PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

RESEARCH COMMENTARY

# A mathematically related singularity and the maximum size of protein domains

**András Szilágyi***

Institute of Enzymology, Biological Research Center, Hungarian Academy of Sciences, Budapest H-1113, Hungary

*ABSTRACT*

***In a paper titled "A topologically related singularity suggests a maximum preferred size for protein domains" (Zbilut et al., Proteins 2007;66:621–629), Zbilut et al. claim to have found a singularity in certain geometrical properties of protein structures, and suggest that this singularity may limit the maximum size of protein domains. They find further support for the singularity in their analysis of G-factors calculated by the PROCHECK program. Here, we show that the claimed singularity is a mathematical artifact with no physical meaning, and we reanalyze the G-factors to show that Zbilut et al.'s results are due to a single outlier in the data. Thus, the existence of an actual singularity in the topological properties of proteins is not supported by the findings of Zbilut et al.***

Proteins 2008; 71:2086–2088.
© 2008 Wiley-Liss, Inc.

*Key words: domain size; protein; compactness; G-factor; singularity.*

In the February 15, 2007, issue of Proteins, Zbilut *et al.* published a paper titled "A topologically related singularity suggests a maximum preferred size for protein domains."[1] In the paper, the authors examine the scaling of certain geometrical properties of protein structures with the length of protein chains, and claim to have found a singularity which could explain why the size of protein domains is limited to about 250–300 residues.

However, a closer examination of what Zbilut *et al.* calculate shows that the singularity they find is nothing but a mathematical artifact.

First, the authors introduce a quantity named REC3D or ρ, which is the number of actual Cα-Cα contacts divided by the maximum possible number of contacts. Mathematically, it is equal to $2C/N^2$, where $C$ is the number of Cα pairs closer than 6 Å (pairs closer than four positions along the sequence are excluded) and $N$ is the chain length. Although Zbilut *et al.* state that REC3D is essentially equivalent to the quantity ρ defined by Chan and Dill,[2] there is a major difference between REC3D and Dill's ρ because when calculating the possible number of contacts, Chan and Dill took the excluded volume effect into account while Zbilut *et al.* take the possible number of contacts to be $N^2/2$, which corresponds to a state where every Cα atom is forced together into a sphere with a 6 Å diameter. In fact, the maximum possible number of contacts scales with $N$ rather than $N^2$, see for example Eqs. (8.3)–(8.5) in Chan and Dill.[3] In globular proteins, the number of contacts per residue is approximately constant for residues in the core and smaller for residues at the surface. Therefore, the scaling of $C/N$ is determined by the surface to volume ratio: it slowly increases with $N$, becoming flat at longer chain lengths [see e.g. Eq. (8) and Fig. 3, top left in Bastolla and Demetrius[4]]. Thus, it is not surprising that the shape of the REC3D ($=2C/N^2$) versus $N$ curve is dominated by the shape of the $1/N$ function (Fig. 1 in Zbilut *et al.*). Thus, the overall shape of this function essentially just reflects the shape of $1/N$ versus $N$, and is not a reflection of any intrinsic topological property of proteins.

By definition, ρ = REC3D must be between 0 and 1. However, in Figure 1 in Zbilut *et al.*, which shows ρ versus $N$, the vertical axis is scaled from 0 to 6. No explanation is given in the paper; only after recalculating the ρ values for a few proteins can the reader find out the reason: the authors multiplied the value of ρ by 100, that is, they

*Correspondence to: András Szilágyi, Institute of Enzymology, Hungarian Academy of Sciences, Karolina ut 29, Budapest H-1113, Hungary. E-mail: szia@enzim.hu
Received 29 October 2007; Revised 4 January 2008; Accepted 21 January 2008
Published online 12 March 2008 in Wiley InterScience (www.interscience.wiley.com). DOI: 10.1002/prot.22000


© 2008 WILEY-LISS, INC.

