Scaling and self-organized criticality in proteins I

J. C. Phillips¹

Department of Physics and Astronomy, Rutgers University, Piscataway, NJ 08854

Contributed by J. C. Phillips, November 8, 2008 (sent for review July 17, 2008)

The complexity of proteins is substantially simplified by regarding them as archetypical examples of self-organized criticality (SOC). To test this idea and elaborate on it, this article applies the Moret-Zebende SOC hydrophobicity scale to the large-scale scaffold repeat protein of the HEAT superfamily, PR65/A. Hydrophobic plasticity is defined and used to identify docking platforms and hinges from repeat sequences alone. The difference between the MZ scale and conventional hydrophobicity scales reflects long-range conformational forces that are central to protein functionality.

hydrophobicity | repeat | scaling

There are many scales (hydrophobicity, charge, helix-forming propensity, etc.) that bioinformatically describe various aspects of amino acid interactions in proteins and their relations to sequence, structure, stability, and functionality, all justified by combining plausible constructs with parameters adjusted through statistical searches. Are all such scales equally useful, and are the substantial differences among them attributable to their inevitably heuristic character? Here, I discuss differences in methodology, with special emphasis on holistic vs. reductionist approaches. I suggest that a better understanding of the limitations and potential of scaling analysis of proteins will result from systematic studies of simple cases first, such as repeat proteins and their associates. Scaling analysis may prove to be a valuable tool in designing proteins with desired functions.

Many physical systems exhibit power-law distributions over limited ranges (hence the popularity of log-log graph paper), and power-law distributions are the characteristic feature of the modern theory of phase transitions near a critical point. Self-organized criticality (SOC) is a methodology that attempts to explain why so many complex systems exhibit power-law distributions and appear to be “accidentally” located near critical points. It is argued that the critical points are dynamical fixed points (“tipping points”) toward which the system evolves without tuning external parameters (1). The critical points are extrema in some property (or properties) with respect to which the system has been optimized, especially with respect to long-range, highly-cooperative interactions, such as conformational changes.

Given the widespread occurrence of power laws, SOC has great intuitive appeal: it has achieved an enduring popularity among theorists (>2,500 papers discussing SOC, >25 current books from a single publisher), notably in modeling the critical stability of sand piles against avalanches, but its concrete applications have been limited largely to seismic phenomena. A cautionary remark: power laws, especially in the context of dissipative reactions, have often been interpreted in terms of SOC, but a simpler interpretation involves only slow sweeping of a control parameter toward a global extremal, as could occur in the context of commercial product optimization (2). However, more recently Boolchand and coworkers (3) discovered a reversibility window associated with an intermediate phase (IP) in the phase diagrams of both chalcogenide and oxide network glasses; near reversibility is strongly suggestive of the concepts of equilibrium statistical mechanics. Such networks resemble proteins, not only geometrically, but also in their exponential (Vogel-Fulcher viscosity) slowing near the glass transition, mimicking the slowness of protein conformational changes, regarded as microscopic phase transitions. The reversibility window is bounded by edges that become more abrupt (4) in the presence

of nonbridging oxygen (analogous to H bonds in proteins). In the reversibility window the internal stress and aging of the networks also nearly vanishes (3), further heightening the similarities to unstressed (nearly ideally compacted) proteins.

Are these abstract SOC-IP models relevant to folded proteins, or should they simply be regarded as yet another inexplicably successful fractal approach to complexity? It has been argued that SOC is a misnomer in the context of life-origin models (5), but bioinformatic studies of graphs of protein domain structures have shown scale-free (and stress-free) properties consistent with SOC (6). These issues have been brought into much sharper focus by the Moret-Zebende (MZ) discovery that solvent-accessible surface areas (SASA) in protein fragments containing $(2N + 1)$ amino acids contract according to power laws in $(2N + 1)$ ($1 \leq N \leq 17$) with different (centered, amino acid-specific) exponents (7). These exponents, derived from bioinformatic scans of 5,526 high-resolution protein fragments, show that each amino acid induces long-range changes in local backbone curvatures [which are smaller (larger) for hydrophilic(phobic) residues]. In general the fluctuations associated with extrema associated with SOC are expected to involve long-wave lengths, such as large-scale conformational changes in microscopic protein-phase transitions.

