Available online at www.sciencedirect.com
SCIENCE DIRECT®
Chemical Physics Letters 405 (2005) 224–228
CHEMICAL PHYSICS LETTERS
www.elsevier.com/locate/cplett

# The optimal size of a globular protein domain: A simple sphere-packing model

Min-yi Shen *, Fred P. Davis, Andrej Sali

*Departments of Biopharmaceutical Sciences and Pharmaceutical Chemistry, California Institute for Quantitative Biomedical Research, University of California, San Francisco, San Francisco, CA 94143, USA*

Received 8 December 2004; in final form 4 February 2005

**Abstract**

We describe a model that relates the optimal size of a globular protein domain to the ratio between hydrophilic and hydrophobic amino acid residues. This model represents a domain as a homogeneous spherical assembly of monodisperse spheres corresponding to the individual residues; the hydrophilic spheres are distributed on the assembly surface, and the hydrophobic spheres are buried in the core. The model predicts that a domain with a 1:1 ratio of hydrophilic and hydrophobic residues is composed of 156 residues. It also predicts that smaller protein domains have more hydrophilic than hydrophobic residues. These predictions are in agreement with the distribution of domain size and residue composition for the experimentally determined protein structures.


Protein domains are considered the basal units of protein structure, function, and evolution [1,2]. These units fold independently, often mediate a specific biological function, and combine modularly to form larger proteins. Several approaches to the definition of domains have been developed based on protein sequence and structure [3]. The two commonly used structure-based domain definition and classification systems are the structural classification of proteins (SCOP) [4] and CATH [5,6]. Experimentally determined structures of proteins are deposited in the protein data bank (PDB) [7]. The PDB currently holds approximately 28,000 structures. Each entry contains on average 2.2 protein chains, and each chain contains on average 2.1 domains.

The characteristic domain size of approximately 160 amino acid residues was first observed in ultracentrifugation experiments in the early 1920s [8]. More recent studies of the distribution of domain sizes in known protein sequences and structures yield similar results, showing that most protein domains range from 100 to 200 residues [9]. For example, the average size of a globular domain from CATH [5,6] is 153 residues [10]. The distributions of domain lengths, and hence the average domain size, do not vary significantly with the specific definitions of domains in SCOP [4], PrISM [11], and CHOPnet [12], despite heterogeneity in the domain definition methods and structure samples to which these methods are applied.

The robustness of the characteristic size of a domain suggests a simple underlying physical principle that is only determined by a few parameters. The earliest theory suggested that the optimal domain size emerges as a consequence of the surface/volume ratio of a sphere [13,14]. This model predicted the characteristic size of a stable domain to be 130 residues. The model was extended by taking into account the composition and size of amino acid residues [15,16]. A more sophisticated physical theory based on both geometry and the free energy of folding predicts the optimal domain size to be 200 residues [17]. Here, we present a simple model that offers additional insight into the distribution and

* Corresponding author. Present address: UCSF Mission Bay Genentech Hall, 600 16th Street, San Francisco, CA 94143-2240, USA. Fax: +1 415 514 4231.
*E-mail address:* smy@salilab.org (M. Shen).

0009-2614/$ - see front matter © 2005 Elsevier B.V. All rights reserved.
doi:10.1016/j.cplett.2005.02.029

average of the domain sizes observed in natural proteins.

We begin by modeling a globular protein domain as an assembly of randomly packed beads that represent individual residues. On average, a randomly packed spherical assembly consisting of $N$ residues of the same size occupies a volume of

$$V_a = \frac{4N\pi a^3}{3\alpha}, \tag{1}$$

where $a$ is the radius of the beads (typically 3.5 Å) representing individual residues and $\alpha$ is the packing ratio of monodisperse three-dimensional (3D) spheres in the random close packing (RCP) state [18–20]. As shown robustly by many simulations and experiments, the packing ratio $\alpha$ is approximately 0.64, which is significantly smaller than the packing ratio of $\pi/\sqrt{18} \approx 0.74$ for the face-centered packing [19,20]. The concept of the RCP state has been superceded by the maximally random jammed (MRJ) state [19], which is physically better defined than the RCP state and has an essentially indistinguishable packing ratio of 0.637 [18,21].

