A Topologically Related Singularity Suggests a Maximum Preferred Size for Protein Domains

ARTICLE *in* PROTEINS STRUCTURE FUNCTION AND BIOINFORMATICS · FEBRUARY 2006

Impact Factor: 2.63 · DOI: 10.1002/prot.21179 · Source: PubMed

Charles L Webber
Loyola University Chicago, Stritch School of …

Alessandro Giuliani
Istituto Superiore di Sanità

A Topologically Related Singularity Suggests a Maximum Preferred Size for Protein Domains

Joseph P. Zbilut,[1]* Gek Huey Chua,[2] Arun Krishnan,[3] Cecilia Bossa,[4] Kristian Rother,[5] Charles L. Webber Jr.,[6] and Alessandro Giuliani[4]

[1]*Department of Molecular Biophysics and Physiology, Rush University Medical Center, Chicago, Illinois 60612*
[2]*Department of Biochemistry, National University of Singapore, Singapore 117597*
[3]*Institute for Advanced Biosciences, Keio University, Tsuruoka-shi, Yamagata 997-0035, Japan*
[4]*Department of Environment and Health, Istituto Superiore di Sanitá, Viale Regina Elena 299, 00161 Roma, Italy*
[5]*Humboldt Universität Berlin, Institut für Biochemie an der Charite, Monbijoustrasse 2, 10117 Berlin, Germany*
[6]*Department of Physiology, Loyola University Medical Center, Maywood, Illinois 60153*

ABSTRACT **A variety of protein physicochemical as well as topological properties, demonstrate a scaling behavior relative to chain length. Many of the scalings can be modeled as a power law which is qualitatively similar across the examples. In this article, we suggest a rational explanation to these observations on the basis of both protein connectivity and hydrophobic constraints of residues compactness relative to surface volume. Unexpectedly, in an examination of these relationships, a singularity was shown to exist near 255–270 residues length, and may be associated with an upper limit for domain size. Evaluation of related G-factor data points to a wide range of conformational plasticity near this point. In addition to its theoretical importance, we show by an application of CASP experimental and predicted structures, that the scaling is a practical filter for protein structure prediction. Proteins 2007;66:621–629.** **© 2006 Wiley-Liss, Inc.**

Key words: domain size; protein; hydrophobicity; compactness; packing; folding; G-factor; length-dependency; singularity; recurrence; critical phenomena

INTRODUCTION

A significant literature exists reporting a general scaling of different geometric features of protein 3D structures with chain length. This scaling has been independently noted, and concomitantly there have been some attempts to explore the physical basis for the scaling.[1–3] Among these, Liang and Dill[1] employed percolation theory to suggest that scaling was related to packing in proteins and is similar to randomly-packed spheres. They concentrate on the accumulation of packing defects giving rise to voids (cavities), thereby decreasing the protein density with increasing length. To examine the effect of voids they chose proteins greater than 200 amino acids in length, since voids are relatively rare below this number. Liang and coworkers[3] used models of self-avoiding walks and concluded that compact packing was only a feature of relatively short polymers (<190 residues), and that, beyond this length, proteins are less densely packed and more prone to void formation. Thus the common observation is the presence of a qualitative difference in the folding configuration near 200 residues. Scaling power laws, however, demonstrate no distinctive features that would suggest a divergence in this region, except for the gradual exponential knee.

Our own interest in the topic developed from an examination of the relationship between protein connectivity and chain length: in a previous article we demonstrated the ability of quantification of contact map features relative to the between α-carbon atoms distances in protein 3D structures to predict secondary structures content and position; however, the overall value for contacts, which we term REC3D, exhibited no correlations with any other calculated variable, but did exhibit a scaling law.[4]

In the present article we further investigate this variable and the possible physical cause. We demonstrate how this measure allows us to detect a general scaling of protein structures pointing to two different topological modes: a “small protein” and a “large protein” mode for respectively less than 180 and greater than 320 residues length with a prominent singularity inside this range corresponding to a zone with a rich repertoire of topological arrangements. We propose an explanation of this singularity based on geometric considerations: as the chain length increases a critical value of the ratio of surface volume (SV) to compactness is achieved and a divergence occurs. This divergence implies a structural instability which may be a natural limit for domain length. Also, from an application point of view, we demonstrate

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

Grant sponsor: NSF; Grant sponsor: NIH; Grant number: NSF DMS #0240230.

*Correspondence to: Joseph P. Zbilut, Department of Molecular Biophysics and Physiology, 1229 Jelke, Rush University Medical Center, 1653 W. Congress Parkway, Chicago, IL 60612.
E-mail: jzbilut@rush.edu

Received 10 May 2006; Accepted 19 July 2006

Published online 11 December 2006 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.21179

