Sequence variability of proteins evolutionarily constrained by solution-thermodynamic function

F. N. Braun

Stockholm Bioinformatics Center, Department of Biochemistry and Biophysics, Stockholm University, 10691 Stockholm, Sweden

(Received 10 June 2003; revised manuscript received 5 September 2003; published 16 January 2004)

Focusing on silk fibroin and hemoglobin molecules as templates, we model protein homolog dispersal across sequence-fitness landscapes determined by solution thermodynamics. Landscapes are constructed by inspecting an idealized theoretical phase topology associated with sequence length and hydrophobic-polar composition, comprising liquid-liquid phase separation, gelation and liquid crystalline self-assembly. We then calculate the distribution of homologs in sequence space as steady states of a simple mutation-selection dynamics. The results are consistent with Swiss-Prot bioinformatic data.

DOI: 10.1103/PhysRevE.69.011903 PACS number(s): 87.14.Ee, 87.23.Kg, 81.30.Dz

I. INTRODUCTION

Much of population genetics' neodarwinian synthesis, evolution as a combination of mutation, drift, and selection, is encompassed by Sewall Wright's concept of a fitness landscape extending over the space of mutation-accessible genotype [1,2]. A very intuitive fundamental corollary of the landscape perspective is that evolving populations generally maintain a degree of genetic variability.

Several authors have applied landscape methods to protein evolution. By adopting some specific choice of sequence-fitness mapping, one is able to predict a ''homolog distribution'' describing sequence-space dispersal of a given protein within a population. Chan and Bornberg-Bauer [3] have reviewed simple exact model (SEM) views of protein folding which are amenable to evolutionary analysis in this respect. Chosen measures of fitness have included native state stability [4,5] and enzymelike ligand binding affinity [6,7].

We aim here to address in an analogous spirit functional contexts which derive instead from the thermodynamics of proteins interacting as ensembles in solution, i.e., a departure from the emphasis thus far on function of a single protein molecule. To this effect, we consider solution-thermodynamic phenomena which can be transparently coupled to sequence. By assigning fitness to distinct phase-topological features, the sequence-space projection of the phase diagram becomes implicitly a fitness landscape.

II. BIOLOGICAL MOTIVATION

A number of aggregation-related phenomena have physiological relevance to globular protein systems, in particular, phase separation and gelation. Lysozyme and eye lens crystallins exhibit a well-studied liquid-liquid phase separation, which is apparently driven by dispersion interactions [8]. In hemoglobin solutions, phase separation is believed to present a possible route to the pathological fibrillization of HbS mutants responsible for sickle cell anemia [9]. This latter contrasts with that of lysozyme etc. in so far as it occurs with increasing temperature (''temperature-reversed'' [10]), readily interpreting as a hydrophobicity-driven effect [11]. Both mechanisms are to an extent generic to globular protein solutions, although whether they are of importance under physiological conditions depends upon system-specific parameters.

The hydrophobic mechanism of hemoglobin is relevant to sequence evolution in so far as the associated critical point is strongly responsive to those genetic mutations which change polarity of the encoded amino-acid residues. As we shall discuss below, it can be treated qualitatively after the fashion of the HP school of SEMs, revolving around a simple two-state characterization of residues, H=hydrophobic vs P=polar.

Eaton and Hofrichter [12] have described the fibrillized phase of HbS solutions as a polymeric gel. In fact, gelation has become something of a blanket term encompassing such statistical-mechanically distinct phenomena as percolation [13] and glass transitions [14]. Nevertheless, light scattering studies of globular proteins consistently suggest that phase separation is generically accompanied by some form of gel-like phase [10], regardless of its precise nature. This generic aspect is sufficiently captured by a simple percolation perspective which we are able to develop within the same theoretical framework introduced to describe phase separation.

In the case of hemoglobin, it is plausible to regard the evolutionary contingency of phase separation and gelation as equivalent to that of sickle cell anemia. We will implement a loose interpretation of hemoglobin as an explicit model template for illustration of our overall method, roughly observing this contingency. Phase separation and gelation will translate as unviable holes in the fitness landscape of the approach, tantamount to assuming that the sickle cell condition is deleterious [15]. Aside from the case of hemoglobin, however, this may represent a fairly general constraint common to many other protein contexts, soluble globular proteins in particular. Cataract formation, for example, another well-studied disease pathology, involves phase separation of the lens crystallins already mentioned.

