Version 2

Protein folding disorders: Toward a basic biological paradigm

Rodrick Wallace, Ph.D.
Division of Epidemiology
The New York State Psychiatric Institute*

April 9, 2010

Abstract

Mechanistic 'physics' models of protein folding fail to account for the observed spectrum of protein folding and aggregation disorders, suggesting that a more appropriately biological paradigm will be needed for understanding the etiology, prevention, and treatment of these diseases. Here, a spontaneous symmetry breaking argument is first applied to the problem, via a rate distortion analysis of the relation between genome coding and the final condensation of the protein molten globule analogous to Tlusty's (2007) exploration of the evolution of the genetic code. In the 'energy' picture, the average distortion between codon message and final protein structure, under constraints driven by evolutionary selection, serves as a temperature analog, so that low values limit the possible distribution of protein forms, producing the canonical folding funnel. A dual 'developmental' perspective sees the rate distortion function itself as the temperature analog, and permits incorporation of chaperones or external factors as catalysts, driving the system to different possible outcomes or affecting the rate of convergence. The developmental formalism is then generalized to a more biologically relevant cognitive paradigm describing the interaction of protein folding with both local cellular machinery and embedding epigenetic and environmental signals. A nonequilibrium empirical Onsager treatment provides an adaptable statistical model for protein folding, in the same manner as a regression equation. This produces quasi-equilibrium 'resilience' states representing normal, corrected, eliminated, and pathological states of protein folding. A straightforward eneralization to long time scales produces diffusion models for the onset of protein folding disorders in which epigenetic or life history factors determine the diffusion coefficient or affect the efficiency of chaperone processes.

Key Words: aging, cognitive paradigm, development, ecological resilience, groupoid, information theory, phase transition, protein folding disorders, rate distortion, spontaneous symmetry breaking

*Affiliation for identification only. Correspondence: Rodrick Wallace, 549 W. 123 St., Apt. 16F, New York, NY, 10027 USA, wallace@pi.cpmc.columbia.edu, rodrick.wallace@gmail.com.

1 Introduction

1.1 Protein folding disorders

The existence of 'global' protein folding and aggregation diseases, in conjunction with the elaborate cellular folding regulatory apparatus associated with the endoplasmic reticulum and other structures (e.g., Scheuner and Kaufman, 2008; Dobson, 2003), makes clear that simple physical 'folding funnel' free energy mechanisms are not fully adequate to describe the process, to understate the matter. This suggests that a more biologically-based model is needed, analogous to Atlan and Cohen's (1998) cognitive paradigm for the immune system. That is, the intractable set of disorders related to protein aggregation and misfolding belies simple mechanistic approaches, although free energy landscape pictures surely capture part of the process. The diseases range from prion illnesses like Creutzfeld-Jakob disease, to amyloid-related dysfunctions like Alzheimer's, Huntington's and Parkinson's diseases, and type 2 diabetes. Misfolding disorders include emphysema and cystic fibrosis. A deeper understanding of protein folding mechanisms, in particular of epigenetic, social, and environmental influences, might contribute to prevention and treatment of these debilitating conditions.

More particularly, the role of epigenetic and environmental factors in type 2 diabetes has long been known (e.g., Zhang et al., 2009; Wallach and Rey, 2009). Haataja et al. (2008), for example, conclude that the islet in type 2 diabetes shows much in common with neuropathology in neurodegenerative diseases where interest is now focused on protein misfolding and aggregation and the diseases are now often referred to as unfolded protein diseases.

Scheuner and Kaufman (2008) likewise examine the unfolded protein response in β cell failure and diabetes. Indeed, their opening paragraph raises the fundamental questions regarding the adequacy of simple energy landscape models of protein folding:

> In eukaryotic cells, protein synthesis and secretion are precisely coupled with the capacity of the endoplasmic reticulum (ER) to fold, process, and traffic proteins to the cell surface. These processes are coupled through several signal transduction pathways collectively known as the unfolded protein

response [that] functions to reduce the amount of nascent protein that enters the ER lumen, to increase the ER capacity to fold protein through transcriptional up-regulation of ER chaperones and folding catalysts, and to induce degradation of misfolded and aggregated protein.

Qiu et al. (2009) address Alzheimer's disease in much the same fashion:

> Alzheimer's dementia is a multifactorial disease in which older age is the strongest risk factor... [that] may partially reflect the cumulative effects of different risk and protective factors over the lifespan, including the complex interactions of genetic susceptibility, psychosocial factors, biological factors, and environmental exposures experienced over the lifespan.

Qiu et al. (2009) explain that mutation effects account for only a small fraction of observed cases, and that the APOE $\epsilon 4$ allele – the only established genetic factor for both early and late onset disease – is a *susceptibility* gene, neither necessary nor sufficient for disease onset. They further describe how many of the same factors implicated in diabetes and cardiovascular disease predict onset of Alzheimer's as well: tobacco use, high blood pressure, high serum cholesterol, chronic inflammation, as indexed by a higher level of serum C-reactive protein, and diabetes itself. Highly significant protective factors include high educational and socioeconomic status, regular physical exercise, mentally demanding activities, and significant social engagement.

Similarly, Fillit et al. (2008) find that lifestyle risk factors for cardiovascular disease, such as obesity, lack of exercise, smoking, and certain psychosocial factors, have been associated with an increased risk for cognitive decline and dementia, concluding that current evidence indicates an association between hypertension, dyslipidemia and diabetes and cognitive decline and dementia.

Goldschmidt et al. (2010) describe pathological protein fibrillation as follows:

> We found that [protein segments with high fibrillation propensity] tend to be buried or twisted into unfavorable conformations for forming beta sheets... For some proteins a delicate balance between protein folding and misfolding exists that can be tipped by changes in environment, destabilizing mutations, or even protein concentration...
>
> In addition to the self-chaperoning effects described above, proteins are also protected from fibrillation during the process of folding by molecular chaperones...
>
> Our genome-wide analysis revealed that self-complementary segments are found in almost all proteins, yet not all proteins are amyloids. The implication is that chaperoning effects have evolved to constrain self-complementary segments from interaction with each other.

These processes and mechanisms seem no less examples of chemical cognition than the immune/inflammatory responses that Atlan and Cohen (1998) describe in terms of an explicit cognitive paradigm, or that characterizes well-studied neural processes. Our own work (Wallace and Wallace, 2008, 2009) introduces a similar, and highly formal, cognitive paradigm for gene expression whose machinery permits the natural incorporation of epigenetic and environmental signals via catalytic mechanisms similar to those of Section 5.3 below. The implication is that progress in understanding, preventing, and treating protein folding and aggregation disorders now requires introduction of a biologically-based cognitive paradigm for the folding process itself.

1.2 The 'standard model' of protein folding

The symmetries and dynamics of protein folding are striking and, in a local sense, fairly well understood (Dill et al. 2007; Wolynes, 1996; Onuchic and Wolynes, 2004). Figure 1, from Goodsell and Olson (2000), shows several typical examples. More general, but less overtly 'symmetric', conformations, however, involve finite tilings of helices, sheets, and attachment loops that would seem better described using groupoid methods, following the arguments of Weinstein (1996): As Wolynes (1996) put the matter, "It is the inexact symmetries of biological molecules that are most striking".

Anfinsen's (1973) thermodynamic hypothesis has strongly dominated thinking on the subject: the native state of a protein has the lowest Gibbs free energy, determined by the interaction of the amino acid sequence and the embedding environment (Wolynes, 1996), with hydrophobic amino acids driven into the center of the 'native' folded protein structure. More recent work (e.g., summarized in Lei and Huang, 2010) suggests that large, complex proteins may have native configurations representing kinetically accessible, rather than thermodynamically minimal, states. Andre et al. (2008) explore the central insight that "...selection is only likely to operate on primordial complexes with sufficient initial interaction energy to at least partially overcome the entropic costs of association of the monomers; evolution can only optimize a complex that is populated sufficiently to confer a benefit on the organism".