TABLE I. Proteins Analyzed

PDF	Length	No. of domains
1DN3	15	nc[a]
1SOL	20	nc
1MEA	28	nc
1CBH	36	nc
1PPT	36	nc
1BBG	40	1
1BDS	43	1
1Q9B	43	1
5RXN	54	1
2OVO	56	1
1BPI	58	1
5PTI	58	1
2CRT	60	1
1D1L	61	1
1UTG	70	1
1HOE	74	1
1UBQ	76	1
451C	82	1
1HIP	85	1
2GN5	87	1
1BTA	89	1
1CYV	98	1
1BU4	104	1
2RNT	104	1
1FD2	106	1
2CDV	107	1
1RNB	110	1
2MHR	118	1
1GD6	119	1
1G96	120	1
1A4V	123	1
1BP2	123	1
1HFX	123	1
1KVY	123	1
7RSA	124	1
1JUG	125	1
1DKJ	129	1
2EQL	129	1
133L	130	1
1I56	130	1
1LZ1	130	1
1BBN	133	1
1IRL	133	1
1AQT	136	2
1ANU	138	1
1B0B	142	1
1DM1	146	1
1DSW	153	1
2MM1	153	1
2A5E	156	1
1B0O	162	1
3DFR	162	1
5TNC	162	2
1L24	164	1
1AMM	174	2
1EL4	195	1
2STV	195	1
1BP4	212	1
9PAP	212	1
1CHG	245	2
12CA	260	1
1BR5	267	2
1RHD	293	2
1XGO	295	2
5CPA	307	1
2GBP	309	2
4PEP	326	2
2LBP	346	2
1AL7	359	1
1ANF	370	2
1PHH	394	2
2CPP	414	1
1CPY	421	2
1PKN	530	3
1FCE	629	3
1CYG	680	4
1A47	683	4
1CXE	686	4
1DOT	686	4
1CE2	689	4
1QFM	710	2
1AA6	715	3
1FDI	715	nc
1FDO	715	nc
1ACC	735	4
1BF2	750	3
1FGH	754	2
7ACN	754	4
1YGE	839	5
1AXR	842	2
1DIK	874	6

[a] nc, not classified.

how this measure is suitable for acting as a prefilter for structure prediction methods.

MATERIALS AND COMPUTATIONAL METHODS

Databases

The article in Ref. 4 details our methodology which we outline here. We started with the dataset of 67 monomeric proteins of Hobohm et al.,[5] and removed entries with ambiguous structural coordinates (Table I). This set was augmented to ensure that longer (but still monomeric) proteins were also represented. This resulted in a collection of 91 proteins (length range 15–874 residues). Although the dataset is relatively small, its useful characteristics include: (1) all proteins have well resolved structures; (2) the range of length is fairly representative; and (3) they cover almost the entire range of secondary structures content (0%–83% of α and β content, respectively). These features allowed us to consider the dataset as a well balanced set of globular proteins with a continuous distribution of the considered characteristics with no outliers biasing the computed correlations.

To verify the findings based on this dataset, for the current article, 1979 single chain PDB files solved by X-ray diffraction were obtained from CATH v2.6.0 (April

2005) in Cath List Format (CLF) 1.0 (http://cathwww.biochem.ucl.ac.uk/latest/lists/index.html) (see Supplementary Material).

REC3D Computation

The α-carbon distance matrix was coded putting a dot for every pair of residues whose α-carbon atoms are at ≤6 Å distance in the 3D structure. Distances up to 11 Å were examined and found not to provide additional information. To eliminate trivial contacts due to the nearness in sequence, only residue pairs at a topological distance greater than three positions were considered.

REC3D is derived from recurrence analysis and is a measure of the "number of contacts" normalized for the "possible contacts" (number of residues couples). It can be viewed as a measure of "compactness," and is computed as

$$R_{i,j} = \Theta(\varepsilon_i - \| \vec{\chi}_i - \vec{\chi}_j \|), \vec{\chi}_i \in \Re^{m},$$
$$i,j = 1, \ldots, N, \text{and}, \text{REC3D} = \frac{1}{N^2}\sum_{i,j=1}^{N} R_{i,j} \tag{1}$$

where R is a recurrence (i.e., an actual contact between α-carbons), N is the number of total contacts $\vec{x}_i$, ε_i is the threshold distance of ≤6 Å, $\| \cdot \|$ is the Euclidean norm of the atom coordinates, and $\Theta(\cdot)$ is the Heaviside function.

It is noted that REC3D is effectively equivalent to Chan and Dill's ρ,[6] a measure of "compactness"; i.e., $\rho = t/t_{\max}$, $0 \leq \rho \leq 1$, where $t_{\max}$ is the maximum number of nonbonded contacts possible for an n-polymer. They observed that their ρ is bounded by 0.16–0.24 for intermediate to very long chains. We confirm this finding (see below). Given the equivalency, and to avoid confusion, from this point onward, we will use the common term, ρ, to refer to REC3D. This variable was found to be completely statistically independent from secondary structure descriptors, thus implying a different aspect of protein structural description. (See Ref. 4 for a detailed discussion of the other contact map quantifiers).

SV Computation

Liang and Dill[1] pointed out that one way to study protein packing was to consider surface/volume relationships in terms of scalings. This was used as a basis for comparing proteins to randomly packed spheres near the percolation threshold. Using a 636 subset of the Hobohm and Sander list,[7] they found that the logarithm of surface volume (SV), log V, scales with the logarithm of the length of the protein, log R with a characteristic slope of 2.47 ± 0.04. We used this finding as the basis for our calculation of SV for the ratio relation (see below).

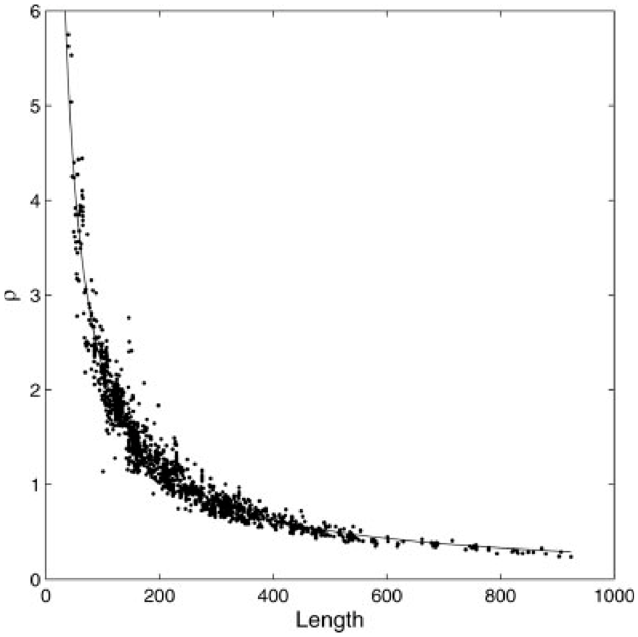

Fig. 1. Scaling of ρ vs. protein chain length in original dataset.