As a counter illustration, a phase which might be positively selected for, we will also incorporate a liquid-crystalline region into our general phase diagram, following a recent theory which relates silk fibroin sequence to self-assembly in solution [16,17]. It is thought that orb-weaver spiders (genus *Nephila*) exploit such a process to optimize the mechanical quality of dragline silks spun from their major ampullate gland [18]. Although Nephila dragline fibroins belong to the class of larger fibrous proteins, their behavior in the gland resembles that of the globular systems above to

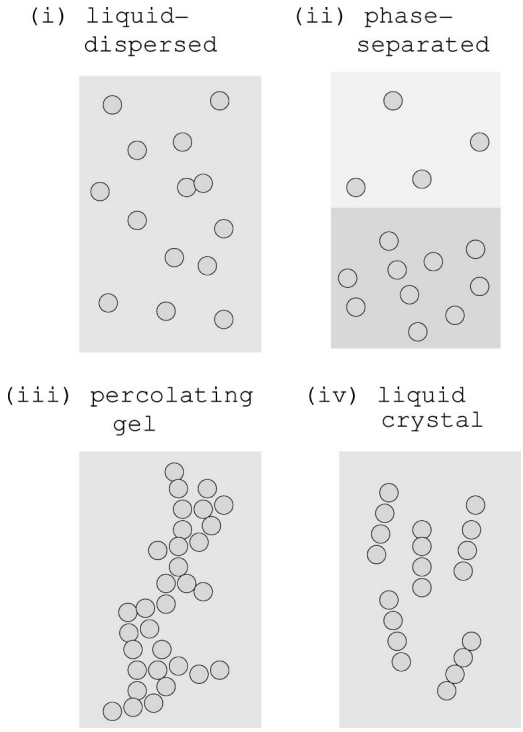

FIG. 1. Protein solution-thermodynamic phenomena which we map to evolutionary fitness in this discussion. Hydrophobicity drives a generic phase separation into two coexisting colloidal-like liquid phases (ii). Hydrophobicity can also drive the onset of gelation, interpreted here in a percolation sense (iii), and self-assembly of supramolecular fibrils/mesogens capable of liquid-crystalline phase ordering (iv).

the extent that individually they adopt a collapsed globular conformation, and avoid phase separation and gelation. Hydrophobicity drives their assembly into supramolecular mesogens, capable of liquid-crystalline ordering (see Fig. 1).

III. COUPLING SEQUENCE TO THERMODYNAMIC FITNESS

In this section we first introduce an idealized mapping between amino-acid sequence and an adhesive protein-protein interaction parameter. The thermodynamic phase space of this interaction is developed in Sec. III B, featuring phase separation, gelation, and a liquid crystalline region as discussed above. Section III C completes the passage to a sequence-fitness landscape.

A. mapping between sequence and globule-globule interaction

The free energy gain of burying hydrophobic (nonpolar) surface area in the interface between protein subunits is of the order of kT/a^2 per unit buried area [19], where $a \sim 1$ nm is a typical amino-acid lengthscale and kT is the thermal energy. Hence, if a fraction f_H of the residues exposed at the surface of a given globular protein are hydrophobic, then the free energy gain in forming a globule-globule contact of area $A \sim a^2$ is

$$\epsilon \sim f_H kT. \tag{1}$$

Let us formulate this as a square-well interaction, having well-depth ϵ,

$$\begin{aligned} v(r) &= \infty, \quad 0 < r < d, \\ &-\epsilon, \quad d < r < d + a, \\ &0, \quad r > d + a, \end{aligned} \tag{2}$$

where d denotes the diameter of the globules. We may assume roughly $d \simeq N^{1/3} a$ for a sequence comprising N residues. We adopt the amino-acid lengthscale a as the well width, which is a natural choice on the assumption that it's hydrophobicity of individual residues which drives the interaction.

Sequence couples to ϵ via the parameter f_H. Thus, if we can map between sequence and f_H, then this will be tantamount to a mapping between sequence and globule-globule interaction. A simple way to approach this problem is to ignore all aspects of the folding problem other than the free energy of coarsely partitioning hydrophobic residues between core and solvent-exposed regions of the globule. We proceed along similar lines to an early approach of Dill [20], as implemented in Ref. [16].

The N residues of the sequence are inscribed on a roughly spherical lattice, which is compact in the sense that all sites are occupied. Of the N residues, it is easy to show that $\approx N_e = 3N^{-1/3}$ lie in the surface such that they are exposed to the solvent. Conversely, $N_b = N - N_e$ residues buried in the core are protected from the solvent. Next, we define n as the number of residues which are hydrophobic (H) as opposed to the $N-n$ polar (P) residues. In a given conformation, the globule partitions such that n_b of the n hydrophobic residues are buried, while $n_e = n - n_b$ are exposed. A ''native partitioning'' follows by minimizing a Gibbs free energy

$$G = n_e \Delta - TS. \tag{3}$$

The first term expresses the hydrophobic effect, with Δ a free energy cost per H residue exposed to the solvent. We set $\Delta = 2kT$ in our calculations. This term drives the H residues into the core, but at a cost in distributional entropy,

$$S/k_B = \ln\left[\frac{N_b!}{n_b!(N_b - n_b)!}\right] + \ln\left[\frac{N_e!}{n_e!(N_e - n_e)!}\right]. \tag{4}$$

The free energy minimization determines $f_H = \overline{n_e}/N_e$, where $\overline{n_e}$ denotes the native partitioning. To complete the mapping between sequence and well-depth ϵ, we write

$$\epsilon = f_H \Delta = \overline{n_e} N^{1/3} \Delta / 3. \tag{5}$$

Figure 2 presents the trend calculated by this method for our two model templates, ''hemoglobin'' having total sequence length $N = 574$ (the tetramer), and Nephila dragline ''silk fibroins'' having $N = 3500$.