Here we will attempt to finesse this general perspective by invoking a rate distortion argument applied to the transmitted signal represented by the translation of the genome into the final, evolutionarily driven, condensation of the molten globule of the resulting amino acid string. The argument, an adaptation of Tlusty's (2007) insights regarding the role of rate distortion constraints in evolutionary process, seems fairly direct. It is based on standard material from statistical physics and information theory, using, respectively, average distortion and the rate distortion function itself, as temperature analogs to produce mirror image 'energy' and 'development' pictures of protein folding.

The final step is to mathematically 'weaken', i.e., generalize, the development perspective, using information sources formally dual to the several chemical cognitive processes involved in protein folding. These then, in a sense, engage in a

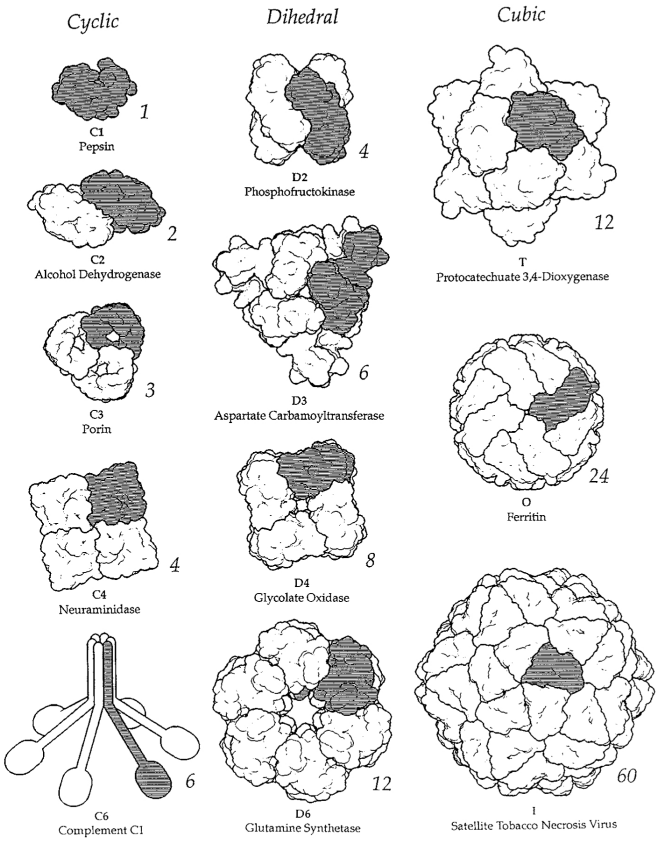

Figure 1: From Godsell and Olson, (2000). Proteins with each of the crystallographic point group symmetries have been found. Point group symbols are included below each protein structure (e.g., C1 and D2), and the number of identical subunits in each group is included below and to the right of the structure. One subunit is shaded in each example. Other, noncrystallographic, point groups are consistent with the entantiomorphic nature of proteins, including complicated cyclic and dihedral symmetries. The more ubiquitous tilings of helices, sheets, and loops can probably be best described using groupoids, most directly seen as disjoint unions of groups, for which a 'product' is only locally defined (e.g., Weinstein, 1996).

local, multifactorial, coevolutionary interaction whose quasi-stable dynamic states generate products that are, respectively, correct, repaired, eliminated, or misfolded/aggregated proteins. This set of processes is analogous to quasi-stable ecosystem resilience modes, in the sense of Holling (1973) or Gunderson (2000), and apparently subject to punctuated transitions between them consequent on epigenetic or environmental perturbations.

The argument generates a new class of statistical models based on the asymptotic limit theorems of information, in the same sense that regression and other parametric models are based on the Central Limit Theorem, and these should prove useful in data analysis as well as providing a new conceptual approach.

We begin with a restatement of some standard material from statistical physics that provides the basis for a subsequent argument-by-abduction.

2 Spontaneous symmetry breaking

Landau's theory of phase transitions (Landau and Lifshitz, 2007) assumes that the free energy of a system near criticality can be expanded in a power series of some 'order parameter' ϕ representing a fundamental measurable quantity, that is, a symmetry invariant. One writes

$$F_0 = \sum_{k=m}^{p(>m)} A_k \phi^k, \tag{1}$$

with $A_2 \approx \alpha(T - T_c)$ sufficiently close to the critical temperature T_c. This mean field approach can be used to describe a variety of second-order effects for $p = 4$ or $p = 6$, $A_3 = 0$ and $A_4 > 0$, and first order phase transitions (requiring latent heat) for either $p = 6, A_3 = 0, A_4 < 0$ or $p = 4$ and $A_3 \neq 0$. These can be both temperature induced (for $m = 2$) and field induced (for $m = 1$).

Minimization of F_0 with respect to the order parameter yields the average value of ϕ, $< \phi >$, which is zero above the critical temperature and non-zero below it. In the absence of external fields, the second-order transition occurs at $T = T_c$, while the first-order, needing latent heat, occurs at $T_c^* = T_c + A_4^2/4\alpha A_6$. In the latter case thermal hysteresis arises between $T_s \equiv T_c + A_4^2/3\alpha A_6$ and T_c. A more accurate approximation involves an expression that recognizes the effect of coarse-graining, adding a term in $\nabla^2\phi$ and integrating over space rather than summing. Regimes dominated by this gradient will show behaviors analogous to those described using the one dimensional Landau-Ginzburg equation, which, among other things, characterizes superconductivity.

The Landau formalism quickly enters deep topological waters (Pettini, 2007, pp. 42-43; Landau and Lifshitz, 2007, pp. 459-466). The essence of Landau's insight was that phase transitions without latent heat – second order transitions – were usually in the context of a significant symmetry change

in the physical states of a system, with one phase, at higher temperature, being far more symmetric than the other. A symmetry is lost in the transition, a phenomenon called spontaneous symmetry breaking. The greatest possible set of symmetries in a physical system is that of the Hamiltonian describing its energy states. Usually states accessible at lower temperatures will lack symmetries available at higher temperatures, so that the lower temperature phase is the less symmetric: The randomization of higher temperatures ensures that higher symmetry/energy states will then be accessible to the system.

At the lower temperature an order parameter must be introduced to describe the system's physical states – some extensive quantity like magnetization. The order parameter will vanish at higher temperatures, involving more symmetric states, and will be different from zero in the less symmetric lower temperature phase.

This can be formalized, following Pettini (2007), as follows. Consider a thermodynamic system having a free energy F which is a function of temperature T, pressure P, and some other extensive macroscopic parameters m_i, so that $F = F(P, T, m_i)$. The m_i all vanish in the most symmetric phase, so that, as a function of the m_i, $F(P, T, m_i)$ is invariant with respect to the transformations of the symmetry group G_0 of the most symmetric phase of the system when all $m_i \equiv 0$.

The state of the system can be represented by a vector $|m >= |m_1, ..., m_n >$ in a vector space $\mathcal{E}$. Now, within $\mathcal{E}$, construct a linear representation of the group G_0 that associates with any $g \in G_0$ a matrix $M(g)$ having rank n. In general, the representation $M(g)$ is reducible, and we can decompose $\mathcal{E}$ into invariant irreducible subspaces $\mathcal{E}_1, \mathcal{E}_2, ..., \mathcal{E}_k$, having basis vectors $|e_i^{(n)} >$ with $n = 1, 2, ...n_i$ and $n_i = dim\mathcal{E}_i$. The state variables m_i are transformed into new variables $\eta_i^{(n)} =< e_i^{(n)}|m >$, where the bracket represents an inner product.