Amino acid residues of different aqueous affinities, namely the hydrophobic and hydrophilic residues, have different spatial distributions in the native protein structures, reflecting the so-called 'minimum condition' for a globular protein domain [9,13]. The hydrophobic residues tend to be buried in the core of a protein, whereas the hydrophilic residues tend to distribute on the surface. This simplified scheme of residue type classification, known as the HP model [22,23], has been applied to the study of many aspects of protein folding kinetics.

Given the HP model, an ideal assembly of $N$ beads has $cN$ hydrophilic beads on the assembly surface and the remaining $(1-c)N$ hydrophobic beads buried in the assembly core, where $c$ is the fraction of hydrophilic residues in the sequence of the domain. The $c$ and $N$ are not independent variables. To satisfy the 'minimum condition' for a globular domain, the $c$ and $N$ must be related to each other. We refer to the corresponding $N$ as optimal ($N_{\text{opt}}$) given a composition ratio $c$ and to the corresponding $c$ as optimal ($c_{\text{opt}}$) given a domain size $N$. When the beads cluster into a spherical assembly, the surface area of the assembly is

$$A_a = 4\pi a^2 \left(\frac{N}{\alpha}\right)^{2/3}. \tag{2}$$

To proceed, we make a 'simplex' approximation that all the hydrophilic surface beads have half of their surface exposed to the solvent. Under this assumption, the total effective surface area of the hydrophilic residues is

$$A_a = 2cN\pi a^2. \tag{3}$$

The optimal value of $N$ for a given residue type composition $c$ can be immediately obtained from equating Eqs. (2) and (3):

$$N_{\text{opt}} = \frac{8}{c^3\alpha^2}. \tag{4}$$

Next, we determine the composition $c$ for the natural protein sequences. The residue types were divided into the hydrophobic and hydrophilic classes, as defined in the Rasmol program [24]. For the entire SwissProt protein sequence database (release 44, July, 2004) [25], the fraction of hydrophilic residues ($c$) is 0.507. When a domain has this composition, our model indicates that the optimal size is approximately 150 residues (Eq. (4), for comparison, $N_{\text{opt}}$ is 156 when $c = 0.5$). This number is in excellent agreement with domain sizes observed in natural proteins (above). Conversely, we can also use the average domain size in the CATH database ($N = 153$ residues) to calculate $c_{\text{opt}}$ of approximately 0.504 (Eq. (4)). Therefore, from both perspectives, our simple model is in a good agreement with the actual data.

Next, we refine the model to estimate the upper and lower bounds on the optimal domain size. These bounds result from more realistic estimates of the minimal and maximal number of surface residues. The surface residues are defined to be the beads that are accessible to an external probe sphere, such as the water molecule with an effective radius of 1.4 Å. The lower bound on the domain size is determined by finding the most economical way to configure a single layer of surface residues while keeping the hydrophobic core inaccessible from water molecules. To shield the hydrophobic core residues from water molecules, the surface residues do not necessarily need to be in close contact. Instead, two surface residues can be up to $2(a + 1.4)$ Å apart and still prevent the water molecule from penetrating into the space between these two residues. Therefore, we model the sparsest packing of surface residues that still shields the core by a group of beads with a separation of $2(a + 1.4)$ Å (Fig. 1). According to this model, the estimated lower bound on the surface residue area is

$$A_a = \frac{cN\pi(a + r_{\text{w}})^2}{\beta}, \tag{5}$$

where $r_{\text{w}}$ and $\beta$ are the characteristic radius of the water molecule and the RCP packing ratio for two-dimensional (2D) disks, respectively. The exact value of the RCP packing ratio is not known as accurately in 2D as it is in 3D, because a dense fluid is less jammed in 2D than in 3D [26]. Reasonable estimates range from 0.86 to 0.89 (the densest 2D packing ratio is $\pi/\sqrt{12} \approx 0.91$). From Eqs. (2) and (5), the lower bound on the optimum size of the globular domain is

$$N_{\text{opt}} = \frac{64\beta^3 a^6}{c^3\alpha^2(a + r_{\text{w}})^6}. \tag{6}$$