It is important to qualify the assumption of compactness as understood in this derivation. Heteropolymer-like considerations [20] suggest a breakdown in compactness below some threshold sequence hydrophobic content n/N, effec-

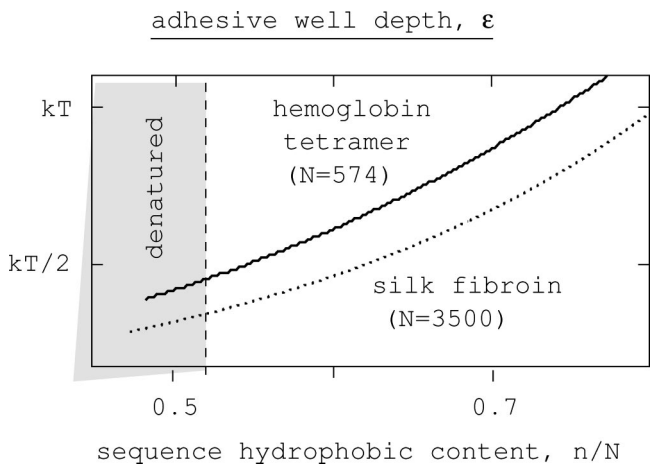

FIG. 2. Phase behavior is driven by an effective adhesive interaction of strength ϵ, which derives from sequence hydrophobic content n/N. The lower curve fixes $N=3500$, reflecting dragline silk fibroins. The upper curve fixes $N=574$, the number of residues in the hemoglobin tetramer. This mapping assumes the interacting proteins adopt a globular-compact conformation. Compactness is lost when the globules denature below some threshold composition, which is roughly independent of N according to the argument given in the Appendix.

tively a denaturation transition. The dashed line of the figure presents an explicit calculation, as outlined in the Appendix.

B. Thermodynamic phase space of adhesive interaction

The square-well fluid is not particularly tractable from an equation of state point of view. However, exact and compact analytical results follow in the limit formulated by Baxter [21],

$$v(r)=\infty, \quad 0<r<d,$$

$$-kT\ln\left[\frac{d+\sigma}{12\tau\sigma}\right], \quad d<r<d+\sigma,$$

$$0, \quad r>d+\sigma, \tag{6}$$

where $\sigma\to 0$.

Baxter's parameter τ^{-1} can by regarded as a measure of adhesive strength, as is clear from its relation to the second virial coefficient

$$\Delta B_2/B_2^{HS}=3d^{-3}\lim_{\sigma\to 0}\int_d^{d+\sigma}[1-\exp(-\beta v(r))]r^2dr$$

$$=-\tau^{-1}/4, \tag{7}$$

where the bare hard-sphere result $B_2^{HS}=2\pi d^3/3$ presents a convenient reference.

The idea of exploiting the analytical tractability of Baxter's limit to represent square-well-like systems is a familiar one in the general colloidal context. We follow here the equivalence prescription of Regnaut and Ravey [22], who fix a correspondence between τ^{-1} and well depth by equating the respective second virial coefficients. To lowest order in a/d we have for the square-well fluid,

$$\Delta B_2/B_2^{HS}\simeq -3(a/d)[\exp(\beta\epsilon)-1].$$

Hence, comparing with Eq. (7),

$$\tau^{-1}=12(a/d)[\exp(\beta\epsilon)-1]. \tag{8}$$

1. Liquid-liquid phase separation

The static structure factor for Baxter particles has an analytical form [22],

$$Q=\frac{(1-\varphi)^4}{[1+2\varphi-\lambda\varphi(1-\varphi)]^2}, \tag{9}$$

where φ is the volume fraction, and λ is the lower root of

$$\frac{\varphi}{12}\lambda^2-\left(\frac{\varphi}{1-\varphi}+\tau\right)\lambda+\frac{1+\varphi/2}{(1-\varphi)^2}=0. \tag{10}$$

A locus of divergence of Q follows straightforwardly, which, in our present colloidal-like context, can be interpreted as a liquid-liquid transition spinodal. The critical point is located at $\tau_c=1.17$, $\varphi_c=0.12$.

2. Gel (percolation) transition

In a cluster-based structural view of an adhesive fluid, the term percolation describes divergence of the mean cluster size, say m. Cluster size is formally related to a pair-connectedness function $P(r)$ analogous to the usual pair distribution function $g(r)$,

$$m=1+c\int P(r)d\mathbf{r}, \tag{11}$$

where c is the particle concentration and $c^2P(\mathbf{r}_1,\mathbf{r}_2)d\mathbf{r}_1d\mathbf{r}_2$ is the probability of finding connected particles in volume elements $d\mathbf{r}_1$ and $d\mathbf{r}_2$ simultaneously. By recasting Baxter's original $g(r)$ method in connectivity language, Chiew and Glandt [23] solve for $P(r)$, yielding

$$m=1/(1-\lambda\varphi)^2, \tag{12}$$

hence percolation along the locus $\lambda\varphi\to 1$.