In terms of irreducible representations $D_i(g)$ induced by $M(g)$ in $\mathcal{E}_i$ we have

$$M(g) = D_1(g) \oplus D_2(g) \oplus, ..., \oplus D_k(g).$$

If at least one of the $\eta_i^{(n)}$ is nonzero, then the system no longer has the symmetry G_0. This symmetry has been broken, and the new symmetry group is G_i, associated with the representation $D_i(g)$ in $\mathcal{E}_i$. The variables $\eta_i^{(n)}$ are the new order parameters, and the free energy is now $F = F(P, T, \eta_i^{(n)})$. For a physical system the actual values of the η as functions of P and T can be variationally determined by minimizing the free energy F.

Two essential features distinguish information systems, like the translation of a genome into a folded protein, from this simple physical model.

First, the dynamics of order parameters cannot always be determined by simplistic minimization procedures in biological circumstances (e.g., Levinthal, 1969): embedding environments can, within contextual constraints (that particularly include available metabolic free energy), write images of themselves via evolutionary selection mechanisms, driving the system toward such structures as the protein folding funnel (e.g., Levinthal, 1968; Wolynes, 1996).

Second, the essential symmetry of information sources is quite often driven by groupoid, rather than group, structures (e.g., Wallace, 2010). One must then engage the full transitive orbit/isotropy group decomposition, and examine groupoid representations (e.g., Bos, 2007; Buneci, 2003) configured about the irreducible representations of the isotropy groups. This observation seems particularly relevant given the usual helix/sheet/connecting loop tilings that characterize most elaborate protein conformations (Wolynes, 1996).

A brief summary of standard material on groupoids is included as a Mathematical Appendix.

3 A little information theory

Here we think of the machinery listing a sequence of codons as communicating with machinery that produces amino acids, folds them *in context*, and produces the final symmetric protein. We then suppose it possible to compare what is actually produced with what should have been produced, perhaps by a simple evolutionary survival mechanism, perhaps via some more sophisticated error-correcting systems. This is not a new idea, and Onuchic and Wolynes (2004), for example, put the matter fully in evolutionary terms:

> Protein folding should be complex... a folding mechanism must involve a complex network of elementary interactions. However, simple empirical patterns of protein folding kinetics... have been shown to exist.
>
> This simplicity is owed to the global organization of the landscape of the energies of protein conformations into a funnel...This organization is not characteristic of all polymers with any sequence of amino acids, but is a result of evolution...
>
> Evolution achieves robustness by selecting for sequences in which the interactions present in the functionally useful structure are not in conflict, as in a random heteropolymer, but instead are mutually supportive and cooperatively lead to a low energy structure. The interactions are 'minimally frustrated'... or 'consistent'...

It is possible to reframe something of this mechanism in formal information theory terms.

Suppose a sequence of signals is generated by a biological information source Y having output $y^n = y_1, y_2, ...$ – codons. This is 'digitized' in terms of the observed behavior of the system with which it communicates, say a sequence of 'observed behaviors' $b^n = b_1, b_2, ...$ – amino acids and their folded protein structure. Assume each b^n is then deterministically retranslated back into a reproduction of the original biological signal, $b^n \to \hat{y}^n = \hat{y}_1, \hat{y}_2, ...$.

Define a distortion measure $d(y, \hat{y})$ which compares the original to the retranslated path. Many distortion measures are possible. The Hamming distortion is defined simply as

$$d(y, \hat{y}) = 1, y \neq \hat{y}$$

$$d(y, \hat{y}) = 0, y = \hat{y}.$$

For continuous variates the squared error distortion is just $d(y, \hat{y}) = (y - \hat{y})^2$.

There are many such possibilities. The distortion between *paths* y^n and $\hat{y}^n$ is defined as $d(y^n, \hat{y}^n) \equiv \frac{1}{n} \sum_{j=1}^{n} d(y_j, \hat{y}_j)$.

A remarkable fact of the Rate Distortion Theorem is that *the basic result is independent of the exact distortion measure chosen* (Cover and Thomas, 1991; Dembo and Zeitouni, 1998).

Suppose that with each path y^n and b^n-path retranslation into the y-language, denoted $\hat{y}^n$, there are associated individual, joint, and conditional probability distributions $p(y^n), p(\hat{y}^n), p(y^n, \hat{y}^n), p(y^n|\hat{y}^n)$.

The average distortion is defined as

$$D \equiv \sum_{y^n} p(y^n) d(y^n, \hat{y}^n). \quad (2)$$

It is possible, using the distributions given above, to define the information transmitted from the Y to the $\hat{Y}$ process using the Shannon source uncertainty of the strings:

$$I(Y, \hat{Y}) \equiv H(Y) - H(Y|\hat{Y}) = H(Y) + H(\hat{Y}) - H(Y, \hat{Y}),$$

where $H(..., ...)$ is the standard joint, and $H(...|...)$ the conditional, Shannon uncertainties (Cover and Thomas, 1991; Ash, 1990).

If there is no uncertainty in Y given the retranslation $\hat{Y}$, then no information is lost, and the systems are in perfect synchrony.

In general, of course, this will not be true.

The *rate distortion function* $R(D)$ for a source Y with a distortion measure $d(y, \hat{y})$ is defined as

$$R(D) = \min_{p(y,\hat{y}); \sum_{(y,\hat{y})} p(y) p(y|\hat{y}) d(y,\hat{y}) \leq D} I(Y, \hat{Y}). \quad (3)$$

The minimization is over all conditional distributions $p(y|\hat{y})$ for which the joint distribution $p(y, \hat{y}) = p(y)p(y|\hat{y})$ satisfies the average distortion constraint (i.e., average distortion $\leq D$).

The *Rate Distortion Theorem* states that $R(D)$ is the minimum necessary rate of information transmission which ensures the communication between the biological vesicles does not exceed average distortion D. Thus $R(D)$ defines a minimum necessary channel capacity. Cover and Thomas (1991) or Dembo and Zeitouni (1998) provide details. The rate distortion function has been calculated for a number of systems.

We reiterate an absolutely central fact characterizing the rate distortion function: Cover and Thomas (1991, Lemma 13.4.1) show that $R(D)$ *is necessarily a decreasing convex function of* D *for any reasonable definition of distortion.*

That is, $R(D)$ *is always* a reverse J-shaped curve. This will prove crucial for the overall argument. Indeed, convexity is an exceedingly powerful mathematical condition, and permits deep inference (e.g., Rockafellar, 1970). Ellis (1985, Ch. VI) applies convexity theory to conventional statistical mechanics.

For a Gaussian channel having noise with zero mean and variance σ^2 (Cover and Thomas, 1991),

$$R(D) = 1/2 \log[\sigma^2/D], 0 \leq D \leq \sigma^2$$
$$R(D) = 0, D > \sigma^2. \quad (4)$$

Recall, now, the relation between information source uncertainty and channel capacity (e.g., Ash, 1990):

$$H[X] \leq C, \quad (5)$$

where H is the uncertainty of the source X and C the channel capacity, defined according to the relation (Ash, 1990)

$$C \equiv \max_{P(X)} I(X|Y), \quad (6)$$

where $P(X)$ is chosen so as to maximize the rate of information transmission along a channel Y.

Note that for a parallel set of noninteracting channels, the overall channel capacity is the sum of the individual capacities, providing a powerful 'consensus average' that does not apply in the case of modern molecular coding.

Finally, recall the analogous definition of the rate distortion function above, again an extremum over a probability distribution.