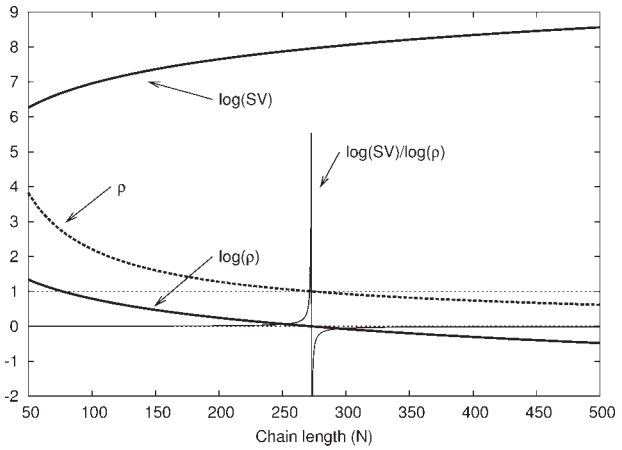


**Figure 1**

*The functions involved in Zbilut et al.'s calculations, plotted versus the chain length N. The following approximations were used to plot the functions: log(SV) = log(10.5N); ρ = $84N^{-0.79}$; log(ρ) = log($84N^{-0.79}$). The function log(SV)/log(ρ) was divided by 2000 to bring it into the plot area. Note that ρ is expressed in percentages. At N = $84^{1/0.79}$ = 273, the value of ρ is 1%, therefore its logarithm is zero. This results in the singular behavior of log(SV)/log(ρ) at N = 273.*

expressed ρ as a percentage. The curve fitting [Eq. (2) in the paper] was also done using 100ρ instead of ρ. Interestingly, this choice, that is using 100ρ instead of ρ, combined with the mathematical transformations the authors apply later, leads to the appearance of the singularity the authors noticed.

The paper starts discussing the ratio of the "surface volume" to ρ. The "surface volume" *SV* is the volume enclosed by the molecular surface of the protein, excluding cavities,[5] and it increases monotonically with *N;* the relationship is linear,[5] a good approximation being $SV \approx 10.5N$ when *SV* is measured in Å$^3$. (Note: the "protein length" used by Zbilut *et al.* to calculate *SV* is not the number of residues but the geometric size of the protein.[5]) Figure 2 in the paper is supposed to show the ratio *SV*/ρ versus the chain length, and it displays a curve with an obvious divergent region, that is, a singularity near a chain length of 274. However, on closer examination of Zbilut *et al.*'s Figure 2, it turns out that it is not *SV*/ρ that is plotted but log(*SV*)/log(ρ), which is an entirely different quantity. Now, the origin of the "singularity" is easy to see; see Figure 1 in this article for a schematic representation of the functions involved. ρ was expressed in percentages, and drops from about 6% to about 0.3% as chain length increases, crossing the value of 1% at a length of 274. Consequently, its logarithm crosses zero at the same length. When log(*SV*) is divided by log(ρ), an obvious "singularity" arises at length 274, because of the division by zero, and, not surprisingly, a function shape reminiscent of $f(x) = -1/x$ appears. However, this singularity simply arises due to the fact that ρ was expressed in percentages before its logarithm was taken, and then this quantity was used as a denominator of a fraction. This procedure creates a singularity at the chain length where the number of contacts, *C*, is 1% of the number of possible contacts (which was taken to be $N^2/2$ in the authors' treatment). Clearly, this singularity is nothing but a mathematical artifact created by the inappropriate mathematical transformations applied to the data. By changing the "unit" of ρ, the singularity can be moved anywhere on the horizontal axis of the plot; it even disappears completely when the original, unscaled ρ is used. Thus, the quantity log(*SV*)/log(ρ) is not physically meaningful because it depends on the scaling of ρ, which can be arbitrarily chosen. The fact that expressing ρ as a percentage results in a singularity at length 274, a reasonably-looking size for a protein domain, is just a coincidence. If one would plot the quantity *SV*/ρ, which the paper actually writes about, no singularity would be seen.

Thus, the singularity the authors found has no physical meaning and is not a reflection of protein topological properties. Next, however, the authors set out to find further signs of the singularity. The main supporting evidence comes from an analysis of G-factors, calculated by the PROCHECK program.[6] It should be noted that the purpose of the PROCHECK program is the evaluation of protein structure quality, and the G-factors primarily reflect the experimental errors rather than the intrinsic geometric or topological properties of proteins. The authors divide their protein set into three groups: group 1 includes proteins shorter than 180 residues, group 2 contains those with lengths 180 to 320, and group 3

