the Materials and Methods subsection “Physicochemical properties of central region exposed to water.”

Discussion

Summary of conditions for prediction of extended proteins

From the analysis of the relationship between amino acid sequence and 3D structure of extended proteins, we determined that the following amino acid sequence conditions must be met for a protein to be predicted as extended: (1)

net charge, $|Q_{\text{Total}}| > 20$; (2) charge density, $D_Q > 0.14$; (3) charge balance, $Q_N \times Q_C > 100$; (4) amphiphilicity and hydrophobicity properties indicating a central helix (see “Physicochemical properties of central helix” section below). We were able to perfectly discriminate extended proteins from compact proteins using these thresholds. We identified 36 extended proteins: 22 calmodulins, 13 troponin Cs, and a transcription factor.

All calmodulins and troponin Cs in our data set were predicted by our criteria. Also, we identified a transcription factor whose amino acid sequence had low homology with calmodulin and troponin C. Our genome analysis identified more than 150 proteins of the human genome that were predicted to be extended, and about half of these were DNA- or RNA-binding proteins annotated as transcription factors, histones, or ribosomal proteins. The four conditions in the preceding paragraph can be used to identify single proteins with extended structures and with no sequence similarity to calmodulin or troponin C. Further results of genome analysis using this algorithm will be reported and discussed in future papers. The present results indicate that extended proteins are stabilized by long-range interaction between the terminals and short-range interaction between the central domain and water molecules.

Most of the predicted extended proteins contained the EF hand motif. There is evidence suggesting that the EF hand motif is responsible for the charge balance of extended proteins. However, recoverin (1REC) and yeast frequenin (1FPW), which have globular forms, each have two EF hand domains (one near the N-terminal, and one near the C-terminal). The charge balance of 1REC is 0 and -3 , and the charge balance of 1FPW is 3 and -9 . This indicates that charge balance involves features of amino acid sequences other than the EF hand motif.

We have developed a method for predicting extended proteins from amino acid sequences. This algorithm, SOSUI dumbbell, is available on the Web as http://bp.nuap.nagoya-u.ac.jp/sosui/sosuidumbbell/dumbbell_submit.html, and is described in Figure 4. Because this tool requires only the amino acid sequence of a protein to determine whether it has an extended structure, it can be applied to sets of all amino acid sequences of entire genomes. We plan to use SOSUI dumbbell to analyze complete genomes in future studies.

Accuracy of prediction of a domain exposed to water

We found that 22 of the proteins we examined were not predicted to have a domain exposed to water; i.e., they did not satisfy our three charge inequality conditions. These proteins did not have as many amphiphilic amino acid residues as extended proteins, whose average amphiphilicity peak was 0.40 (< 0.90). Three of the 22 proteins that were

SOSUIdumbbell: Submit a protein sequence

Enter a title or comment for the sequence :

Enter your sequence with one-letter symbol (by copy & paste) :
 (Minimum: 100 a.a., Maximum: 500 a.a.)

Query sequence

Execution button

To execute the query, press this button :

To clear the form, press this button :

[Sample Sequences]

1. [Calmodulin \(human\)](#)
2. [Myoglobin \(human\)](#)

Dumbbell-type Protein Database by SOSUIdumbbell

sosu@proteome.bi.tuat.ac.jp

SOSUIdumbbell Result

Title: Calmodulin(human)
 This is a dumbbell-type protein.
 N = 148 a.a.
 C = -23

N-terminal domain	Central helix	C-terminal domain
1-64	65-91	92-148

Option

[Go back to SOSUIdumbbell home page](#)

Figure 4. SOSUIdumbbell Web site: http://bp.nuap.nagoya-u.ac.jp/sosu/sosuidumbbell/dumbbell_submit.html. This Web site can be used to determine whether a query sequence is an extended protein. If the query sequence is predicted to be an extended protein, the results page shows the central helical component exposed to water.