The hierarchical ordering of these dimensionless holistic exponents (not energies) is qualitatively similar (7) to that of reductionist hydrophobicity scales based on ordering transference energies of isolated amino acids (an MZ fragment with $N = 0$) from water to organic solvents (8, 9), which involve shortest wave lengths. Reductionist isolation methods have often been criticized on logical grounds, but in practice replacing them with reliable and transferable results holistically is difficult. In other words, the entirely unexpected discovery of the MZ self-similar (power-law) hydrophobicity scale based on SASA is of great interest not only in terms of applications, but also in terms of scaling methods in general and connections between SOC and evolution. A deeper discussion of SASA and geometrical aspects of the highly original MZ scale is given elsewhere (10).

Previously, it has been noted that weak, large-scale interactions (like water-protein interactions) are generally more important than strong (intraprotein) ones in determining properties and functions of large, complex systems; the classic example is the Hueckel model of (aromatic) polycyclic planar hydrocarbons (benzene, anthracene, naphthalene, etc.), where weak out-of-plane π polarizability interactions dominate strong in-plane σ interactions, even in feeding back to determine variations of in-plane σ bond lengths (10). However, in general one cannot find classical force field (CFF) screening parameters to describe the effects of weak water interactions on strong bare amino acid interactions. Even in a pair lattice model with only 2 kinds of interactions and 2 kinds of amino acids, the 2 adjustable parameters are fixed by the diagonal interactions, leaving no flexibility to treat the off-diagonal interactions (11, 12), a situation that rapidly deteriorates with 20 aa interacting in virial clusters of 3 or more amino acids. Power laws and SOC (7, 10) close this apparently insoluble (open-ended or divergent)

Author contributions: J.C.P. designed research, performed research, and wrote the paper.

The author declares no conflict of interest.

¹E-mail: jcpillips8@comcast.net.

© 2009 by The National Academy of Sciences of the USA

series expansion problem by treating conformational changes as microscopic-phase transitions.

Given that SOC is a plausible model for relating protein sequences to some combination of structure, stability, and functionality (10), and that MZ have discovered a long-range hydrophobicity scale compatible with SOC, what should we look for? We should look first for proteins with large SASA, where interactions with water are critical to functionality, such as nonglobular superhelical repeat proteins (10, 13). We should not look first at transition states associated with unfolding, leading to dysfunctional aggregation and amyloid formation; the latter depends primarily on at least 3 factors, hydrophobicity, charge, and helix-forming propensity (14). Nor should we search bioinformatically for a “universal” model that ignores functionality entirely, because each protein family with similar functions in all likelihood shares a specific SOC that is different from other families with different functionalities. As for methods, the reader definitely should not look for bioinformatic statistics; these are already contained in the MZ hydrophobicity scale, based on 5,526 ultra-high resolution helical fragments (7). In effect this (already bioinformatic) scale enables us to analyze long-range water–protein interactions holistically in a multiply curved (or winding) amino acid sequence space, without the limitations imposed by reductionist deconstructions based on individual water–organic solvent transference.

An Instructive Precedent: An Absolute Chemical Scale

By restricting our attention to the context of functionalities of protein families dominated by interactions with water, often (but not necessarily) sharing similar secondary structures, we maximize the knowledge to be gained from scaling. There is ample precedent for this restrictive approach in semiconductor science, a very limited context, yet technologically an incomparably successful field with a database second only to proteins. There a clever combination of deductive and inductive methods produced extremely accurate results 40 years ago, which have not been replicated since by deductive or inductive methods alone (10), not even with the largest, fastest, and most modern supercomputers. The discussion was further restricted to the context of all 70 simple binary octet compounds $A^N B^{8-N}$, where there are 2 broad classes to be considered (10). In 6-fold coordinated salts with ionic bonding (NaCl, CaO), increasing binding energy increases the average energy gap and reduces polarizability, but in 4-fold coordinated semiconductors with covalent bonding (Si, GaAs), increasing binding energy reduces the average energy gap and increases polarizability (because of quantum interference of atomic orbitals in the bonding overlap region). By exploiting the deductive f-sum rule, the dielectric theory of ionicity derives a chemical configuration coordinate that converts the thermodynamically first-order phase transition into one that is second order in this coordinate. This chemical configuration coordinate is the dimensionless fractional ionicity: it is 100% successful in predicting the chemical-phase transition involved in structural separation of these compounds, using as inductive input only the valence electron density and the electronic polarizability. The theory also explains chemical trends in many other properties of these (technologically extremely successful, hence well-studied) materials, including the second-order collapse of shear resistance (called enhanced conformational flexibility by protein scientists) in the AgX salts ($X = \text{Cl, Br, I}$) as the ionicity increases to its critical value. It shows (10) that these chemical trends (previously known qualitatively) can be predicted with high accuracy (usually >10%, for example in heats of formation and elastic constants) by using a configuration coordinate that is accurate to >1%.