Our own work (Wallace and Wallace, 2008) focuses on the homology between information source uncertainty and free energy density. More formally, if $N(n)$ is the number of high probability 'meaningful' – that is, grammatical and syntactical – sequences of length n emitted by an information source X, then, according to the Shannon-McMillan Theorem, the zero-error limit of the Rate Distortion Theorem (Ash, 1990; Cover and Thomas, 1991; Khinchin, 1957),

$$H[X] = \lim_{n\to\infty} \frac{\log[N(n)]}{n}$$

$$= \lim_{n\to\infty} H(X_n|X_0, ..., X_{n-1})$$

$$= \lim_{n\to\infty} \frac{H(X_0, ..., X_n)}{n+1},$$

(7)

where, again, $H(...|...)$ is the conditional and $H(..., ...)$ is the joint Shannon uncertainty.

In the limit of large n, $H[X]$ becomes homologous to the free energy density of a physical system at the thermodynamic limit of infinite volume. More explicitly, the free energy density of a physical system having volume V and partition function $Z(\beta)$ derived from the system's Hamiltonian – the energy function – at inverse temperature β is (e.g., Landau and Lifshitz 2007)

$$F[K] = \lim_{V\to\infty} -\frac{1}{\beta}\frac{\log[Z(\beta, V)]}{V} \equiv$$

$$\lim_{V\to\infty} \frac{\log[\hat{Z}(\beta, V)]}{V},$$

with $\hat{Z} = Z^{-1/\beta}$. The latter expression is formally similar to the first part of equation (7), a circumstance having deep implications: Feynman (2000) describes in great detail how information and free energy have an inherent duality. Feynman, in fact, defines information precisely as the free energy needed to erase a message. The argument is surprisingly direct (e.g., Bennett, 1988), and for very simple systems it is easy to design a small (idealized) machine that turns the information within a message directly into usable work – free energy. Information is a form of free energy and the construction and transmission of information within living things consumes metabolic free energy, with nearly inevitable losses via the second law of thermodynamics. If there are limits on available metabolic free energy there will necessarily be limits on the ability of living things to process information.

Figure 2 presents a schematic of the mechanism: As the complexity of a dynamic physiological information process rises, that is, as H increases, its free energy content increases linearly. The metabolic free energy needed to construct and maintain the physiological systems that instantiate H should, however, be expected to increase nonlinearly with it, hence the 'translation gap' of the figure. Section 5 of Wallace (2010) gives a fairly elementary derivation of such a relation in terms of rate distortion theory. Figure 2 suggests that H may indeed be a good, if highly nonlinear, index of large-scale free energy dynamics.

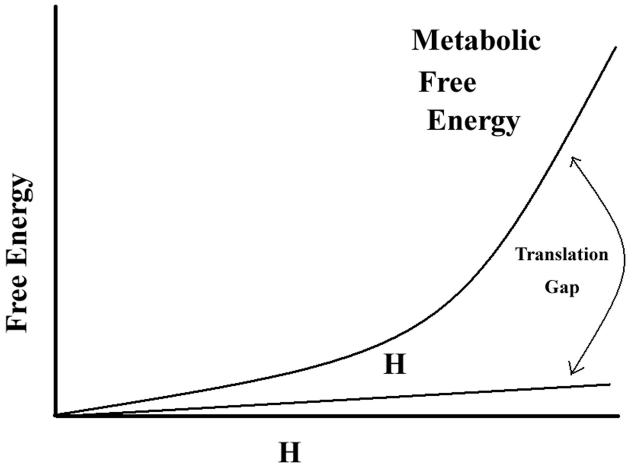

Figure 2: Nonlinear increase in metabolic free energy needed to maintain and generate linear increase in the information source uncertainty of a complex physiological process. H is seen to 'leverage' metabolic expenditures, parameterizing a more complicated nonequilibrium thermodynamics. See Wallace (2010) for an explicit calculation in a somewhat different system.

Conversely, information source uncertainty has an important heuristic interpretation that Ash (1990) describes as follows:

> [W]e may regard a portion of text in a particular language as being produced by an information source. The probabilities $P[X_n = a_n|X_0 = a_0, ...X_{n-1} = a_{n-1}]$ may be estimated from the available data about the language; in this way we can estimate the uncertainty associated with the language. A large uncertainty means, by the [Shannon-McMillan Theorem], a large number of 'meaningful' sequences. Thus given two languages with uncertainties H_1 and H_2 respectively, if $H_1 > H_2$, then in the absence of noise it is easier to communicate in the first language; more can be said in the same amount of time. On the other hand, it will be easier to reconstruct a scrambled portion of text in the second language, since fewer of the possible sequences of length n are meaningful.

In sum, if a biological system characterized by H_1 has a richer and more complicated internal communication structure than one characterized by H_2, then necessarily $H_1 > H_2$ and system 1 represents a more energetic process than system 2, and by the arguments of figure 2, may trigger even greater metabolic free energy dynamics.

By equations (5), (6), and (7), the Rate Distortion Function, $R(D)$ is likewise a free energy measure, constrained by the availability of metabolic free energy.

4 The energy picture

Ash's comment leads directly to a model in which the average distortion between the initial codon stream and the final form of the folded amino acid stream, the protein, becomes a dominant force, particularly in an evolutionary context in which fidelity of codon expression has survival value. The direct model examines the distortion between the codon stream and the folded protein structure.

Suppose there are n possible folding schemes. The most familiar approach, perhaps, is to assume that a given distortion measure, D, under evolutionary selection constraints, serves much as an external temperature bath for the possible distribution of conformation free energies, the set $\{\mathcal{H}_1, ..., \mathcal{H}_n\}$. That is, high distortion, represented by a low rate of transmission of information between codon machine and amino acid/protein folding machine, permits a larger distribution of possible symmetries – the big end of the folding funnel – according to the classic formula

$$Pr[\mathcal{H}_j] = \frac{\exp[-\mathcal{H}_j/\lambda D]}{\sum_{i=1}^{n} \exp[-\mathcal{H}_i/\lambda D]}, \tag{8}$$

where $Pr[\mathcal{H}_j]$ is the probability of folding scheme j having conformational free energy $\mathcal{H}_j$.

We are, in essence, assuming that $Pr[\mathcal{H}_j]$ is a one parameter distribution in the 'intensive' quantity D.

The free energy Morse Function associated with this probability is

$$F_R = -\lambda D \log[\sum_{i=1}^{n} \exp[-\mathcal{H}_i/\lambda D]]. \tag{9}$$

Applying a spontaneous symmetry breaking argument to F_R generates topological transitions in folded protein structure as the 'temperature' D decreases, i.e., as the average distortion declines. That is, as the channel capacity connecting codon machines with amino acid/protein folding machines increases, the system is driven to a particular conformation, according to the 'protein folding funnel'.

5 The developmental picture

The developmental approach of Wallace and Wallace (2009) permits a different perspective on protein folding.

We now are concerned with developmental pathways in a 'phenotype space' that, in a series of steps, take the amino acid string $\mathbf{S}_0$ at time 0 to the final folded conformation $\mathbf{S}_f$ at some time t in a long series of distinct, sequential, intermediate configurations $\mathbf{S}_i$.

Let $N(n)$ be the number of possible paths of length n that lead from $\mathbf{S}_0$ to $\mathbf{S}_f$. The essential assumptions are:

[1] This is a highly systematic process governed by a 'grammar' and 'syntax' driven by the folding funnel, so that it is possible to divide all possible paths $x_n = \{\mathbf{S}_0, \mathbf{S}_1, ..., \mathbf{S}_n\}$ into two sets, a small, high probability subset that conforms to the demands of the folding funnel topology, and a much larger 'nonsense' subset having vanishingly small probability.

[2] If $N(n)$ is the number of high probability paths of length n, then the 'ergodic' limit

$$H = \lim_{n\to\infty} \log[N(n)]/n \tag{10}$$

both exists and is independent of the path x. This is, essentially, a restatement of the Shannon-McMillan Theorem (Khinchin, 1957).