RESULTS

The graph linking the length of the proteins from our original dataset and their ρ resulted in the relation reported in Figure 1 expressible by means of a power law,

$$\rho = 84(\text{Length}^{-0.79})(r = 0.94; P < 0.0001). \tag{2}$$

As can be seen from the figure, the power law describes a bend near 100 residues, but no other prominent features. This scaling is consistent with many other plots depicting the length dependency of other protein variables such as volume, hydrodynamic radius, and packing density. We conjectured that a common topological feature of globular proteins may be responsible for this scaling. In this regard we set out to determine if there was a fixed relationship between contacts and SV relative to length.

For the original dataset, the ratio of SV relative to ρ was computed. Figure 2 demonstrates this relationship. The results point to a divergence beginning near length 180 and ending around 320 residues length. To get a sharper identification of this putative critical region, we computed the "theoretical" ratio using as the denominator the derived ρ scaling of Eq. (2); i.e., SV/84 (Length$^{-0.79}$). As can be seen, the divergence appears at length 274. Noteworthy is also that the ratio slowly increases to the singular point, discontinuously drops below the baseline, and then slowly returns to a baseline (Fig. 3). Thus, this region exhibits classic singularity behavior in terms of a discontinuity.[8,9]

To confirm that this critical point is not dependent upon the database, the same procedures were applied to the CATH data. A scaling was also confirmed as defined by ρ = 152(Length$^{-0.91}$) (r = 0.98; P < 0.0001) (Fig. 4).

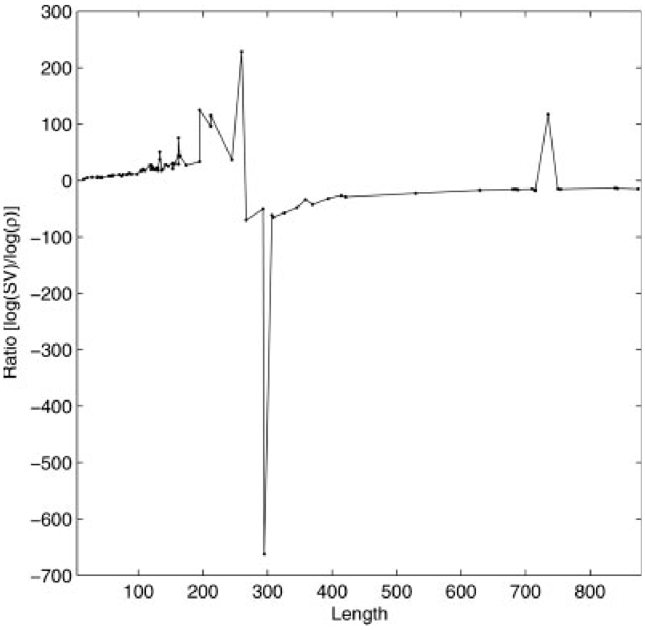

Fig. 2. Ratio of original dataset.

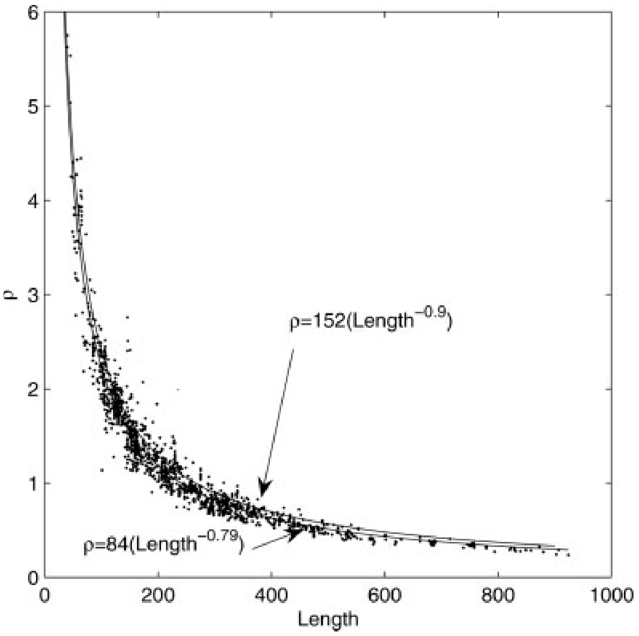

Fig. 4. Scaling based on CATH database.

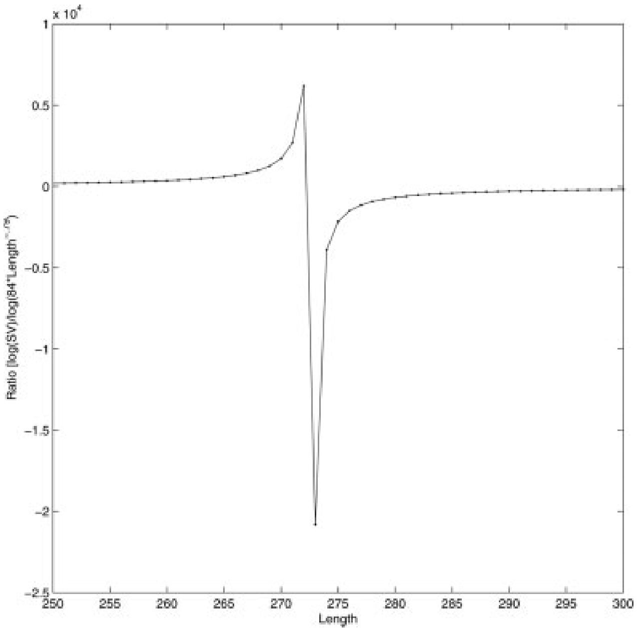

Fig. 3. Theoretical ratio based on SV/84*(Length$^{-0.79}$). Classic, singular divergence is noted at length 274.