predicted as proteins without a central region, one calmodulin (ICTR), and two troponin Cs (1DTL, 1TNX) were expected to be predicted as extended proteins. The structural data for these three proteins were incomplete, and lacked structural data for the domain linking the two terminal domains, probably due to the large fluctuation of these regions. This indicates that prediction of domains exposed to water requires discrimination of extended structures using full-length amino acid sequences by bioinformatics. Nine protein structures that did not correspond to calmodulin or troponin C, and which lacked a central region, were determined to be globular, and included actin-binding proteins, hydrolases, transport proteins, and other types of proteins. Some of these structures had two close globular domains linked by a short loop that was shorter than 19 residues. Ten other structures were determined to be DNA- or RNA-binding proteins, including transcription factors, nucleosomal proteins, and ribosomal proteins. Some of these proteins resembled extended proteins. However, predicted exposed central regions were located near a terminal domain, indicating that the terminal domain was smaller than 50 residues. Extension of these DNA- or RNA-binding proteins may be the result of interaction with DNA or RNA molecules.

Structural change of extended proteins

We found two types of large changes in conformation of extended proteins involved in molecular complexes, by analysis of structural data. Each complex contains a molecule with a charge that is opposite in sign to that of the extended protein. For example, when calmodulin is bound to a polypeptide molecule, its extended structure (Fig. 5, middle left) changes to a collapsed form (Fig. 5, bottom left). Fluctuation of the structure of calmodulin bound to a peptide molecule has been observed in analysis of 3D NMR structural data (Ikura et al. 1992; Meador et al. 1992). In contrast, transcription factors with a large positive net charge bind to double-stranded DNA in an extended form when they recognize a specific nucleotide sequence (Pavletich and Pabo 1993), and the structures of most single extended proteins are unknown (Fig. 5, middle and bottom right). Calmodulin, which has a large negative net charge, can combine with four calcium ions (two at each terminal), implying that long-range repulsion between the two terminal domains is weakened. Conversely, when zinc finger motifs bind zinc ions, the protein becomes more positively charged, which increases its ability to bind to DNA molecules. In each of these cases, long-range electrostatic in-

Figure 5. Two possible structures of extended proteins in bound state. The unbound extended protein molecule forms a stable extended structure with electrostatic repulsion between the terminal domains (*middle left*). When the extended protein binds to a globular molecule, interaction between the terminals of the extended protein weakens, causing it to collapse (*bottom left*). When the extended protein binds to an extended molecule, it remains in an extended structure (*bottom right*).

teraction appears to be the dominant factor in the structural change of the extended protein. Investigation of physical properties of amino acid sequences can provide useful information, not only about protein folding but also about structural changes.

Materials and methods

3D structure data

Two data sets were generated from PDB release 96, excluding redundant amino acid sequences in identical PDB IDs. The lengths of the proteins surveyed ranged from 100 to 500 residues. We analyzed the 3D structural data of 7234 proteins, including full-length data of 15 calmodulins and 11 troponin Cs, to determine which had structures typical of extended proteins. We also analyzed the amino acid sequences of 26,075 proteins, in a data set including 22 calmodulins (15 identified from single protein data, and seven identified from protein complex data) and 13 troponin Cs (11 identified from single protein data, and two identified from protein complex data) to determine which were predicted to be extended proteins.

Definition of extended protein structure

First, we defined the structure of extended proteins and selected typical extended structures from a data set of the PDB. Structural data of extended proteins was compared with physical properties of their amino acid sequences. Then, we used 3D structures of

single proteins (7234 proteins) in PDB release 96 to make a complementary data set of extended-type proteins and other types of proteins. In this analysis, proteins in complexes were excluded because their conformation may be stabilized by protein-protein interaction, which apparently involves interaction between specific domains. Size of proteins was limited to a range of 100–500 residues. A typical extended protein is composed of three segments: N- and C-terminus segments, and a central helix. The central helix is longer than 19 residues, and is near the midpoint of the amino acid sequence. The domains at the N- and C-terminal regions contain more than 50 residues each. Selection of extended-type proteins was also based on the spatial relationship between the central helix and the N- and C-terminus domains (Fig. 1A). We imagined a virtual plane perpendicular to the central helix at its midpoint. With a protein that is globular overall, the virtual plane will pass through domains of both the N- and C-terminal regions. Briefly, proteins were classified as extended type if the virtual plane did not pass through the N- or C-terminal segment; i.e., $W_{NN} = W_{CC} = 1.0$ and $W_{NC} = W_{CN} = 0$ (Fig. 1B). This scheme selects proteins with typical extended structures. Eleven of the 7234 single proteins satisfied these structural criteria for extended proteins: calmodulin (1CLL, 1CLM, 1OSA, 4CLN) and troponin C (1NCX, 1NCY, 1NCZ, 1TN4, 1TOP, 2TN4, 4TNC).