The essentially complete success of the dielectric theory of ionicity is based on several subtle points. Had the energies in the ionic salt and covalent semiconductor structures been calculated

separately by conventional reductionist methods, the large cancellations in most of the energy differences would have amplified the effects of small errors in either first principles quantum calculations (even though the separate energies are nearly 100% accurate), or (much worse) parameterized CFF models, leading to $\approx 70\%$ accuracy at best in predicting chemical trends in properties. [These reductionist cancellation problems were already evident in CFF (Born-Mayer) models of binary salts in the 1930s, which is why they were abandoned in favor of quantum theories (10).] The use of the electronic polarizability to construct covalent and ionic energy gaps turned out to be especially felicitous, because the former is measured at optical wavelengths and holistically weights long-range weak interactions more heavily compared with short-range nearest-neighbor strong ones. In proteins energy cancellations are especially challenging, because protein folding energies typically average <1 K/(degree of freedom), and errors in CFF energies are typically of order 100 K/(degree of freedom).

Today, the successes of the dielectric scale of ionicity are so extensive that it is regarded as the Kelvin scale (analogous to his absolute temperature scale) of quantum chemical calculations, which presently are able to reproduce its results with 90% accuracy for the 30 easiest $A^N B^{8-N}$ compounds (10). Here, I argue that a properly constructed (holistic) hydrophobicity scale (restricted to protein contexts) also exists that is dimensionless and can be expected to be almost equally successful on the much larger protein–water canvas, where the need for it is correspondingly greater. Of course, the success of such a scale must be confirmed in practice, but it would be wrong to attempt to do so on a large number of proteins with widely different structures and functionalities: a case-by-case approach is necessary to establish the contexts for which this scale is optimal and to recognize distinctive peak-to-valley hydrophobicity patterns associated with specific functionalities.

Hydrophobic Order Parameters

When the secondary structure of helices, strands, and loops is known, it is convenient to introduce averages over these structural features; these averages are analogous to order parameters defined by averages over sublattices in mixed magnetic crystals. Given the MZ alphanumeric table for hydrophobicity, $-\gamma(\text{aa})$, one can define a suitable average hydrophobicity $\Psi(M, S) = \langle -\gamma(\text{aa})M \rangle$, where the average is taken either over M consecutive residues, or over part of, or an entire secondary element (such as a helix S). This average is similar to the one used by Kyte and Doolittle for membrane proteins (8). A second intersecondary, especially interhelical quantity Φ , is a measure of hydrophobic plasticity or flexibility,

$$\Phi(S) = \Sigma [(\gamma_{\text{aa}}(R(S) - \gamma_{\text{aa}}(R(S + 1)))^2 + (\gamma_{\text{aa}}(R(S) - \gamma_{\text{aa}}(R(S - 1)))^2] / 2M \quad [1]$$

Here, R denotes an amino acid and $\gamma_{\text{aa}}(R)$ is its hydrophobicity. For applications to helical repeats, consecutive repeats are aligned in the standard matrix tableau dictated by the consensus set; the sum is over matched helices of maximum length M , so that both Ψ and Φ are normalized. Note that consensus sites usually contribute 0 to Φ , but that alignment of other sites may either increase or decrease Φ , and several tests have shown that Φ is minimized by standard consensus alignments in cases where the structures of repeat proteins are known. (However, Φ can also be used to align sequences by minimizing the mismatch that Φ measures, thus overcoming some of the ambiguities associated with residue insertions or deletions.) Even in the context of adjacent (nearly parallel nonrepeats, no consensus sites) helices (such as merely H-bonded β strands), Φ potentially could be useful. A general remark: overall there is a tendency toward

arms and relatively soft $\Psi(A)$ “rudder” arms, with a hydro hinge (both arms hydrophilically softened below hydroneutral at 0.155) at the central repeat $S = 8$. The A arms are associated with the exposed convex outer surface, whereas the B arms belong to the concave inner surface, which functions as the scaffolding support for catalytic and regulatory domains in the tumor suppressor protein phosphatase PP2A (18, 20, 21).