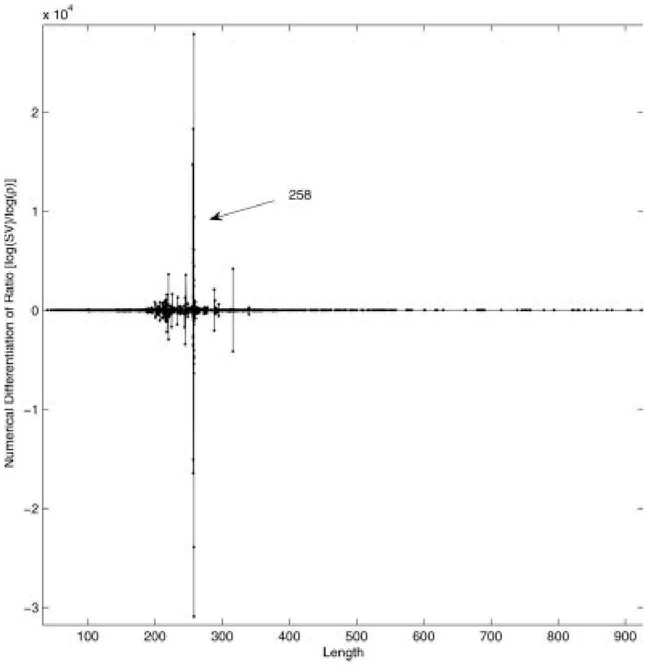

Fig. 5. Numerical differentiation of CATH-based ratio.

Thus the exponent of the scaling was not significantly different from the original. Examination of the plot of the numerically differentiated ratio confirmed the location of the criticality at a length of 258 (Fig. 5).

DISCUSSION

The appearance of a divergence is reminiscent of a critical point often found in phase transitions in condensed matter physics. Liang and Dill[1] referred to this concept in their analysis, but not in reference to the present phenomenon. This divergence is apparent in the ratio between SV and contact compactness thus intimating a topological invariant at the basis of this observation. Although the reduced and larger databases differ slightly in the location of the divergence, the locations are roughly equivalent given the experimental errors incurred in the 3D coordinates as well as statistical errors of curve fitting. The question remains, however, as to whether the finding is simply a result of the partic-

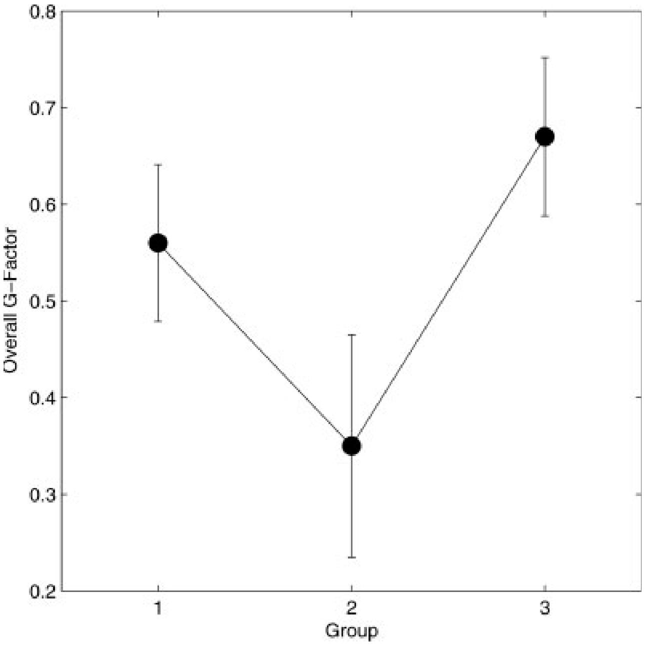

Fig. 6. Variance of overall G-factors for lengths partitioned as: group 1 = 1–179; group 2 = 180–320; and group 3 >320. Group 2 represents lengths near the singularity. Groups are based on the CATH chains.

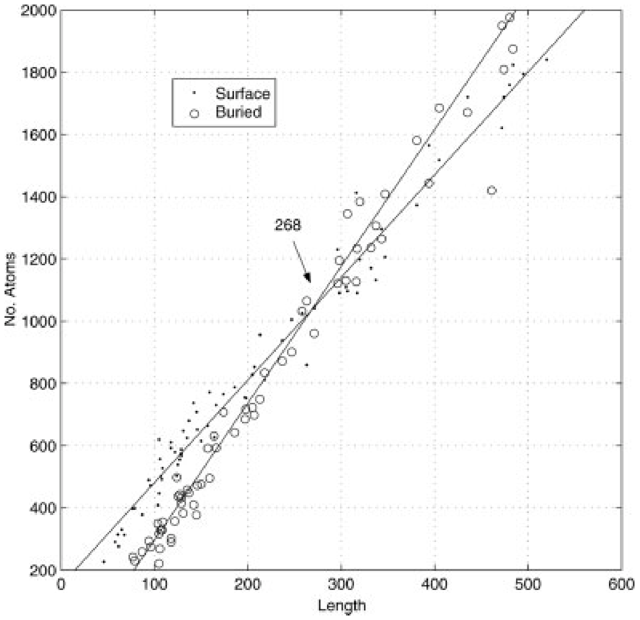

Fig. 7. Plot of numbers of atoms on the surface and atoms buried vs. length. Regression lines are significant at the level of $P < 0.0001$.

ular computational method used. To answer this question, we turned to two different sources: (1) the analysis of G-factors and (2) packing theory, and protein shell density.