That is, the folding of a particular protein, from its amino acid string to its final form, is not a random event, but represents a highly – evolutionarily – structured (i.e., by the folding funnel) 'statement' by an information source having source uncertainty H.

5.1 Symmetry arguments

A formal equivalence class algebra can now be constructed by choosing different origin and end points $\mathbf{S}_0, \mathbf{S}_f$ and defining equivalence of two states by the existence of a high probability meaningful path connecting them with the same origin and end. Disjoint partition by equivalence class, analogous to orbit equivalence classes for dynamical systems, defines the vertices of the proposed network of developmental protein 'languages'. We thus envision a *network of metanetworks*. Each vertex then represents a different equivalence class of developmental information sources. This is an abstract set of metanetwork 'languages'.

This structure generates a groupoid, in the sense of Weinstein (1996). States a_j, a_k in a set A are related by the groupoid morphism if and only if there exists a high probability grammatical path connecting them to the same base and end points, and tuning across the various possible ways in which that can happen – the different developmental languages – parameterizes the set of equivalence relations and creates the (very large) groupoid.

There is an implicit hierarchy. First, there is structure *within the system having the same base and end points*. Second, there is a complicated groupoid structure defined by sets of dual information sources surrounding the variation of base and end points. We do not need to know what that structure is in any detail, but can show that its existence has profound implications.

We begin with the simple case, the set of dual information sources associated with a fixed pair of beginning and end states.

5.1.1 The first level

Taking the serial grammar/syntax model above, we find that not all high probability meaningful paths from $\mathbf{S}_0$ to $\mathbf{S}_f$ are actually the same. They are structured by the uncertainty of the associated dual information source, and that has a homological relation with free energy density.

Let us index possible information sources connecting base and end points by some set $A = \cup\alpha$. Argument by abduction from statistical physics is direct. The minimum channel capacity needed to produce average distortion less than D in the energy picture above is $R(D)$. We take the probability of a particular H_α as determined by the standard expression

$$P[H_\beta] = \frac{\exp[-H_\beta/\mu R]}{\sum_\alpha \exp[-H_\alpha/\mu R]}, \tag{11}$$

where the sum may, in fact, be a complicated abstract integral.

A basic requirement, then, is that the sum/integral always converges.

Thus, in this formulation, there must be structure *within* a (cross sectional) connected component in the base configuration space, determined by R. Some dual information sources will be 'richer'/smarter than others, but, conversely, must use more available channel capacity for their completion.

5.1.2 The second level

While we might simply impose an equivalence class structure based on equal levels of energy/source uncertainty, producing a groupoid (and possibly allowing a Morse Theory approach), we can do more *by now allowing both source and end points to vary*, as well as by imposing energy-level equivalence. This produces a far more highly structured groupoid.

Equivalence classes define groupoids, by standard mechanisms. The basic equivalence classes – here involving both information source uncertainty level and the variation of $\mathbf{S}_0$ and $\mathbf{S}_f$, will define transitive groupoids, and higher order systems can be constructed by the union of transitive groupoids, having larger alphabets that allow more complicated statements in the sense of Ash above.

Again, given a minimum necessary channel capacity R, we propose that the metabolic-energy-constrained probability of an information source representing equivalence class G_i, H_{G_i}, will again be given by

$$P[H_{G_i}] = \frac{\exp[-H_{G_i}/\kappa R]}{\sum_j \exp[-H_{G_j}/\kappa R]}, \tag{12}$$

where the sum/integral is over all possible elements of the largest available symmetry groupoid. By the arguments of Ash above, compound sources, formed by the union of underlying transitive groupoids, being more complex, generally having richer alphabets, as it were, will all have higher free-energy-density-equivalents than those of the base (transitive) groupoids.

Let

$$Z_G = \sum_j \exp[-H_{G_j}/\kappa R]. \tag{13}$$

We now define the *Groupoid free energy* of the system, a Morse Function F_G, at channel capacity R, as

$$F_G[R] = -\tfrac{1}{\kappa R}\log[Z_G[R]]. \tag{14}$$

These free energy constructs permit introduction of the spontaneous symmetry breaking arguments above, but now an *increase* in R (with corresponding decrease in average distortion D) permits richer system dynamics – higher source uncertainty – resulting in more rapid transmission of the 'message' constituting convergence from $\mathbf{S}_0$ to $\mathbf{S}_f$.

5.2 Folding speed and mechanism

Dill et al. (2007) describe the conundrum of folding speeds as follows:

> ...[P]rotein folding speeds – now known to vary over more than eight orders of magnitude – correlate with the topology of the native protein: fast folders usually have mostly local structure, such as helices and tight turns, whereas slow folders usually have more non-local structure, such as β sheets (Plaxco et al., 1998)...

A simple groupoid probability argument reproduces this result. Assume that protein structure can be characterized by some groupoid representing, at least, the disjoint union of the groups describing the symmetries of component secondary structures – e.g., helices and sheets. Then, in equation 11, the set $A = \cup\alpha$ grows in size – cardinality – with increasing structural complexity. If channel capacity is capped by some mechanism, so that (at least) R grows at a lesser rate than A, by some measure, then

$$P[H_\beta] = \frac{\exp[-H_\beta/\mu R]}{\sum_\alpha \exp[-H_\alpha/\mu R]}$$

must decrease with increase in the number of possible states α, i.e., with increase in the cardinality of R, producing progressively lower rates of convergence to the final state.

These matters lead to the next central question: How can folding rates be modulated?

5.3 Catalysis of protein folding

Incorporating the influence of embedding contexts – epigenetic or chaperone effects, or the effects of (broadly) toxic exposures – can be done here by invoking the Joint Asymptotic Equipartition Theorem (JAEPT)(Cover and Thomas, 1991). For example, given an embedding contextual information source, say Z, that affects protein development, then the developmental source uncertainty H_{G_i} is replaced by a joint uncertainty $H(X_{G_i}, Z)$. The objects of interest then become the jointly typical dual sequences $y^n = (x^n, z^n)$, where x is associated with protein folding development and z with the embedding context. Restricting consideration of x and z to those sequences that are in fact jointly typical allows use of the information transmitted from Z to X as the splitting criterion.

One important inference is that, from the information theory 'chain rule' (Cover and Thomas, 1991), $H(X,Y) = H(X) + H(Y|X) \leq H(X) + H(Y)$, while there are approximately $\exp[nH(X)]$ typical X sequences, and $\exp[nH(Z)]$ typical Z sequences, and hence $\exp[n(H(x)+H(Y))]$ independent joint sequences, there are only about $\exp[nH(X,Z)] \leq \exp[n(H(X) + H(Y))]$ jointly typical sequences, so that the effect of the embedding context, in this model, is to lower the *relative* free energy of a particular protein channel.

Thus the effect of epigenetic/catalytic regulation or toxic exposure is to channel protein into pathways that might otherwise be inhibited or slowed by an energy barrier. Hence the epigenetic/catalytic/toxic information source Z acts as a *tunable catalyst*, a kind of second order enzyme, to enable and direct developmental pathways. This result permits hierarchical models similar to those of higher order cognitive neural function (e.g, Wallace, 2005).

This is indeed a relative energy argument, since, metabolically, two systems must now be supported, i.e., that of the 'reaction' itself and that of its catalytic regulator. 'Programming' and stabilizing inevitably intertwined, as it were.

Protein folding, in the developmental picture, can be visualized as a series of branching pathways. Each branch point is a developmental decision, or switch point, governed by some regulatory apparatus (if only the slope of the folding funnel) that may include the effects of toxins or epigenetic mechanisms.

A more general picture emerges by allowing a distribution of possible 'final' states $\mathbf{S}_f$. Then the groupoid arguments merely expand to permit traverse of both initial states and possible final sets, recognizing that there can now be a possible overlap in the latter, and the catalytic effects are realized through the joint uncertainties $H(X_{G_i}, Z)$, so that the guiding information source Z serves to direct as well the possible final states of X_{G_i}.