3. Liquid crystallinity

Supramolecular self-assembly is common in protein solutions, and occurs via a range of different mechanisms. In their capacity as sterically anisotropic mesogens, the assembled structures are sometimes capable of spontaneously ordering into a liquid-crystalline phase.

An example of such behavior is thought to occur during the early stages of the orb-weaver spider's dragline spinning process. In modeling this process, Braun and Viney [16] assume that the constituent fibroins, while dispersed in the gland prior to extrusion, adopt a collapsed conformation. These ''fibrous globules'' assemble reversibly into supramolecular rodlike structures, driven by an adhesive interac-

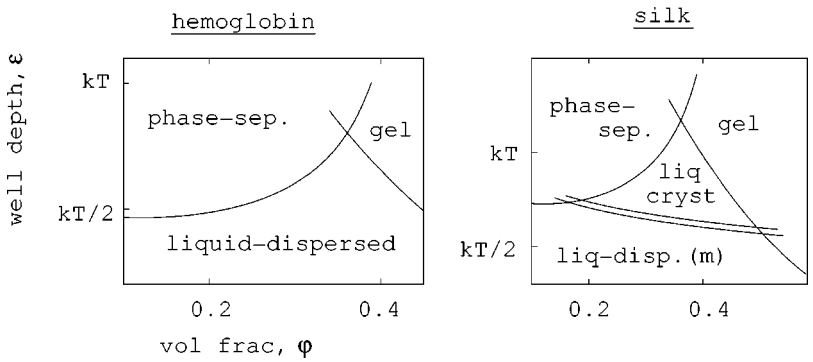

FIG. 3. Phase space of adhesive interaction. The hemoglobin diagram reproduces a schematic topology outlined by Muschol and Rosenberger from experimental data [10]. These features are arguably quite generic, since sequence length $N=574$ is the only hemoglobin-specific parameter in the calculation. The silk phase diagram is sensitive to an additional parameter $s=10$ specifying the morphology of supramolecular mesogens. The liquid-dispersed (m) phase now comprises disordered supramolecular mesogens, separated by a biphasic region from an ordered liquid-crystalline phase of the nematic type.

tion having the same sequence-hydrophobic origin as our present ϵ.

Subject to a morphological constraint s, specifying the number of monomeric strands in the cross section, the rods have a mean axial ratio

$$L \simeq s^{-3/2}\varphi^{1/2}\exp(E_{sc}/2kT), \tag{13}$$

where $E_{sc}\sim s\epsilon$ is an effective scission energy, the cost of slicing a rod in two.

Beyond some critical value of L the rods undergo a transition from a disordered isotropic state to an ordered liquid-crystalline state of the nematic type. Nematic-isotropic coexistence follows according to a well-known theory of Onsager [24], by solving criteria defining, respectively, the isotropic (I) and nematic (N) nodal lines bounding the biphasic region,

$$L(\varphi_I)=3.3/\varphi_I, \quad L(\varphi_N)=4.5/\varphi_N. \tag{14}$$

C. Phase-determined fitness landscape

In Fig. 3 we have calculated the various transition loci discussed above to present unified phase diagrams describing hemoglobin and silk fibroin solutions, respectively. The toy nature of this description should be stressed. The physical chemistry of both hemoglobin and silk fibroins is of course far more complex in their respective biological environments, the cell (see, e.g., Ref. [9]) and the gland, involving salts, other macromolecules, nonequilibrium, and so forth. Moreover, we have neglected further phase topology, for example glassy and crystalline solid phases, which could conceivably be developed within the idealized framework.

Despite their simplicity, these phase diagrams have phenomenological merit in so far as they present a straightforward implicit relation between sequence and phase, via the mapping between adhesive interaction and sequence defined in Sec. III A. Given the volume fraction of a protein in solution, its length and hydrophobic content, we can determine the phase.

Figure 4 establishes the link to evolution, by assigning fitness values to the respective phases. The result is a fitness landscape stretching over the space of hydrophobic content n/N vs solution volume fraction. Phase-separation, gelation, and denaturation reside at sea level, i.e., are deleterious, having zero fitness. The liquid-dispersed phase appears by contrast as a viable plateau. In the case of silk, liquid crystallinity is represented as a higher secondary plateaulike outcrop [25].