The regulatory subunit B56 γ 1 (itself a 16-repeat) attaches (19) to the B arms of the 6 repeats 2–7. The catalytic subunit attaches to the B arms of repeats 11–15, which are associated with large oscillations in Ψ . Thus, the docking structure of the scaffold-regulatory-catalytic heterotrimer, shown in Fig. 2, again from ref. 20 for the reader's convenience, partially reflects the inter-repeat spatial structure of the scaffold alone (the weakening of the helical structure in repeats 8–10 corresponds to the interface or hydrohinge between the regulatory and catalytic units). However, at the same time the aberrant nonparallel packing between repeats 3 and 4, and 12 and 13, apparently caused by divergences from the consensus sequence, probably reflect only the marginal stability of isolated PR65/A. These “breaks” received detailed attention in discussions of the isolated structure, including many amino acid-specific packing interactions (18, 19), but they appear to be irrelevant to the heterotrimer interfaces, which, of course, are tailored by functionality and stabilized in the complex. The break at repeat 12 disappears in the heterotrimeric structure, in which the overall shape of PR65/A changes from a twisted hook to a more regular horseshoe (20).

Next, I turn to the phantom (A, B) hydroplasticity $\Phi_{A,B}$ patterns (Fig. 1B), which are quite different from the spatial packing patterns. Apart from softness near the N end, there is little structure between the outward-facing phobic A arms, but the inward-facing phobic B arms show distinct fragility Φ_B peaks, at $S = 13, 4$, and 12 (in that order). One of the helical consensus sites in the B arm is 24, which is occupied by Val in 11 of the 15 repeats. This hydrophobic amino acid is missing from repeats 1 (N end), 4, 12, and 13, an essentially perfect correlation with the fragility peaks 13, 4, and 12. The marginal fragility of B arms of $S = 8$ and 9 is presumably correlated with the hydrohinge seen in Ψ in Fig. 1A. Overall it is clear that the phantom $\Psi(A, B)$ and $\Phi(A, B)$ long-range adaptive patterns of the helical arms of PR65/A, shown in Fig. 1, provide a better interpretation of the locations and functions of the heterotrimer protein interfaces (20) than the interrepeat spatial patterns (18, 19) of isolated PR65/A. The hydroflexibility of the helical arms enables the intrarepeat loops, which are in close proximity to the regulatory B and catalytic C subunits (20), to adopt optimal conformations. This is also apparent from the narrowing of the A horseshoe in the trimeric complex compared with the free A subunit.

It should be stressed that the essentially perfect matching of the large-scale hydroplasticity patterns in Fig. 1B with the heterotrimer structure (20) can scarcely be accidental, and that this matching is not easily recognized by using conventional methods based on CFF of any kind (10). This success implies that the large-scale structure of the heterotrimer interface is determined both implicitly by the self-organized evolutionary factors that designed the 3 proteins involved and explicitly by long-range protein–water interactions that are accurately described by the MZ scale. By looking closely at the short-range scaffold contacts of intrarepeat turns to the regulatory B and catalytic C subunits shown in figures 3 and 2 of ref. 20, one can see that they can be described as an alternating mixture of rigid hydrophilic salt interactions (anionic Glu and Asp with cationic Arg and Lys) with hydrophobic spacers; the latter could make the large-scale contacts more flexible and adaptive.