G-Factors

G-factors provide a measure of how "normal," or alternatively how "unusual," a given stereochemical property is.[10] G-factors are essentially just a log-odds score based on the observed distributions of these stereochemical parameters. When applied to a given residue, a low G-factor indicates that the property corresponds to a low-probability conformation. So, for example, residues falling in the disallowed regions of the Ramachandran plot will have a low (or very negative) G-factor. Thus, we conjectured that the region near the singularity would exhibit a comparatively large variance for overall G-factors (reflecting the large range of ρ "swings"). Choice of boundaries for the critical region was guided by the differentiated ratio of Figure 5: the differentiation is smooth except for the section delimited by length 180–320.

The overall average G-factor score for the proteins was obtained from PROCHECK (http://www.biochem.ucl.ac.uk/~roman/procheck/procheck.html). A comparison of these factors for the chains in the critical region near the singularity (length 180–320, group 2; group 1 = proteins <180; and group 3 proteins >320) revealed a significant difference in their variability (Bartlett's test of variance: χ^2, $P < 0.0001$) (Fig. 6). The large variability of G-factors in the neighborhood of the singularity points to a particularly high plasticity in terms of topological features of proteins near the divergence. This observation is in line with the modal value for length of proteins (225) in the SWISS-PROT database (http://ca.expasy.org/sprot/relnotes/relstat.html). It is not clear why most proteins are in this region; nonetheless, it may have to do with the fact that the outermost amino acids exhibit an unusually large extent of topological movement, perhaps providing increased possibilities for ligand interactions.[11]

Packing Theory and Shell Density

Protein connectivity is in some ways related to sphere packing; i.e., how densely can objects fill a volume. Donev et al.,[12] in considering the packing problem using candies as experimental objects, have pointed out that ellipsoids can pack very densely (~0.74). However, to some degree, this is dependent upon their aspect ratio, with the packing increasing as the spheroid becomes more amorphous. The authors suggested that as the objects deviate progressively from spheres, there is an increased number of contacts required to eliminate local and collective degrees of freedom. And in making more contacts between the objects, the packing becomes denser.

Packing, Shells, and Density

Rother et al.[13] observed that packing in a protein molecule varies with "depth" of the shell. They noted that the innermost shell was not as densely packed as the next outermost shell. This is consistent with the observation of Sandelin[2] that the fraction of hydrophobic residues remain fairly constant relative to residue length. The implication is that the hydrophobic forces maintain a strong influence on packing with a variation of density going from the core to the outer shells. In other words

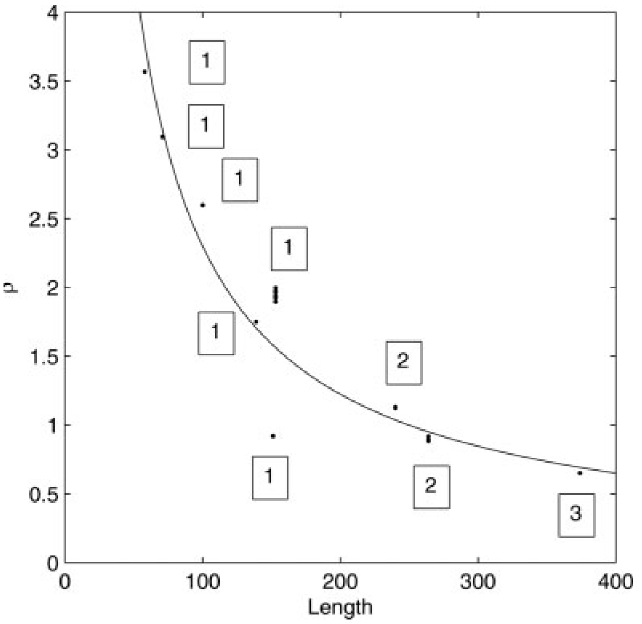

Fig. 8. ρ values for fibrous proteins vs. length compared to line of scaling. Boxed numbers refer to number of domains.

TABLE II. Fibrous Chains

PDB	Chain	Length	No. of domains
1ksr	O	100	1
2btf	A	374	3
2btf	P	139	1
1h6w	A	151	1
1vih	O	71	1
1qiu	A	264	2
1qiu	B	264	2
1qiu	C	264	2
1qiu	D	264	2
1qiu	E	264	2
1qiu	F	264	2
1pu0	A	153	1
1pu0	B	153	1
1pu0	C	153	1
1pu0	D	153	1
1pu0	E	153	1
1pu0	F	153	1
1pu0	G	153	1
1pu0	H	153	1
1pu0	I	153	1
1pu0	J	153	1
1okx	A	240	2
1okx	B	240	2
1kun	O	58	1

the "hydrophobic effect" progressively diminishes going toward the exterior.