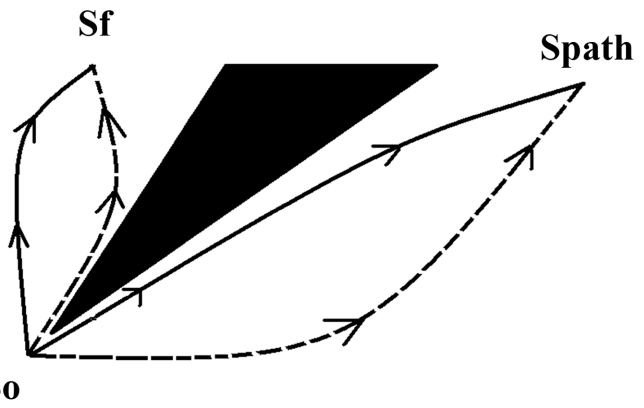

Figure 3: Given an initial state $\mathbf{S}_0$ and a critical period casting a path-dependent developmental shadow, there are two different directed homotopy equivalence classes of deformable paths leading, respectively, to the normal folded protein state $\mathbf{S}_f$ and the pathological state $\mathbf{S}_{path}$. These sets of paths form equivalence classes defining a topological groupoid.

5.4 Extending the model

The most natural extension of the developmental model of protein folding would be in terms of the directed homotopy classification of ontological trajectories, in the sense of Wallace and Wallace (2008, 2009). That is, developmental trajectories themselves can be classified into equivalence classes, for example those that lead to a normal final state $\mathbf{S}_f$, and those that lead to pathological aggregations or misfoldings, say some set $\{\mathbf{S}^i_{path}\}, i = 1, 2, ...$ This produces a dynamic directed homotopy groupoid topology whose understanding might be useful across a broad spectrum of diseases.

Figure 3 illustrates the concept. The initial developmental state $\mathbf{S}_0$ can, in this picture, 'fall' down two different sets of developmental pathways, separated by a critical period 'shadow' preventing crossover between them. Paths within one set can be topologically transformed into each other without crossing the filled triangle, and constitute a directed homotopy equivalence classes. The lower apex of the triangle can, however, start at many possible critical period points along any path connecting $\mathbf{S}_0$ and $\mathbf{S}_f$, following the arguments of Section 12 of Wallace and Wallace (2009).

Onset of a path that converges on the conformation $\mathbf{S}_{path}$ is, according to the model, driven by a genetic, epigenetic, or environmental catalysis event, in the sense of Section 5.3. The topological equivalence classes define a groupoid on the developmental system.

6 Toward a cognitive paradigm for protein folding

We now take the developmental perspective as the foundation for generating an empirically-based statistical model – effectively a cognitive paradigm for protein folding – that incorpo-

rates the embedding contexts of epigenetic and environmental signals. Atlan and Cohen (1998), in the context of a study of the immune system, argue that the essence of cognition is the comparison of a perceived signal with an internal, learned picture of the world, and then choice of a single response from a large repertoire of possible responses. Such choice inherently involves information and information transmission since it always generates a reduction in uncertainty, as explained in Ash (1990, p. 21). Thus structures that process information are constrained by the asymptotic limit theorems of information theory, in the same sense that sums of stochastic variables are constrained by the Central Limit Theorem, allowing the construction of powerful statistical tools useful for data analysis.

More formally, a pattern of incoming input $\mathbf{S}_i$ describing the folding status of the protein – starting with the initial codon stream $\mathbf{S}_0$ of equation Section 5 – is mixed in a systematic algorithmic manner with a pattern of otherwise unspecified 'ongoing activity', including cellular, epigenetic and environmental signals, $\mathbf{W}_i$, to create a path of combined signals $x = (a_0, a_1, ..., a_n, ...)$. Each a_k thus represents some functional composition of internal and external factors, and is expressed in terms of the intermediate states as

$$\mathbf{S}_{i+1} = f([\mathbf{S}_i, \mathbf{W}_i]) = f(a_i) \tag{15}$$

for some unspecified function f. The a_i are seen to be very complicated composite objects, in this treatment that we may choose to coarse-grain so as to obtain an appropriate 'alphabet'.

In a simple spinglass-like model, $\mathbf{S}$ would be a vector, $\mathbf{W}$ a matrix, and f would be a function of their product at 'time' i.

The path x is fed into a highly nonlinear decision oscillator, h, a 'sudden threshold machine' pattern recognition structure, in a sense, that generates an output $h(x)$ that is an element of one of two disjoint sets B_0 and B_1 of possible system responses. Let us define the sets B_k as

$$B_0 = \{b_0, ..., b_k\},$$

$$B_1 = \{b_{k+1}, ..., b_m\}.$$

Assume a graded response, supposing that if $h(x) \in B_0$, the pattern is not recognized, and if $h(x) \in B_1$, the pattern has been recognized, and some action $b_j, k+1 \leq j \leq m$ takes place. Typically, the set B_1 would represent the final state of the folded protein, either normal or in some pathological conformation, that is sent on in the biological process or else subjected to some attempted corrective action. Corrections may, for example, range from activation of 'heat shock' protein repair to more drastic clean-up attack.

The principal objects of formal interest are paths x triggering pattern recognition-and-response. That is, given a fixed initial state $a_0 = [\mathbf{S}_0, \mathbf{W}_0]$, examine all possible subsequent paths x beginning with a_0 and leading to the event $h(x) \in B_1$. Thus $h(a_0, ..., a_j) \in B_0$ for all $0 < j < m$, but $h(a_0, ..., a_m) \in B_1$. B_1 is thus the set of final possible states, $\mathbf{S}_f \cup \{\mathbf{S}_{path}\}$ from figure 3 that includes both the final 'physics' state $\mathbf{S}_f$ and the set of possible pathological conformations.

Again, for each positive integer n, let $N(n)$ be the number of high probability grammatical and syntactical paths of length n which begin with some particular a_0 and lead to the condition $h(x) \in B_1$. Call such paths 'meaningful', assuming, not unreasonably, that $N(n)$ will be considerably less than the number of all possible paths of length n leading from a_0 to the condition $h(x) \in B_1$.

While the combining algorithm, the form of the nonlinear oscillator, and the details of grammar and syntax, can all be unspecified in this model, the critical assumption that permits inference of the necessary conditions constrained by the asymptotic limit theorems of information theory is that the finite limit

$$H = \lim_{n \to \infty} \frac{\log[N(n)}{n} \tag{16}$$

both exists and is independent of the path x.

Call such a pattern recognition-and-response cognitive process *ergodic*. Not all cognitive processes are likely to be ergodic in this sense, implying that H, if it indeed exists at all, is path dependent, although extension to nearly ergodic processes seems possible (e.g., Wallace and Fullilove, 2007).

Invoking the spirit of the Shannon-McMillan Theorem, as choice involves an inherent reduction in uncertainty, it is then possible to define an adiabatically, piecewise stationary, ergodic (APSE) information source $\mathbf{X}$ associated with stochastic variates X_j having joint and conditional probabilities $P(a_0, ..., a_n)$ and $P(a_n | a_0, ..., a_{n-1})$ such that appropriate conditional and joint Shannon uncertainties satisfy the classic relations of equation (7).

This information source is defined as *dual* to the underlying ergodic cognitive process.

Adiabatic means that the source has been parameterized according to some scheme, and that, over a certain range, along a particular piece, as the parameters vary, the source remains as close to stationary and ergodic as needed for information theory's central theorems to apply. *Stationary* means that the system's probabilities do not change in time, and *ergodic*, roughly, that the cross sectional means approximate long-time averages. Between pieces it is necessary to invoke various kinds of phase transition formalisms, as described more fully in e.g., Wallace (2005).