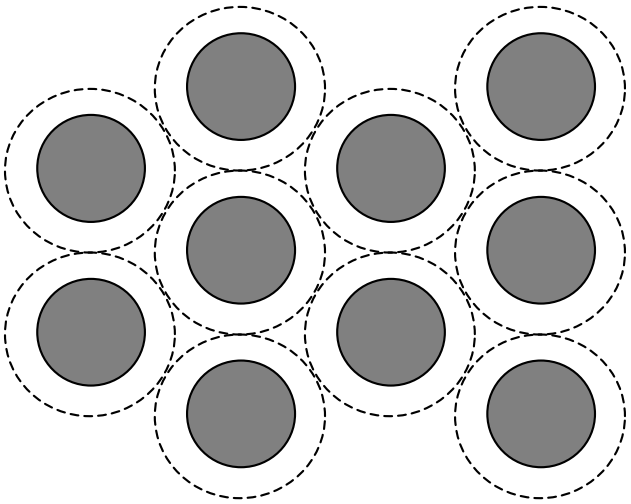

Fig. 1. The sparsest packing of surface beads that still shields the buried beads from water. Gray disks represent the 2D surface projections of the residue beads, with a maximum separation of $r_w$ as denoted by the dashed circles.

For reasonable parameter values ($c = 0.5$, $a = 3.5$ Å, $r_w = 1.4$ Å, $\alpha = 0.64$, and $\beta = 0.88$), the lower bound on the domain size is approximately 117 residues.

As mentioned previously, the lower bound on the optimal domain size was calculated by assuming that the residues in the outermost exposed shell are hydrophilic and the rest of them are hydrophobic. In contrast, we estimate the upper bound on the optimal domain size by assuming that the potentially accessible residues in the first two outermost shells are hydrophilic while the rest of them are hydrophobic. We rationalize this choice as follows. Initially, we have to determine how deeply a residue can reside from the surface of a well-packed spherical assembly while remaining solvent accessible. The first shell residues are defined to have an average depth of zero (Fig. 2), but the second shell residues have a variety of burial depths. A residue that is buried in a second shell right below a first shell residue would have no chance of being exposed to the solvent because it is totally eclipsed by the first shell residues. A model for a maximally buried surface residue describes a pair of surface residues separated by a water sphere, which is in close contact with a second shell residue buried to a depth of $\kappa_{max}a$ (Fig. 2). With simple algebra, we obtain $\kappa_{max} = 2a/(a + r_w) \approx 1.43$, which implies the depth of the maximally buried solvent accessible residues is approximately $1.43a$, which corresponds to 5.0 Å for the 3.5 Å residue radius $a$. This maximum burial depth of 5.0 Å agrees well with the conventional definition of a 'surface residue'.

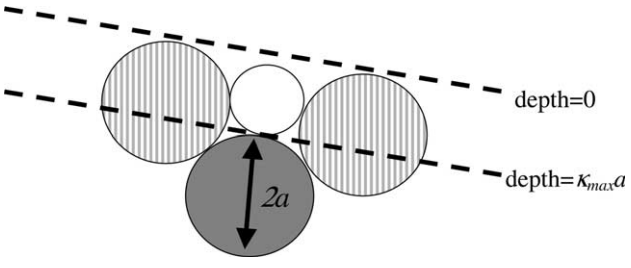


Fig. 2. The densest surface packing. The white, striped and gray spheres denote a water molecule, first shell residues, and an accessible second shell residue, respectively. The maximum burial depth is $\kappa_{max}a$.