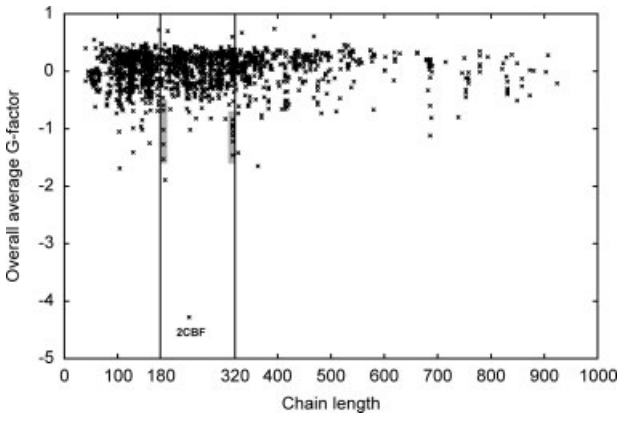


**Figure 2**

*Overall average G-factors, calculated by the PROCHECK program, plotted as a function of chain length, for the 1979 protein structures taken from the supplementary material to Zbilut et al.[1] Vertical lines at lengths 180 and 320 indicate the boundaries of the three groups Zbilut et al. defined. The single outlier in the data set is labeled "2CBF." The gray rectangles enclose the data points corresponding to multiple poor-quality structures of human dihydrofolate reductase (length: 186) and L. casei thymidylate synthase (length: 316).*

includes those longer than 320 residues. They then show that the average overall G-factor of structures in group 2 is lower, and its variance is greater, than that of the other two groups. However, a reanalysis of the actual data (see Fig. 2) shows that most of the difference between the variances is due to a single outlier in group 2, namely, the PDB structure 2CBF, which has a resolution of 3.1 Å, and is of very poor quality (overall G-factor −4.28). The mean G-factors are 0.057, 0.032, and 0.067, with standard deviations 0.28, 0.34, and 0.29 in groups 1, 2, and 3, respectively. When the single outlier is removed from group 2, the mean G-factor increases to 0.039 and the standard deviation drops to 0.3, which no longer differs significantly from that of the other two groups. In addition, because Zbilut *et al.* took no precautions to eliminate redundancies in the data set, there are two proteins in group 2 with multiple poor-quality structures: human dihydrofolate reductase (length: 186; PDB entries 1OHK, 1OHJ, 1HFR, 1HFQ, and 1HFP) and thymidylate synthase from *Lactobacillus casei* (length: 316; PDB entries 1LCB, 1VZD, 1VZC, 1LCE, 1LCA, 1VZA, 1VZB, and 1VZE). These structures come from related experiments and contain similar stereochemical errors. Just keeping one representative in each of these two groups of structures (and removing the rest) increases the mean G-factor in group 2 to 0.057 and lowers the standard deviation to 0.26. Group 2 is then completely indistinguishable from the other two groups on the basis of G-factors. Therefore, the result of Zbilut *et al.*'s G-factor analysis is not robust against outliers (it essentially hinges on a single outlier), and the reanalysis does not support the hypothesis that topological protein properties show any divergence or singularity in the 180–320 length range.

The remaining findings in Zbilut *et al.*'s paper do not support the "singularity" hypothesis either; any change we see is just a smooth, continuous change; no sign of any divergent behavior appears.

Why the size of protein domains tends to be limited to about 300 residues (although this is not a hard limit; several known domains, for example pyruvate formate lyase, are of over 700 residues long) is an intriguing question. The answer may lie in the physics underlying the stability of proteins or the mechanisms of protein folding, but evolutionary reasons, and the role of domains as functional modules may also play a role. Although Zbilut *et al.*'s suggestion about a topologically related singularity may not hold water, further research into the problem is certainly justified.

## ACKNOWLEDGMENTS

András Szilágyi was supported by a Bolyai János fellowship. The author wishes to thank Professor Péter Závodszky for advice and support, and Zsuzsanna Dosztányi, Gergely Gyimesi, Bálint Mészáros, and Dániel Györffy for helpful discussions.

## REFERENCES

1. Zbilut JP, Chua GH, Krishnan A, Bossa C, Rother K, Webber CLJ, Giuliani A. A topologically related singularity suggests a maximum preferred size for protein domains. Proteins 2007;66:621–629.
2. Chan HS, Dill KA. Compact polymers. Macromolecules 1989;22: 4559–4573.
3. Chan HS, Dill KA. The effects of internal constraints on the configurations of chain molecules. J Chem Phys 1990;92:3118–3135.
4. Bastolla U, Demetrius L. Stability constraints and protein evolution: the role of chain length, composition and disulfide bonds. Protein Eng Des Sel 2005;18:405–415.
5. Liang J, Dill KA. Are proteins well-packed? Biophys J 2001;81:751–766.
6. Laskowski RA, McArthur MW, Moss DS, Thornton JM. PROCHECK: a program to check the stereochemical quality of protein structures. J Appl Crystallogr 1993;26:283–291.