Structure is now subsumed *within the sequential grammar and syntax of the dual information source* rather than within

the set of developmental paths of figure 3 and the added catalysis arguments of Section 5.3.

This transformation in perspective carries heavy computational burdens, as well as providing deeper mathematical insight, as cellular machineries, and phenomena of epigenetic or environmental catalysis, are now included within a single model.

The energy and development pictures of Sections 4 and 5 were 'dual' as simply different aspects of the convexity of the rate distortion function with average distortion. This model seems qualitatively different, as we are now invoking a 'black box' information theory statistical model involving grammar and syntax driven by an asymptotic limit theorem, the Shannon-McMillan Theorem. The set of nonequilibrium empirical generalized Onsager models derived from it, as in Wallace and Wallace (2008, 2009), is based on the information source uncertainty H as a free energy-analog (e.g., Wallace and Wallace, 2009), thus having a significantly different meaning from those above, and are more similar to regression models fitted according to the Central Limit Theorem. In a manner similar to the treatment in Wallace (2005), the system becomes subject to 'biological' renormalizations at critical, highly punctuated, transitions.

The most evident assumption at this point is that there may be more than a single cognitive protein folding process in operation, e.g., that the action of chaperones and other corrective mechanisms involve separate cognitive processes $\{H_1, ..., H_m\}$ that interact via some form of crosstalk. Following the direction of Wallace and Wallace (2009) we invoke a complicated version of an internal system of empirical Onsager relations, assuming that the different cognitive processes represented by these dual information sources *become each others primary environments*, a broadly, if locally, coevolutionary phenomenon, in the sense of Diekmann and Law (1996). We write

$$H_k = H_k(K_1, ..., K_s, ..., H_j, ...) \quad (17)$$

where the K_s represent other relevant parameters and $k \neq j$. In a generalization of the statistical model, we would expect the dynamics of such a system to be driven by an empirical recursive network of stochastic differential equations. Letting the K_s and H_j all be represented as parameters Q_j, with the caveat that H_k not depend on itself, we are able to define an entropy-analog based on the homology of information source uncertainty with free energy as

$$S_k = H_k - \sum_i Q_i \partial H_k / \partial Q_i, \quad (18)$$

whose gradients in the Q define local (broadly) chemical forces. In close analogy with other nonequilibrium phenomena we obtain a complicated recursive system of phenomenological Onsager relation stochastic differential equations:

$$dQ_t^j = \sum_i [L_{j,i}(t, ..., Q_k, ...)dt + \sigma_{j,i}(t, ..., Q_k, ...)dB_t^i] \quad (19)$$

where, again, for notational simplicity, we have expressed both parameters and information sources in terms of the same symbols Q^k. The dB_t^i represent different kinds of 'noise' having particular forms of quadratic variation that may represent a projection of environmental factors under something like a rate distortion manifold (Glazebrook and Wallace, 2009a, b).

As usual for such systems, there can be multiple quasi-stable points within a given system's $\{..., H_k..., ..., K_j, ...\}$ representing a class of generalized resilience modes (Holling, 1973; Gunderson, 2000; Wallace and Wallace, 2008) accessible via punctuation as various possible outcomes of the protein folding process: normal, repaired, eliminated, and pathological. These states can, in theory, be found by setting equation 19 to zero, as the noise terms preclude unstable equilibria. As described elsewhere (e.g., Diekmann and Law, 1996; Champagnat et al., 2006; Wallace and Wallace 2009), however, far more complicated 'coevolutionary' behaviors can be expected that we will not explore further: here we enter deep biological waters whose exploration will require a significant extension of our general formal perspective. Glazebrook and Wallace (2009a, b) provide something of a mathematical roadmap.

The essential point is that, under resilience theory, 'perturbations' of various sorts can be expected to shift the system between different quasi-stable folding modes, and once shifted, correction may be exceedingly difficult or impossible, as these are, broadly, developmental processes having significant path dependence.

7 Aging and protein folding: extending the time scale

7.1 Onsager models

The developmental perspective above, although focused on the relatively short time frames of protein metabolism – in the range from microseconds to minutes – is suggestive. The principal 'risk factor' for a large array of protein folding disorders is biological age – for humans, in the range of decades – and a simplified version of the previous section may provide a life-course perspective, that is, a developmental model over a far longer timescale.

Equations 3-7 suggest that the rate distortion function, $R(D)$, is itself a free energy measure, as it represents the minimum channel capacity needed to assure average distortion equal to or less than D. Let us now consider the principal

branch in figure 3, the set of paths from $\mathbf{S}_0$ to $\mathbf{S}_f$, representing normal protein folding, taken as a communication channel having a given rate distortion function. The arguments of the previous section suggest that there will be an empirical Onsager relation in the gradient of the *rate distortion disorder*, an entropy-analog,

$$S_R \equiv R(D) - DdR(D)/dD \tag{20}$$

such that, over a life-history timeline,

$$dD/dt = f(dS_R/dD) \tag{21}$$

for some appropriate function f.

For a Gaussian channel, having $R(D) = (1/2)\log(\sigma^2/D)$, $S_R(D) = (1/2)\log(\sigma^2/D)+1/2$, the simplest possible Onsager relation becomes

$$dD/dt = -\mu dS_R dD = \mu/2D \tag{22}$$

with the explicit solution

$$D = \sqrt{\mu t}. \tag{23}$$

For an appropriate timescale – necessarily many orders of magnitude longer than the time of folding itself – the average distortion, representing the degree of misfolding, simply grows as a diffusion process in time. This is the simplest possible aging model, in which μ represents the accumulated impacts of epigenetic and broadly environmental effects including toxic exposures, nutrition, the richness of social interaction, and so on, over a lifetime.

A somewhat less simplistic model takes the Onsager relation as constrained by the availability of metabolic free energy, M, that powers active chaperone processes,

$$dD/dt = -\mu dS_R/dD - \kappa M = \mu/2D - \kappa M \tag{24}$$

where κ represents the efficiency of use of metabolic energy. This equation has the equilibrium solution (when $dD/dt = 0$)

$$D_{equlib} = \mu/2\kappa M. \tag{25}$$

Here aging is represented by a decay in the efficiency of those chaperone processes, i.e., a slow decline in κ, that may involve idiosyncratic dynamics, ranging from punctuated phase transitions to autocatalytic runaway effects, since D, in equation 8, acts as a temperature analog for a system able to undergo symmetry breaking.

More complicated models of this nature can be found in Wallace and Wallace (2010).

7.2 A metabolic model

Again, the 'dual' treatment focuses on $R(D)$, assuming that the probability density function for $R(D)$ at a given intensive index of embedding metabolic energy, M, can be described using an approach like equations 8 and 11:

$$Pr[R(D), \kappa M] = \frac{\exp[-R(D)/\kappa M]}{\int_{D_{min}}^{D_{max}} \exp[-R(D)/\kappa M] dD} \tag{26}$$

where κM represents the synergism between the intensity and physiological availability of the embedding free energy. At a fixed value of κM, again taking a life course timeframe as opposed to a folding timeframe, the mean of R is

$$< R >= \int_{D_{min}}^{D_{max}} R(D) Pr[R(D), \kappa M] dD. \tag{27}$$

For the Gaussian channel, $R(D) = (1/2)\log(\sigma^2/D), 0 \leq D \leq \sigma^2$, we obtain directly

$$< R >= \kappa M/(1 + 2\kappa M). \tag{28}$$

A decline in κ can, again, trigger complicated phase change dynamics for this system, as R itself, according to equation 11, can act as a temperature analog in a symmetry breaking argument, causing sudden, punctuated, changes in the underlying protein folding mechanisms.