With the average residue density and radius of the spherical assembly of $3\alpha/(4\pi a^3)$ and $(N/\alpha)^{1/3}a$, respectively, the expected number of hydrophilic residues in the first and second shell is

$$N_{\text{hydrophilic}} = \alpha\kappa^3 - 3\alpha^{2/3}\kappa^2 N^{1/3} + 3\alpha^{1/3}\kappa N^{2/3}. \quad (7)$$

If we again assume that the hydrophilic residues correspond to one half of the residues ($N_{\text{hydrophilic}}(N_{\text{opt}}) = cN_{\text{opt}}$, $c = 0.5$) in the assembly, the upper bound on the optimal domain size is 213 residues. The explicit expression for the upper bound on the optimal size is complicated and is not presented here.

In summary, our highly simplified sphere-packing model suggests that the optimal size of a globular domain ranges from 117 to 213 residues, with an average of 165 residues. These values are close to the results obtained based on the simplex approximation (Eq. (4)).

The optimal domain size depends on the residue type composition (Fig. 3). The left hand side of the plot corresponds to the case of over-representing hydrophilic residues in a protein domain. In such a case, the domain size should be smaller to retain greater surface/core ratio. In other words, for domains smaller than the optimal size, the simplex approximation predicts the following dependence of the optimal hydrophobic content $c'_{\text{opt}} = 1 - c_{\text{opt}}$ on the domain size $N$:

$$c'_{\text{opt}} = 1 - \sqrt[3]{\frac{8}{\alpha^2 N}}. \quad (8)$$

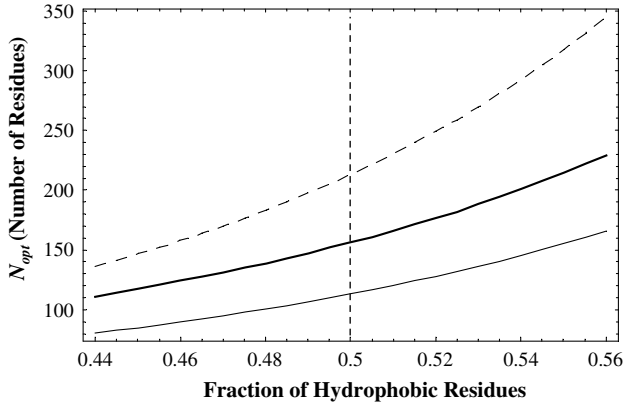


Fig. 3. The dependence of the optimal domain size on the residue type composition. Top (dashed), middle (bold), and bottom curves indicate the upper bound, simplex approximation, and lower bound estimates, respectively. The 1:1 hydrophilic/hydrophobic ratio is indicated by a vertical dashed line.

Next, we compare the dependence of the domain size on the content of hydrophobic residues as observed in the ASTRAL 1.65 database [27,28] with that calculated from Eq. (8) (Fig. 4). The two curves agree qualitatively, sharing a nearly identical minimal domain size of ~20 residues and the increase of domain size as a function of hydrophobic content. On the other hand, larger domains are not observed to have greater hydrophobic content as the theory predicts, suggesting they may retain the nearly 1:1 hydrophilic–hydrophobic ratio by an increase of the surface area.

Above, we assumed a perfectly spherical assembly of residues with an entirely polar coat. However, many domains are not spherical. Moreover, domains in multi-domain proteins and protein assemblies interact with each other via substantially hydrophobic interfaces [29]. Therefore, to make our model more realistic, we now generalize it to an ellipsoid domain with an aspect ratio $\varepsilon(\varepsilon = 1$ for a sphere) and only a fraction $f(0 \leqslant f \leqslant 1)$ of its coat covered by polar residues. Following the arguments for the simplex approximation, the optimal domain size is

$$N_{\text{opt}} = \begin{cases} \frac{f^3}{c^3\alpha^2\varepsilon^2}\left(1+\frac{\varepsilon^2\sin^{-1}\left(\sqrt{1-1/\varepsilon^2}\right)}{\sqrt{\varepsilon^2-1}}\right)^3, & \varepsilon > 1, \\ \frac{8f^3}{c^3\alpha^2}, & \varepsilon = 1, \\ \frac{f^3}{c^3\alpha^2\varepsilon^2}\left(1+\frac{\varepsilon^2\sinh^{-1}(\sqrt{1/\varepsilon^2-1})}{\sqrt{1-\varepsilon^2}}\right)^3, & 0 < \varepsilon < 1. \end{cases} \quad (9)$$