Rother et al.[13] also performed a study of packing density based upon surface atoms vs. buried atoms. If a plot is taken of these two quantities vs. residue length, the regression line for buried atoms crosses the line of surface atoms in the neighborhood of 268 residues (Fig. 7). It is important to note that this study was based on a completely different data set with different variables, and suggests a coincident location for the singularity with respect to our 258 estimate. Linking this evidence to the general theory of Donev et al.[12] would suggest that the deepest areas of globular proteins have a relatively fixed packing governed by hydrophobic forces. However, the remaining shells do not have this limitation and may have fewer contacts. This implies that progressively fewer contacts are preserved as indeed the graph of ρ vs. length shows. Nonetheless, at a critical point the numbers of buried residues increases beyond those of the surface abolishing the possibility of obtaining the necessary contacts to maintain close packing.

Shen et al.[14] taking into consideration hydrophobicity and packing, and using a simple theoretical sphere packing model of the type described above, came to the conclusion that the optimal domain size ranged from 117 to 213 residues. We suspect that use of spherical packing underestimates various shells of protein packing as the work of Rother et al.[13] suggests, and the optimal domain size would increase with decreasing sphericity.

Fibrous Proteins

An obvious question that arises is whether fibrous proteins behave according to the scaling law. Somewhat surprisingly, the identified fibrous examples obey the scaling law (Fig. 8) (Table II). The reason for this may redound to the fact that they have similar hydrophobicity patterns irrespective of their length. In a previous work we demonstrated that such proteins typically had repetitive hydrophobicity patterns.[15] If one accepts the findings of Rother et al.[13] and Sandelin[2] that hydrophobicity remains relatively constant relative to length, the scaling should be expected: in these proteins relatively few "heterogeneous" amino acids are included in their chains. Thus their compactness is essentially related to their typical hydrophobic clusters. It is also useful to recall that although ρ, compactness, and p, packing density are sometimes correlated, they are somewhat different. Packing density, p, considers the amount of space within the van der Waals envelope of the molecule divided by the total volume of space that contains the molecule; whereas ρ is a measure of contacts. Thus as Liang et al. point out, the relationship between these two measures is nonlinear, with the number of voids influencing packing density.[16]

Numbers of Domains

A plot of the number of domains vs. length generally confirms a progressive increase in the numbers of domains, although there is overlap (Fig. 9). If grouped according to length by proximity to the singularity one sees significant differences among all three groups (Fig. 10) (ANOVA, for all group contrasts, $P < 0.0001$). However, the difference between Groups 1 and 2 is not as

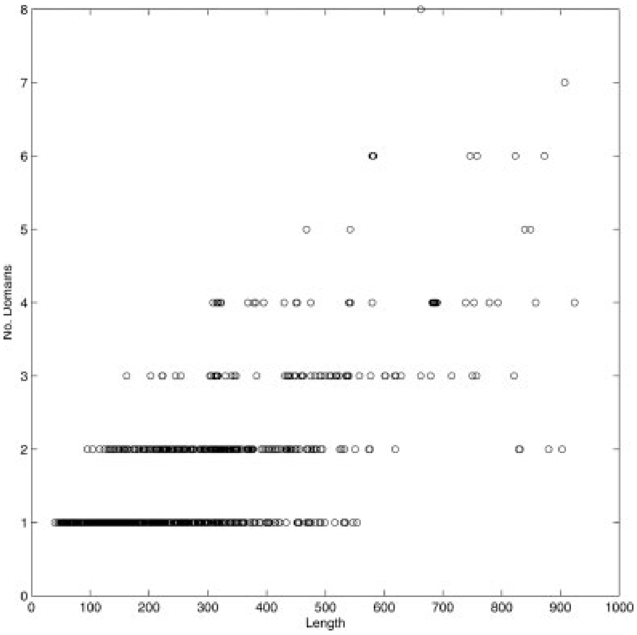

Fig. 9. Numbers of domains for CATH database vs. length.

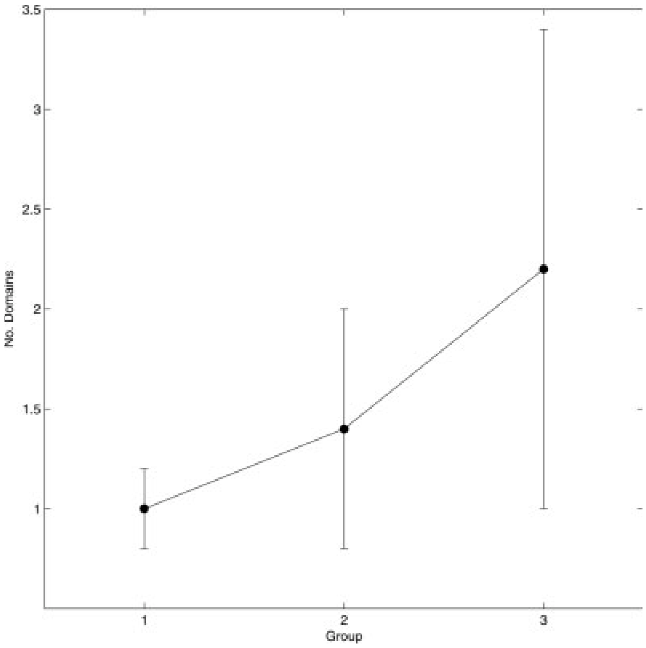

Fig. 10. ANOVA for numbers of domains factored by group scheme as in Figure 6.