Physicochemical properties of central region exposed to water

To examine the central domain, we used the hydropathy and amphiphilicity indices of amino acids determined by Kyte and Doolittle (1982) and Mitaku et al. (2002), respectively. Minimum amphiphilicity was set at 0.9, and maximum hydrophobicity was set at -1.15 . The side lobes of hydrophobicity peaks were > -0.25 , indicating sequences at least 10 amino acids long. Using these conditions, we were able to locate the edges of the central region exposed to water. The central region was determined as the area in which the proportion of the lateral segments in the terminals was < 1.50 for hydrophobicity and > 0.67 for amphiphilicity.

Acknowledgments

We thank Dr. Nobuhiro Hayashi for providing helpful suggestions. This research was supported in part by a grant-in-aid for special projects in genome science from the Ministry of Education, Sports, Science and Technology (Mombukagakusho) of Japan.

References

- Babu, Y.S., Bugg, C.E., and Cook, W.J. 1988. Structure of calmodulin refined at 2.2 Å resolution. *J. Mol. Biol.* **204**: 191–204.
- Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Res.* **28**: 235–242.
- Chou, J.J., Li, S., Klee, C.B., and Bax, A. 2001. Solution structure of Ca(2+)-calmodulin reveals flexible hand-like properties of its domains. *Nat. Struct. Biol.* **8**: 990–997.
- Hardin, C., Pogorelov, T.V., and Luthey-Schulten, Z. 2002. Ab initio protein structure prediction. *Curr. Opin. Struct. Biol.* **12**: 176–181.
- Hirokawa, T., Boon-Chieng, S., and Mitaku, S. 1998. SOSUI: Classification and secondary structure prediction system for membrane proteins. *Bioinformatics* **14**: 378–379.
- Houdusse, A., Love, M.L., Dominguez, R., Grabarek, Z., and Cohen, C. 1997. Structures of four Ca²⁺-bound troponin C at 2.0 Å resolution: Further insights into the Ca²⁺-switch in the calmodulin superfamily. *Structure* **5**: 1695–1711.
- Ikura, M., Clore, G.M., Gronenborn, A.M., Zhu, G., Klee, C.B., and Bax, A.

1992. Solution structure of a calmodulin-target peptide complex by multi-dimensional NMR. *Science* **256**: 632–638.
- Jones, D.T. and Ward, J.J. 2003. Prediction of disordered regions in proteins from position specific score matrices. *Proteins* **53**: 573–578.
- Kauzmann, W. 1959. Some factors in the interpretation of protein denaturation. *Adv. Protein Chem.* **14**: 1–63.
- Kyte, J. and Doolittle, R.F. 1982. A simple method for displaying the hydrophobic character of a protein. *J. Mol. Biol.* **157**: 105–132.
- Meador, W.R., Means, A.R., and Quincho, F.A. 1992. Target enzyme recognition by calmodulin: 2.4 Å structure of a calmodulin-peptide complex. *Science* **257**: 1251–1255.
- Mitaku, S., Hirokawa, T., and Tsuji, T. 2002. Amphiphilicity index of polar amino acids as an aid in the characterization of amino acid preference at membrane-water interfaces. *Bioinformatics* **18**: 608–616.
- Pavletich, N.P. and Pabo, C.O. 1993. Crystal structure of a five-finger gli-DNA complex: New perspectives on Zinc fingers. *Science* **261**: 1701–1707.
- Ward, J.J., Sodhi, J.S., McGuffin, L.J., Buxton, B.F., and Jones, D.T. 2004. Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J. Mol. Biol.* **337**: 635–645.
- Wright, P.E., and Dyson, H.J. 1999. Intrinsically unstructured proteins: Re-assessing the protein structure-function paradigm. *J. Mol. Biol.* **293**: 321–331.
- Zhang, M., Tanaka, T., and Ikura, M. 1995. Calcium-induced conformational transition revealed by the solution structure of apo calmodulin. *Nat. Struct. Biol.* **2**: 758–767.