Eq. (9) relates four variables: the hydrophilic residue content, surface polarity, protein size, and eccentricity of a protein. For a constant domain size, the hydrophilic content $c$ reaches the minimum value at $\varepsilon = 1$. This conclusion resembles the prediction by Fisher [16], which states that a spherical structure has the minimum polarity ratio for any given protein size due to the isoperimetric theorem. Conversely, the assembly can be spherical if and only if $N\alpha^2(c/f)^3 = 8$; any non-negative deviation from this relation will lead to aspheric proteins (cf. $N\alpha^2(c/f)^3 \geqslant 8$). Indeed, real small domains (25–50 residues) usually have greater hydrophobic residue content than predicted by the simplex theory for a spherical structure (Fig. 4). Because small domains do not have a truly buried core [16], the excess hydrophobic residues must be distributed on the protein surface, which leads to $c/f \approx 1$, such that $N\alpha^2(c/f)^3 \approx 0.41N$ is always greater than 8 for $25 \leqslant N \leqslant 50$. From this derivation, we deduce that the actual small domains have aspheric shape mainly due to the excess surface hydrophobicity.

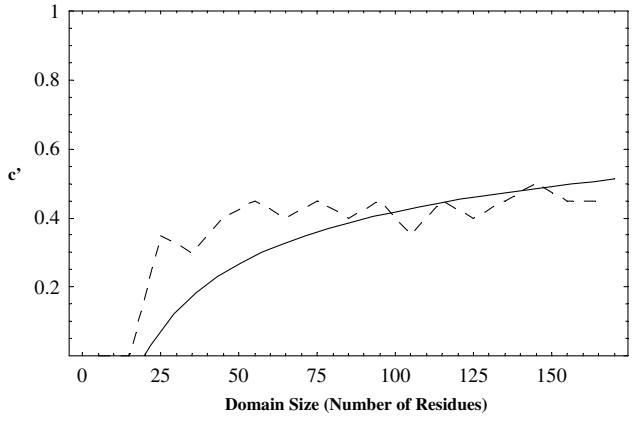


Fig. 4. The dependence of the hydrophobicity on the optimal domain size for domains smaller than 170 residues. The simplex approximation (Eq. (8); solid line). For single domains in SCOP classes a, b, c, and d in ASTRAL 1.65 (dashed line).

In conclusion, we found that the optimal domain size depends strongly on the 3D RCP packing ratio and the hydrophilic/hydrophobic ratio. In contrast, the optimal domain size depends relatively weakly on the sizes of the water molecule and amino acid residues (cf. the simplex model is independent of the amino acid residue and water radii). The 3D random packing ratio should be considered as a rather universal constant, as it is characteristic of many packing problems. Hence, the only 'tunable' parameter in this model is the hydrophilic-to-hydrophobic residue ratio. Does the optimal surface/core ratio arising from geometry as defined in our model steer the evolution of the actual hydrophilic/hydrophobic ratio observed in real proteins, which subsequently determines the domain size? One practical application of our model may be to provide some guidance to the algorithms that aim to define the domains in protein sequences [4–6,11,12].

## Acknowledgments

We thank Profs. Ken Dill and Karl Freed for commenting on this manuscript. Helpful discussions with Dr. Huafeng Xu, Dr. Frank Alber, Dr. Damien Devos, David Eramian and Dr. Bianxiao Cui are also acknowledged. This work was supported in part by the NIH/NIGMS Grant to A.S. (R01 GM54762).