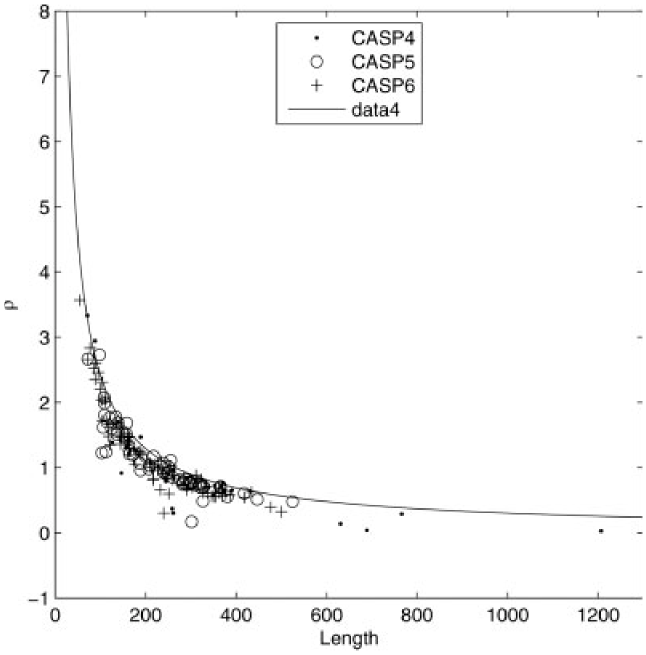

Fig. 11. Actual crystal structure ρ values for CASP4,5,6 vs. length compared to the scaling law.

great as between 2 and 3. The reasons for the overlap are not clear. First, there is the obvious effect of experimental and statistical error. A second putative cause is the fact that many of the chains are complexed with metallic ions, ligands, or both. Clearly, such complexed proteins may affect protein stability and contacts. Finally, some proteins may have unique amino acid compositions such that they favor structures such as bridges. More generally, we still do not have a unique definition of what a domain is.[17] From a purely formal perspective a domain must be interpreted as a "module" or a "cluster" of the network of contacts; i.e., a domain is made by those residues for which the number of intracluster contacts is much greater than the number of intercluster contacts.[18] This formal definition of domain, borrowed from graph theory, is in our opinion the most natural one for interpreting our results but is still not widely accepted in protein science. Thus, we prefer to rely on the classical definition of domains based on super-secondary structure and this could have some effect on the observed results.

Utility

The existence of a scaling and critical domain length has obvious interest from a theoretical consideration of protein structure formation. A more practical use of this scaling may include an initial check of the consistency of proposed sequence/structure models. To this end we applied the above formula to the most recent CASP exercises (CASPs 4, 5 and 6; http://predictioncenter.gc.ucdavis.edu) and immediately recognized that the actual crystal structures of the model proteins follow the proposed scaling law (Fig. 11), thus adding to the validity of the observed relation. More importantly, when applied to the space of predictions made by the different CASP predictors, the scaling evidences a large scattering with respect to the X-ray solved structures.

Figure 12 shows a plot of the predictions for CASP6 against the scaling curve. It can be observed that almost all of the correct models fall in a tight band around the scaling curve. This result gives us an immediate oppor-

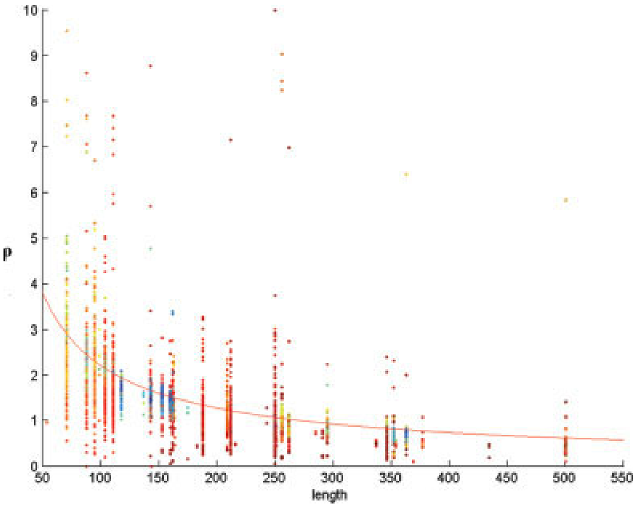

Fig. 12. Plot of ρ for CASP6 predictions vs. length: ρ values for the predictions for CASP6 plotted against the theoretical scaling relationship. Each point represents a predicted structure and is colored by its distance from the actual structure using GDT_TS as the metric. The colors range from the smallest distance (blue) to the largest distance (red). As can be observed, almost all of the good models lie in a narrow band along the scaling curve. See Ref. 19 for a discussion of GDT_TS.

tunity to have in advance, a filter, for judging the correctness of a proposed sequence structure prediction: if the compactness of the model is far from the reference line (obtainable by the simple knowledge of protein length) it implies that the model is erroneous.

Thus it can be seen that the scaling relationship between the percentage of residue contacts and the length of a protein can be used to discriminate to a significant extent between correct and bad models on the basis of the length of the proteins which, of course, is known to the modeler in advance. However, this discrimination covers only an half of the problem: a prediction that is far from the scaling law is likely a bad prediction but the inverse is not true. On the other hand, it was observed that almost all the "good" models tend to follow the scaling relationship. Hence, although we can claim that models that lie far from the scaling curve are "bad" models, no such claim for the goodness of a model can be made for those lying close to the scaling curve. This method however does give researchers a quick filter to weed out truly incorrect predictions before knowing the actual structure, and thus substantially improve the quality of sequence/structure models. [A website has been created where interested readers can submit their structure files online to get the predicted REC3D (ρ) value and plot it against the scaling curve. The URL is http://gosper.iab.keio.ac.jp]

CONCLUSIONS

Taking these observations together it is suggested that as protein size grows, the number of contacts per residue decreases (ρ) until reaching the singular point, where the compactness introduces asymmetries and forces a divergence with the number of contacts per residues remaining more or less constant with length. This behavior suggests that proteins go from a very compact and dense globular form to a type of "lattice" form (in the lattice the number of neighbors of a given element is constant despite the dimensions of the lattice). This may correspond to a shift from single to multidomain organization in which each residue has contacts only with residues of the same domain. This interpretation is supported by the observation that, besides the lack of a consensus about what a "domain" is, after the singularity the majority of analyzed proteins are defined as "multidomain" by CATH definition (Figs. 9 and 10).