In a more general approach, further solution-thermodynamic dimensions to the landscape would naturally

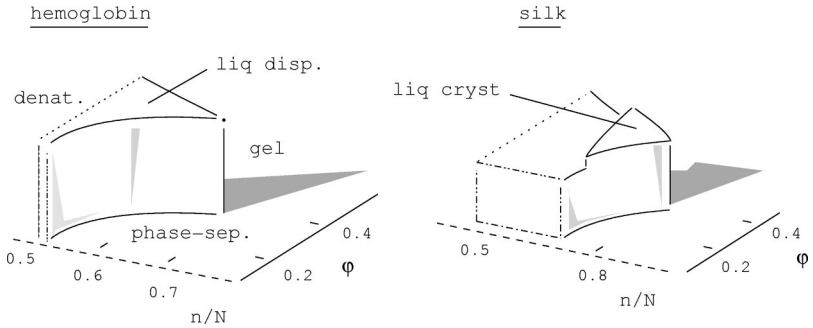

FIG. 4. Sequence-concentration-fitness landscapes inferred from Figs. 2 and 3. Phase-separation, gelation, and denaturation are assigned zero fitness, such that they represent unviable sea-level areas of the landscape. We represent the liquid-dispersed phase as a viable plateau. In the hemoglobin scenario, evolution is dominated by negative (purifying) selection, i.e., removal from the population of mutants which stray from the liquid-dispersed plateau. On the silk landscape, the contrasting positive mode of selection dominates, driving the population onto the higher liquid-crystalline outcrop.

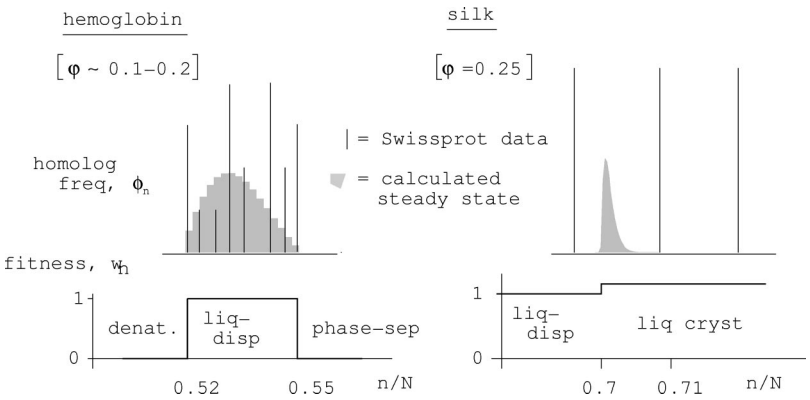

FIG. 5. Calculated steady states (shaded) of the simple mutation-selection dynamics of Eq. (15), applied over constant-φ projections w_n of the landscapes. The agreement with taxon-sampled homolog distributions from the Swiss-prot database (impulses) is good for both hemoglobin and silk. Human hemoglobin is situated at the phase-separation edge of w_n, in line with its sickle cell propensity. The positively selected silk population is marginally optimal in so far as it crowds at the edge of the liquid-crystalline plateau.

develop—temperature, salinity, and so forth. Note that the landscape is athermal at the present level of description, because the adhesive interaction from which it derives has no enthalpic part. In fact, measured hydrophobic free energies of transfer do generally have a (small) enthalpic part [26], which we have neglected in setting $\Delta = 2kT$, in Sec. III A.

IV. THE HOMOLOG DISTRIBUTION

We interpret the homolog distribution as the normalized frequency ϕ_n of proteins in an infinite population, having n hydrophobic residues vs $N-n$ polar residues. Following in the spirit of Bornberg-Bauer and Chan [4], evolution will be understood in the sense of deterministic iteration of ϕ_n through discrete generational steps $t=1,2$, etc.

At each step, a fraction θ of the population mutates

$$\Gamma \phi_n(t+1) = (1-\theta) w_n \phi_n(t) + \theta \sum_{n'} M_{nn'} w_{n'} \phi_{n'}(t), \tag{15}$$

where Γ is a normalization factor guaranteeing $\Sigma_n \phi_n(t+1) = 1$. The matrix $M_{nn'}$ defines the relative probabilities of sequence mutations $n' \to n$. Subject to the condition that a given gene can only undergo at most a single HP flipping substitution per generation (i.e., a gene with sequence n' can only mutate to $n=n'+1$ or $n=n'-1$), we can show straightforwardly

$$M_{nn'} = \left(\frac{n'}{N}\right) \delta_{n,n'-1} + \left(1 - \frac{n'}{N}\right) \delta_{n,n'+1}. \tag{16}$$

For the fitness function w_n we take some physiologically plausible constant-φ projection of the solution-thermodynamic landscape discussed above. In so far as we do not treat φ as an evolutionary degree of freedom, we defer the possibility of linkage to regulatory genes controlling expression [27,28].

A. Hemoglobin

In the case of hemoglobin, the interval $\varphi \sim 0.1-0.2$ is reasonably physiological [29]. With reference to Fig. 4, we note that over this interval the phase-separation spinodal is almost flat, remaining around $n/N \simeq 0.55$ as it passes through the critical point at $\varphi_c = 0.12$.

The fitness projection exhibits hence a deleterious edge at $n/N \simeq 0.55$ as shown in Fig. 5. This is consistent with the human hemoglobin sequence data available from the Swiss-Prot database. By counting hydrophobic residues, we estimate for the healthy wild type $n/N \simeq 0.55$ such that it resides on the plateau right up against the edge. The sickle cell mutant HbS differs from the wild type by a single P to H substitution (Glu to Val) [30]. According to our picture, this is sufficient to cause phase separation, pushing the wild type to the right and off the edge.