The 3-aa turns at the vertex of the groove formed by the intrarepeat turns connecting the A and B arms of PR65/A are solvent-exposed, and it is useful to note that they form 2 distinct hydrophilic groups. Thus, 10 of the turns are strongly hydro-

Fig. 2. Diagram of the overall structure of the heterotrimeric PP2A holoenzyme, from ref. 20. The scaffold subunit A (importin β), the regulatory subunit B, and the catalytic subunit C are in blue, green, and orange, respectively. The details of the matching between repeats A2–A7 and the B subunit, and A11–A15 and the C subunit, are shown in figures 3 and 2, respectively, of ref. 20. [Reproduced with permission from ref. 20 (Copyright 2007, Macmillan Publishers: Nature).]

philic, $\langle\Psi_3\rangle \approx 0.10$, whereas 5 are close to hydroneutral ($\langle\Psi_3\rangle \approx 0.14$, repeats 3, 4, 8, 10, and 14; note that 4 of 5 of these turns are located near 1 of the 3 Φ_B peaks). The latter may promote stability of the marginally stable scaffolding structure in the unbound state. It appears that these vertex turns determine a characteristic length in repeat proteins that makes $\langle\Psi_3\rangle$ a valuable configuration coordinate for repeats composed of L-shaped building blocks.

Four oncogenic site mutations associated with lung and colon tumors have been discussed in terms of the static structure (19) but here one can propose an alternative interpretation. Two of the mutations occur in helical sites, and these 2 ($P^{65} \rightarrow S$ and $L^{101} \rightarrow P$) involve substantial destabilizing decreases in hydrophobicity by reducing helical rigidity. The other 2 occur in the short turns, $D^{504} \rightarrow G$ (repeat 13) and $V^{545} \rightarrow A$ (repeat 14). The former increases $\langle\Psi_3\rangle$ (504), thus hydrophobically strengthening the typically hydrophilic repeat 13 turn, whereas the latter reduces $\langle\Psi_3\rangle$ (545) and hydrophilically weakens the atypically nearly hydroneutral repeat 14 turn; both mutations regress $\langle\Psi_3\rangle$ for these turns toward the mean value (0.118) for all 15 turns. All of these small changes could combine to favor more rapid production of the oncogenic unbound protein, while disrupting its functionality, by averaging barrier heights. Although the individual mutagenic changes are small, acting cooperatively they holistically (collectively) smooth water–protein interactions at both the helices and the solvent-exposed turns and could easily accelerate oncogenic production; this cooperative phantom mechanism is spatially invisible and is also inaccessible to conventional CFF simulations.

In another article (33) we will see that large-scale binding of transport repeat proteins is both hydrophobically- and charge-driven, but note here that the $(K + R - D - E)$ charges on the A and B arms of PR65/A are small [near ± 1 for (B, A) helical arms] and there are no long basic loops: large-scale charge effects are weak for proteins binding to this scaffold. Another repeat with mechanically driven binding interactions and the characteristic L shape with A and B helical arms is the ankyrin D34 section. Ankyrins are superhelical springs or multifunctional adaptors that link specific proteins to the membrane-associated, spectrin-actin cytoskeleton (22); in D34 A and B arm charges are small and evidence almost no correlation. Hydrophobic analysis of D34 has been given elsewhere (10, 23). Solenoidal repeats also function as scaffolds and support a wide variety of functions through gradually more divergent sequences (13, 24). The spatial curvature created by the superhelical nature