This shift from single domain to multidomain architecture after the singularity with the consequent independence of protein contact density from protein length tends to support our definition of a maximal domain size around 180–320 residues. This observation prompts us in accordance to Ref. 16, to a general definition of domain analogous to the one used for defining modules inside networks (or clusters in a multidimensional space) as a group of residues whose internal connectivity (number of contacts between residues pertaining to the same domain) greatly outperforms the average connectivity of the system.

Although we have suggested a topological basis for maximal domain size it should be recognized that to a degree, it appears that the innermost hydrophobic core sets the basis for this feature. Certainly, Chan and Dill alluded to this more than a decade ago.[20] Sandelin indicates that hydrophobic residues remain constant irrespective of protein length, while the hydrophobic core size decreases. Additionally, Miao et al.[21] suggest that there is an optimal hydrophobic cluster fraction to ensure collapse. Thus it would seem that the remaining shells rearrange themselves relative to the hydrophobic core. Yet all of the rearrangements ultimately converge to the fact that the SV/ρ relationship eventually limits domain size to ~258 residues. Consequently, it would seem that the clear determinant is the hydrophobic core cluster, with the remaining residues accommodating a variable number of inter-residue contacts.

ACKNOWLEDGEMENTS

JPZ thanks Prof. Zeev Schuss of the Applied Mathematics Department, Tel-Aviv University, and Jie Liang of the Bioinformatics Program at the University of Illinois Chicago for useful discussions; and Erik Sandelin of the Bioinformatics Center, Stockhoms Universitet for the data of Ref. 2. CGH carried out the calculations. AK and AG initiated the study of scaling. CB was involved with the analysis. KR analyzed protein shell packing. CW developed the algorithms. JPZ developed the analysis and writing.

REFERENCES

1. Liang J, Dill KA. Are proteins well packed? Biophys J 2001;81: 751–766.
2. Sandelin E. On hydrophobicity and conformational specificity in proteins. Biophys J 2004;86:23–30.
3. Zhang J, Chen R, Tang C, Liang J. Origin of scaling behavior of protein packing density: a sequential Monte Carlo study of compact long chain polymers. J Chem Phys 2003;118:6102–6109.
4. Webber CL, Giuliani A, Zbilut JP, Colosimo A. Elucidating protein secondary structures using α carbon recurrence quantifications. Proteins: Struct Funct Genet 2001;44:292–303.
5. Hobohm U, Scharf M, Schneider R, Sander C. Selection of representative protein data sets. Protein Sci 1992;1:409–417.
6. Chan HS, Dill KA. Compact polymers. Macromolecules 1989;22: 4559–4573.
7. Hobohm U, Sander C. Enlarged representative set of protein structures. Protein Sci 1994;3:522–524.
8. Zbilut JP. Unstable singularities and randomness. Amsterdam: Elsevier; 2004.
9. Zak M, Zbilut JP, Meyers RE. From instability to intelligence. Berlin: Springer; 1997.Lecture notes in physics, New series M, Monographs, Vol.49.
10. Engh RA, Huber R. Accurate bond and angle parameters for X-ray protein structure refinement. Acta Crystallogr A 1991;47: 392–400.
11. Zavodszky MI, Kuhn LA. Side chain flexibility in protein ligand binding. Protein Sci 2005;14:1104–1114.
12. Donev A, Cisse I, Sachs D, Variano EA, Stillinger FH, Connelly R, Torquato S, Chaikin PM. Improving the density of jammed disordered packings using ellipsoids. Science 2004;303:990–993.
13. Rother K, Preissner R, Goede A, Frömmel C. Inhomogeneous molecular density: reference packing densities and distribution of cavities within proteins. Bioinformatics 2003;19:2112–2121.
14. Shen M, Davis FP, Sali A. The optimal size of globular protein domain: a simple sphere-packing model. Chem Phys Lett 2005;405: 224–228.
15. Zbilut JP, Colosimo A, Conti F, Colafranceschi M, Manetti C, Valerio MC, Webber CL, Jr, Giuliani A. Protein aggregation/folding: the role of deterministic singularities of sequence hydrophobicity as determined by nonlinear signal analysis of acylphosphatase and Aβ(1–40). Biophys J 2003;85:3544–3557.
16. Liang J, Zhang J, Chen R. Statistical geometry of packing defects of lattice chain polymer from enumeration and sequential Monte Carlo method. J Chem Phys 2002;117:3511–3521.
17. Saini HK, Fischer D. Meta-DP: domain prediction meta-server. Bioinformatics 2005;21:2917–2920.
18. Kannan N, Vishveshwara S. Identification of side-chain clusters in protein structures by a graph spectral method. J Mol Biol 1999;292:441–464.
19. Vendovas C, Zemla A, Fidelis K, Moult J. Assessment of progress over the CASP experiments. Proteins: Struct Funct Genet 2003;53:585–595.
20. Chan HS, Dill KA. Origins of structure in globular protein. Proc Natl Acad Sci USA 1990;87:6388–6392.
21. Miao J, Klein-Seetharaman J, Meierovitch H. The optimal fraction of hydrophobic residues required to ensure protein collapse. J Mol Biol 2004;344:797–811.