A second lower edge of w_n, located at $n/N = 0.52$ in the figure, corresponds to the onset of denaturation. The resulting steady-state homolog distribution, calculated from Eq. (16), is shown directly above (shaded), its peak frequency wedged roughly midway between the w_n edges. Note that this steady state is independent of mutation rate θ, as holds generally for a plateau landscape in which the only mode of selection is ''negative.'' Negative selection refers to the removal of unviable mutants from a population [31].

We see substantiated the idea that protein function is manifest in some signature variability. Gu [32] and Naylor and Gerstein [33] have developed the corollary that a shift in variability measured between clades is indicative of diverging function. Their interpretation is based on variability profiles for each individual amino-acid site along a sequence. Our focus on the simpler HP compositional space presents an alternative perspective.

Superimposed against the predicted homolog distribution, we show an ''experimental'' distribution, taxon-sampled from the Swiss-prot database. The data set comprises 20 species from the primate clade [33]. With the exception of the anemia-prone human, gorilla and langur tetramers, this data set lies a relatively safe phylogenetic distance away from the pathological edge, roughly as anticipated by the model. However, the comparison should be regarded as informal only. We defer nontrivial fundamental aspects, such as to what extent a clade-wide distribution can be considered equilibrated in the steady-state sense of an infinite population dynamics.

B. Silk

We set $\varphi = 0.25$, approximately corresponding to the *in vitro* volume fraction at which Nephila major-ampullate solutions exhibit liquid crystallinity [34]. The resulting ho-

molog population, also shown in Fig. 5, sits at the extreme edge $n/N=0.7$ of the liquid-crystalline region of w_n overlooking the lower liquid-dispersed plateau. Only three dragline-sequenced members of the Nephila clade are available for the comparison [35], but the location and relative narrowness of the calculated distribution is again roughly corroborated.

The ascent of the population onto the fitter liquid-crystalline outcrop constitutes an adaptive response. This mode of evolutionary dynamics is known as positive selection, distinct from the negative mode which dominates the hemoglobin steady state. A positively selected steady state is sensitive in general to the balance between mutation rate θ and the selection differential δw [2]. Figure 5 is calculated in the limit $\delta w \gg \theta$. A qualitatively contrasting steady state results, on the other hand, if the selection differential falls below the so-called error threshold $\delta w \sim \theta$ [36]. Selection is then effectively washed out by ''neutral'' mutation [37]. In the dynamics specified by Eq. (15), neutral mutation drives the homolog population towards the binomial limit,

$$\phi_n = (\pi N/2)^{-1/2} \exp\left[-\frac{2}{N}(n-N/2)^2 \right] \quad \text{(neutral steady state)}. \tag{17}$$

This neutral distribution is equivalently a measure of intrinsic designability of a given HP composition. According to a definition of Bornberg-Bauer and Chan, the maximally designable ''prototype'' over a neutral mutation space is that having the most nearest neighbors, i.e., sequences lying a single mutational step away. Thus in the present approach, the binomial mean $N/2$ corresponds trivially to ''prototype composition.''

In so far as populations tend to evolve towards design prototypes, a possible explanation for the emergence of marginal folding stability, as generally observed for globular proteins, is simply that the neutral prototype is unstable [38]. The silk example demonstrates that analogous considerations can also apply in respect of solution-thermodynamic fitness. The prototype lies far away from the fittest liquid-crystalline outcrop of the landscape. The homolog distribution is able to maximize its design entropy, nevertheless, by crowding towards the edge of the outcrop overlooking the prototype. We can describe Nephila silk fibroins as ''marginally'' optimal in this sense.

V. CONCLUSION

A comprehensive union of population genetics with protein folding, solution/cell biochemistry etc. is not generally feasible as a quantitative means of predicting protein evolution [39]. Simple exact models (SEM's), making tractable an idealized union of sorts, have been exploited in phenomenological investigations of genomic mechanisms such as duplication [5,7], recombination [3], and in the elucidation of general trends in bioinformatic data [40].

The particular framework developed here, while retaining elements of the traditional SEM emphasis on folding, broaches protein evolutionary contexts having to do with their macroscopic phase behavior. This presents on one hand an extension of the SEM paradigm towards a population genetics of proteins in the quasimacroscopic environment of the cell. In a template illustration, we constrained the evolution of model hemoglobins by invoking phenomena associated with sickle cell anemia; phase separation and gelation.

Despite the importance to disease pathologies, solution thermodynamics is not exactly stringent in our description. Proteins thus constrained are evolutionarily quite plastic, having mutational access to a large extent of nondeleterious sequence space. This has relevance to the important genomic mechanism so-called neofunctionalization, where redundant gene duplicates assume some novel function [41]. In the case of eye lens crystallins, weak solution-thermodynamic constraints similar to those envisaged here have apparently facilitated their co-option from disparate duplicated sources, ranging from heat-shock proteins to metabolic enzymes [42].