1. Bak P, Tang C, Wiesenfeld K (1987) Self-organized criticality: An explanation of the $1/f$ noise. *Phys Rev Lett* 59:381–384.
2. Sornette D (1994) Sweeping of an instability: An alternative to self-organized criticality to get power laws without parameter tuning. *J Phys I [French]* 4:209–221.
3. Chakravarty S, Georgiev DG, Boolchand P, Micoulaut M (2005) Aging, fragility, and the reversibility window in bulk alloy glasses. *J Phys Condens Matter* 17:L1–L7.
4. Rompicharla K, et al. (2008) Abrupt boundaries of intermediate phases and space filling in oxide glasses. *J Phys Condens Matter* 20:202101.
5. Abel DL, Trevors JT (2006) Self-organization vs. self-ordering events in life-origin models. *Phys Life Rev* 3:211–228.
6. Dokholyan NV (2005) The architecture of the protein domain universe. *Gene* 347:199–206.
7. Moret MA, Zebende GF (2007) Amino acid hydrophobicity and accessible surface area. *Phys Rev E* 75:011920.
8. Kyte J, Doolittle RF (1982) A simple method for displaying the hydropathic character of a protein. *J Mol Biol* 157:105–132.
9. Chandler D (2005) Interfaces and the driving force of hydrophobic assembly. *Nature* 437:640–647.
10. Phillips JC (2008) Protein sequence, structure, stability, and functionality, <http://xxx.lanl.gov/abs/0802.3641>.
11. Salvi G, De Los Rios P (2003) Effective interactions cannot replace solvent effects in a lattice model of proteins. *Phys Rev Lett* 91:258102.
12. Seno F, Trovato A, Banavar JR, Maritan A (2008) Maximum entropy approach for deducing amino acid interactions in proteins. *Phys Rev Lett* 100:078102.
13. Kobe B, Kajava AV (2000) When protein folding is simplified to protein coiling: The continuum of solenoid protein structures. *Trends Biochem Sci* 25:509–515.
14. Pawar AP, DuBay KF, Zurdo J, Vendruscolo M, Dobson CM (2005) Prediction of “aggregation-prone” and “aggregation-susceptible” regions in proteins associated with neurodegenerative diseases. *J Mol Biol* 350:379–392.
15. Andrade M, Petosa C, O'Donoghue S (2001) Comparison of ARM and HEAT protein repeats. *J Mol Biol* 309:1–18.
16. Cutter AD, Wasmuth JD, Blaxter ML (2006) The evolution of biased codon and amino acid usage in nematode genomes. *Mol Biol Evol* 23:2303–2315.
17. Kobe B, Kajava AV (2001) The leucine-rich repeat as a protein recognition motif. *Curr Opin Struct Biol* 11:725–732.
18. Kobe B, Gleichmann T, Horne J, Jennings IG, Scotney PD, Teh T (1999) Turn up the HEAT. *Structure (London)* 7:R91–R97.
19. Groves MR, Hanlon N, Turowski P, Hemmings BA, Barford D (1999) The structure of the protein phosphatase 2A PR65/A subunit reveals the conformation of its 15 tandemly repeated HEAT motifs. *Cell* 96:99–110.
20. Cho US, Xu WQ (2007) Crystal structure of a protein phosphatase 2A heterotrimeric holoenzyme. *Nature* 445:53–57.
21. Mumby M (2007) PP2A: Unveiling a reluctant tumor suppressor. *Cell* 130:21–24.
22. Michaely P, Tomchick DR, Machius M, Anderson RGW (2002) Crystal structure of a 12 ANK repeat stack from human ankyrinR. *EMBO J* 21:6387–6396.
23. Phillips JC (2008) Functionality and protein–water interactions, <http://xxx.lanl.gov/abs/0803.0091>.
24. Karpenahalli MR, Lupas AN, Soding J (2007) TPRpred: A tool for prediction of TPR-, PPR- and SELI-like repeats from protein sequences. *BMC Bioinform* 8:2.
25. Lubchenko V (2008) Competing interactions create functionality through frustration. *Proc Natl Acad Sci USA* 105:10635–10636.
26. Shakhnovich E (2006) Protein folding thermodynamics and dynamics: Where physics, chemistry, and biology meet. *Chem Rev* 106:1559–1588.
27. Israelachvili JN, Pashley RM (1982) The hydrophobic interaction is long-range, decaying exponentially with distance. *Nature* 300:341–342.
28. Yamaotsu N, Oda A, Hirano S (2008) Determination of ligand-binding sites on proteins using long-range hydrophobic potential. *Biol Pharm Bull* 31:1552–1558.
29. Yan H, et al. (2008) Thermoreversible lysozyme hydrogels: Properties and an insight into the gelation pathway. *Soft Matter* 4:1313–1325.
30. Parmeggiani F, et al. (2008) Designed armadillo repeat proteins as general peptide-binding scaffolds: Consensus design and computational optimization of the hydrophobic core. *J Mol Biol* 376:1282–1304.
31. Ley TJ, et al. (2008) DNA sequencing of a cytogenetically normal acute myeloid leukemia genome. *Nature* 456:66–72.
32. Hofmann K, Stoffel W (1992) PROFILEGRAPH: An interactive graphical tool for protein sequence analysis. *Bioinformatics* 8:331–337.
33. Phillips JC (2009) Scaling and self-organized criticality in proteins II. *Proc Natl Acad Sci USA* 10.1073/pnas.0811308106.