## References

[1] C. Vogel, M. Bashton, N.D. Kerrison, C. Chothia, S.A. Teichmann, Curr. Opin. Struct. Biol. 14 (2004) 208.
[2] C.P. Ponting, R.R. Russell, Annu. Rev. Biophys. Biomol. Struct. 31 (2002) 45.
[3] S. Veretnik, P.E. Bourne, N.N. Alexandrov, I.N. Shindyalov, J. Mol. Biol. 339 (2004) 647.
[4] L. Lo Conte, B. Ailey, T.J. Hubbard, S.E. Brenner, A.G. Murzin, C. Chothia, Nucleic Acids Res. 28 (2000) 257.
[5] C.A. Orengo, A.D. Michie, S. Jones, D.T. Jones, M.B. Swindells, J.M. Thornton, Structure 5 (1997) 1093.
[6] C.A. Orengo, F.M. Pearl, J.M. Thornton, Meth. Biochem. Anal. 44 (2003) 249.

[7] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, Nucleic Acids Res. 28 (2000) 235.
[8] T. Svedberg, Nature 123 (1929) 871.
[9] E.N. Trifonov, I.N. Berezovsky, Curr. Opin. Struct. Biol. 13 (2003) 110.
[10] S.K. Burley, Nat. Struct. Biol. 7 (Suppl.) (2000) 932.
[11] A.S. Yang, B. Honig, J. Mol. Biol. 301 (2000) 679.
[12] J. Liu, B. Rost, Nucleic Acids Res. 32 (2004) 3522.
[13] S.E. Bresler, D.L. Talmud, Compt. Rend. Acad. Sci. URSS 43 (1944) 310.
[14] W. Kauzmann, Adv. Prot. Chem. 14 (1959) 1.
[15] R.E. Gates, H.F. Fisher, Proc. Natl. Acad. Sci. USA 68 (1971) 2928.
[16] H.F. Fisher, Proc. Natl. Acad. Sci. USA 51 (1964) 1285.
[17] K.A. Dill, Biochemistry 24 (1985) 1501.
[18] A. Donev, F.H. Stillinger, P.M. Chaikin, S. Torquato, Phys. Rev. Lett. 92 (2004).
[19] S. Torquato, T.M. Truskett, P.G. Debenedetti, Phys. Rev. Lett. 84 (2000) 2064.
[20] W.M. Visscher, M. Bolsterl, Nature 239 (1972) 504.
[21] A. Donev, I. Cisse, D. Sachs, E. Variano, F.H. Stillinger, R. Connelly, S. Torquato, P.M. Chaikin, Science 303 (2004) 990.
[22] K.Z. Yue, K.A. Dill, Phys. Rev. E 48 (1993) 2267.
[23] K.A. Dill, K.M. Fiebig, H.S. Chan, Proc. Natl. Acad. Sci. USA 90 (1993) 1942.
[24] R.A. Sayle, E.J. Milner, Trends Biochem. Sci. 20 (1995) 374, Amino acids A, G, I, L, M, F, P, W and V are classified as hydrophobic. For more information: http://info.bio.cmu.edu/Courses/BiochemMols/RasFrames/PSTABLE.HTM.
[25] A. Bairoch, B. Boeckmann, S. Ferro, E. Gasteiger, Brief Bioinform. 5 (2004) 39.
[26] A. Donev, S. Torquato, F.H. Stillinger, R. Connelly, J. Appl. Phys. 95 (2004) 989.
[27] J.M. Chandonia, G. Hon, N.S. Walker, L. Lo Conte, P. Koehl, M. Levitt, S.E. Brenner, Nucleic Acids Res. 32 (Database issue) (2004) D189.
[28] J.M. Chandonia, N.S. Walker, L. Lo Conte, P. Koehl, M. Levitt, S.E. Brenner, Nucleic Acids Res. 30 (2002) 260.
[29] J. Janin, S. Miller, C. Chothia, J. Mol. Biol. 204 (1988) 155.