Finally, weak constraints allow sequence to evolve towards design prototypes. ''Marginal'' solution-thermodynamic fitness, illustrated by our silk model template, emerges in situations where selection for phase behavior competes against prototype.

ACKNOWLEDGMENTS

It is a pleasure to thank David Liberles, Christopher Viney, and Erik Sandelin for discussions, and Bergen University's Computational Biology Unit for their hospitality during preparation of the manuscript. Funding for this work was provided by the Swedish Foundation for Strategic Research.

APPENDIX: DENATURING TRANSITION

Consider a heteropolymerlike conformational free energy similar to that first introduced by Dill [20] comprising three terms

$$F = F_1(\text{hydrophobicity}) + F_2(\text{hydration}) + F_3(\text{polymeric elasticity}). \tag{A1}$$

The first term expresses an effective attraction between the n hydrophobic residues, governed by a Flory parameter χ,

$$F_1/NkT = -\rho\chi\left(\frac{n}{N}\right)^2, \tag{A2}$$

where ρ denotes residue packing fraction.

The hydration term accounts for the distributional entropy of solvent molecules trapped within the conformation,

$$F_2/NkT = \frac{1-\rho}{\rho}\ln(1-\rho). \tag{A3}$$

The third term expresses the cost of stretching the random coil conformation entropically favored by polymeric connectivity, $R^2 \sim Na^2$, where R is the mean end-to-end distance and a is a characteristic residue dimension

$$F_3/kT = \frac{3R^2}{2Na^2}. \tag{A4}$$

Writing $\rho = 3Na^3/4\pi R^3$, and expanding F_2, we obtain for Eq. (A1)

$$F[\rho]/NkT \simeq A\rho + B\rho^2 + C\rho^{-2/3}, \tag{A5}$$

where

$$A = 1/2 - \chi\left(\frac{n}{N}\right)^2, \quad B = 1/6, \quad C = \frac{3}{2}\left(\frac{3}{4\pi}\right)^{2/3} N^{-4/3}. \tag{A6}$$

The general form of Eq. (A5) admits various scaling regimes [43]. If $A=0$, the protein is in the random coil state, scaling as $\overline{\rho} \sim N^{-1/2}$. If $A>0$ the protein swells to a lower packing fraction, scaling as $\overline{\rho} \sim N^{-4/5}$. We label this swollen state "denatured." From Eq. (A6), denaturation occurs below

$$\frac{n}{N} < \frac{1}{\sqrt{2\chi}} \quad \text{denatured.} \tag{A7}$$

In the converse situation $A<0$, the protein is globular compact, $R^3 \sim Na^3$

$$\frac{n}{N} > \frac{1}{\sqrt{2\chi}} \quad \text{globular} \tag{A8}$$

(Note, since $\overline{\rho} \sim N/R^3$, there is no N scaling of $\overline{\rho}$ in this compact regime).

The parameter χ is related to, but not the same as Δ of Sec. III A. Recall Δ is set to $2kT$ in our calculations of adhesive interaction ϵ. Choice of χ affects the model fit to the Swiss-prot hemoglobin data, Fig. 5. We obtained a good fit with $\chi = 1.85kT$, corresponding to denaturation at $n/N = 0.52$ (Fig. 2).

[1] S. Wright, in *Proceedings of the Sixth International Congress on Genetics*, edited by D. F. Jones (Brooklyn Botanic Gardens, New York, 1932), Vol. 1.

[2] B. Drossel, Adv. Phys. **50**, 209 (2001).

[3] H.S. Chan and E. Bornberg-Bauer, Appl. Bioinform. **1**, 121 (2002).

[4] E. Bornberg-Bauer and H.S. Chan, Proc. Natl. Acad. Sci. U.S.A. **96**, 10689 (1999).

[5] D. M. Taverna and R. A. Goldstein, in *Proceedings of the 2000 Pacific Symposium on Biocomputing 2000, Honolulu*, edited by R. B. Altman, A. K. Dunker, L. Hunter, and T. E. Klein (World Scientific, Singapore, 2000).

[6] B.P. Blackburne and J. Hirst, J. Chem. Phys. **115**, 1935 (2001).

[7] F. N. Braun and D. A. Liberles, Intl. J. Biol. Macromol. **33**, 19 (2003).

[8] O. Annunziata, O. Ogun, and G.B. Benedek, Proc. Natl. Acad. Sci. U.S.A. **100**, 970 (2003).

[9] O. Galkin, K. Chen, R.L. Nagel, R.E. Hirsch, and P.G. Vekilov, Proc. Natl. Acad. Sci. U.S.A. **99**, 8479 (2002).

[10] M. Muschol and F. Rosenberger, J. Chem. Phys. **107**, 1953 (1997).

[11] F.N. Braun, J. Chem. Phys. **116**, 6826 (2002).

[12] W.A. Eaton and J. Hofrichter, Adv Protein Chem **40**, 63 (1990).

[13] A. Coniglio, J. Phys.: Condens. Matter **13**, 9039 (2001).

[14] J. Bergenholtz, M. Fuchs, and T. Voigtmann, J. Phys.: Condens. Matter **12**, 6575 (2000).

[15] We neglect diploidy, linkage effects, malarial resistance etc.

[16] F.N. Braun and C. Viney, Int. J. Biol. Macromol. **32**, 59 (2003).

[17] Related experimental observations concerning the relation between sequence hydrophobicity and the physical chemistry of silk glands have been presented by H.J. Jin and D.L. Kaplan, Nature (London) **424**, 1057 (2003).

[18] F. Vollrath and D.P. Knight, Nature (London) **410**, 541 (2001).

[19] B. Vallone, A.E. Miele, P. Vecchini, E. Chaincone, and M. Brunori, Proc. Natl. Acad. Sci. U.S.A. **95**, 6103 (1998).

[20] K.A. Dill, Biochemistry **24**, 1501 (1985).

[21] R.J. Baxter, J. Chem. Phys. **49**, 2770 (1968)

[22] C. Regnaut and J.C. Ravey, J. Chem. Phys. **91**, 1211 (1989).

[23] Y.C. Chiew and E.D. Glandt, J. Phys. A **16**, 2599 (1983).

[24] L. Onsager, Ann. N.Y. Acad. Sci. **51**, 627 (1949).

[25] In constructing the secondary plateau of Fig. 4, we subsume the nematic-isotropic biphasic region for convenience into the liquid-dispersed phase.

[26] K.A. Dill, D.O.V. Alonso, and K. Hutchinson, Biochemistry **28**, 5439 (1989).

[27] F. Jacob and J. Monod, J. Mol. Biol. **3**, 318 (1961).

[28] G.A. Wray *et al.*, Mol. Biol. Evol. **20**, 1377 (2003).

[29] Physiological concentrations lie in the range $c \sim$ 10–30 g/dl [9]. The corresponding volume fraction can be estimated according to $\varphi \sim Na^3 \times cN_A/M$, substituting $M \sim 64$kDa for the molecular weight of the tetramer (N_A is Avogadro's number).

[30] The exact position of the HbS substitution in the surface of the native tertiary structure is a classic result of x-ray crystallography. That this presents a rather well-defined hydrophobic sticky patch is of course captured only very crudely by the low-resolution of the present approach.

[31] For an overview of the distinction between negative and positive selection, see D.A. Liberles and M. L Wayne, Genome Biology **3**, 1018 (2002).

[32] X. Gu, Mol. Biol. Evol. **16**, 1664 (1999).

[33] The number of hydrophobic residues in the tetramer n is inferred by counting over the separately archived α and β subunits, then doubling. Our dataset comprises 20 hemoglobin-sequenced primates listed by G.J.P. Naylor and M. Gerstein, J. Mol. Evol. **51**, 223 (2000). Ruffed lemur, brown lemur, galago, slender loris, slow loris, tarsier, spider monkey, capuchin, marmoset, tamarin, gorilla, human, colobus, langur, green monkey, macaque, mandril, mangabey, yellow baboon, gelada.

[34] K. Kerkam, C. Viney, D. Kaplan, and S. Lombardi, Nature (London) **349**, 596 (1991).

[35] *Nephila* dragline silk comprises two fibroins, spidroin I and spidroin II. According to the theory of Ref. [16], liquid crystalline mesogens in the gland are assembled exclusively from spidroin II. The Swiss-prot data presented in Fig. 5 is extrapolated from available spidroin II sequence fragments, *N.clavipes*, *N.madagascariensis*, and *N.senegalensis*. See J. Gatesy, C. Hayashi, D. Motriuk, J. Woods, and R. Lewis, Science **291**, 2603 (2001).

[36] Assuming 1 generation $\sim$1 yr, and a per-site HP flipping rate of 1 pauling, i.e., 10^{-9} site $^{-1}$y^{-1} [37], we estimate $\theta \sim 10^{-6}$ for $N \sim 3500$ fibroins. Hence, positive selection for liquid crystallinity in the *Nephila* silk gland confers a fitness advantage at least of the order of $\delta w \sim 10^{-6}$.

[37] M. Kimura, *The Neutral Theory of Molecular Evolution* (Cambridge University Press, Cambridge, 1983).

[38] D.M. Taverna and R.A. Goldstein, Proteins **46**, 105 (2002).

[39] S.A. Benner, G. Cannarozzi, D. Gerloff D, M. Turcotte, and G. Chelvanayagam, Chem. Rev. (Washington, D.C.) **97**, 2725 (1997).

[40] A. Irbaeck and E. Sandelin, Biophys. J. **79**, 2252 (2000); E. Sandelin, Biophys. J. (to be published).

[41] S. Ohno, *Evolution by Gene Duplication* (Springer, Berlin, 1970).

[42] R. A. Raff, *The Shape of Life* (University of Chicago Press, Chicago, 1996).

[43] J. D. Bryngelson and E. M. Billings, in *Physics of Biological Systems*, edited by H. Flybjerg *et al.* (Springer, Berlin, 1997).