Note that solving this equation for R in terms of κM produces a 'metabolic singularity' much like that proposed in figure 2.

Note also that taking the nonequilibrium Onsager relation

$$dD/dt = -\mu dS_R/dD - \frac{\mu}{2\sigma^2}\exp[\frac{2\kappa M}{1+2\kappa M}] \quad (29)$$

instead of $dD/dt - \mu dS_R dD - \kappa M$ as just above, gives

$$R_{eq} = \kappa M/(1+2\kappa M), \quad (30)$$

so that the two approaces are indeed dual.

8 Discussion and conclusions

The fidelity of the translation between genome and final protein conformation, measured by an average distortion measure, or its dual, the minimum channel capacity needed to limit average distortion to a given level, serve as evolutionarily-sculpted temperature analogs, in the sense of Onuchic and Wolynes (2004), to determine the possible phase transitions defining different degrees of protein symmetry. The protein folding funnel follows a spontaneous symmetry breaking mechanism with average distortion as the temperature analog, or, in the developmental picture, greater channel capacity leads more directly to the final state $\mathbf{S}_f$. These symmetries may perhaps be characterized by finite groupoid tilings as well as by the kinds of structures shown in figure 1.

The various outcomes to the full protein folding process – normal, corrected, eliminated, pathological – emerge, in the expanded 'Onsager relation' statistical model based on a cognitive paradigm, as distinct 'resilience' modes of a complicated internal cellular ecosystem, subject to punctuated transitions driven, in some cases, by signals from embedding epigenetic and ecological structures. Increase in the rate of folding disorders with age emerges through a long-time generalization of the Onsager model.

In a sense this work extends Tlusty's (2007) elegant topological exploration of the evolution of the genetic code, suggesting that rate distortion considerations are central to a broad spectrum of molecular biological phenomena, although different measures may come to the fore under different perspectives.

The cognitive paradigm introduced here opens a unified biological vision of protein folding and its disorders that may relate the etiology of a large set of misfolding and aggregation diseases more clearly to both cellular and epigenetic processes and environmental stressors. This would be, in the current reductionist sandstorm, no small thing. A cognitive paradigm subsumes epigenetic and environmental catalysis of protein conformation 'development' within a single grammar and syntax, and allows both normal folding and its pathologies to both be viewed as 'natural' outcomes, a perspective more consistent with rates of folding and aggregation disorders observed within an aging population.

Most basically, however, such a cognitive paradigm, as we have constructed it, will likely serve as the foundation for a new class of statistical tools – based on the asymptotic limit theorems of information theory rather than on the Central Limit Theorem alone – that should be useful in the analysis of data related to protein misfolding and aggregation disorders.

9 References

Andre, I., C. Strauss, D. Kaplan, P. Bradley, and D. Baker, 2008, Emergence of symmetry in homooligomeric biological assemblies, *Proceedings of the National Academy of Sciences*, 105:16148-16152.

Anfinsen, C., 1973, Principles that govern the folding of protein chains, *Science*, 181:223-230.

Ash, R., 1990, *Information Theory*, Dover, New York.

Atlan, H., and I. Cohen, 1998, Immune information, self-organization, and meaning, *International Immunology*, 10:711-717.

Bennett, C., 1988, Logical depth and physical complexity. In Herkin, R. (ed.), *The Universal Turing Machine: A Half-Century Survey*, Oxford University Press, pp. 227-257.

Bos, R., 2007, Continuous representations of groupoids. arXiv:math/0612639.

Brown, R., 1987, From groups to groupoids: a brief survey, *Bulletin of the London Mathematical Society*, 19:113-134.

Buneci, M., 2003, *Representare de Groupoizi*, Editura Mirton, Timosoara, Romania.

Cannas Da Silva, A., and Weinstein, A., 1999, *Geometric Models for Noncommutative Algebras*, American Mathematical Society, Providence, RI.

Champagnat, N., R. Ferriere, and S. Meleard, 2006, Unifying evolutionary dynamics: From individual stochastic processes to macroscopic models, *Theoretical Population Biology*, 69:297-321.

Cover, T., and H. Thomas, 1991, *Elements of Information Theory*, Wiley, New York.

Dembo, A., and O. Zeitouni, 1998, *Large Deviations and Applications*, 2nd edition, Springer, New York.

Diekmann U., and R. Law, 1996, The dynamical theory of coevolution: a derivation from stochastic ecological processes, *Journal of Mathemaical Biology*, 34:579-612.

Dill, K., S. Banu Ozkan, T. Weikl, J. Chodera, and V. Voelz, 2007, The protein folding problem: when will it be solved? *Current Opinion in Structural Biology*, 17:342-346.

Dobson, C., 2003, Protein folding and misfolding, *Nature*, 426:884-890.

Ellis, R., 1985, *Entropy, Large Deviations, and Statistical Mechanics*, Springer, New York.

Feynman, R., 2000, *Lectures on Computation*, Westview, New York.

Fillit, H., D. Nash, T. Rundek, and A. Zukerman, 2008, *American Journal of Geriatric Pharmacotherapy*, 6:100-118.

Glazebrook, J.F., and R. Wallace, 2009a, Small worlds and red queens in the global workspace: an information-theoretic approach, *Cognitive Systems Research*, 10:333-365,

Glazebrook, J.F., and R. Wallace, 2009b, Rate distortion manifolds as models for cognitive information, *Informatica*, 33:309-345.

Goldschmidt, L., P. Teng, R. Riek, and D. Eisenberg, 2010, Identifying the amylome, proteins capable of forming amyloid-like fibrils, *Proceedings of the National Academy of Sciences*, 107:3487-3492.

Goodsell, D., and A. Olson, 2000, Structural symmetry and protein function, *Annual Reviews of Biophysics and Biomolecular Structure*, 29:105-153.

Gunderson, L., 2000, Ecological resilience in theory and application, *Annual Reviews of Ecological Systematics*, 31:425-439.

Haataja, L., T. Gurlo, C. Huang, and P. Butler, 2008, Islet amylod in type 2 diabetes, and the toxic oligomer hypothesis, *Endocrine Reviews*, 29:303-316.

Holling, C., 1973, Resilience and stability of ecological systems, *Annual Reviews of Ecological Systematics*, 4:1-23.

Khinchin, A., 1957, *Mathematical Foundations of Information Theory*, Dover, New York.

Landau, L., and E. Lifshitz, 2007, *Statistical Physics, Part I*, Elsevier, New York.

Lei, J., and K. Huang, 2010, Protein folding: A perspective from statistical physics.

arXiv:10025013v1.

Levinthal, C., 1968, Are there pathways for protein folding? *Journal de Chimie Physique et de Physicochimie Biologique*, 65:44-45.

Levinthal, C., 1969. In *Mossbauer Spectroscopy*, Debrunner et al. (eds.), University of Illinois Press, Urbana, pp. 22-24.

Onuchic, J., and P. Wolynes, 2004, Theory of protein folding, *Current Opinion in Structural Biology*, 14:70-75.

Pettini, M., 2007, *Geometry and Topology in Hamiltonian Dynamics*, Springer, New York.

Plaxco, K., K. Simons, and D. Baker, 1998, Contact order, transition state placement and the refolding rates of single domain proteins, *Journal of Molecular Biology*, 277:985-994.

Protter, P. 1990, *Stochastic Integration and Differential Equations: A new approach*, Springer, New York.

Qiu, C., M. Kivipelto, and E. von Strauss, 2009, Epidemiology of Alzheimer's disease: occurrence, determinants, and strategies toward intervention, *Dialogues in Clinical Neuroscience*, 11:111-128.

Rockafellar, R., 1970, *Convex Analysis*, Princeton University Press, Princeton, NJ.

Scheuner, D., and R. Kaufman, 2008, The unfolded protein response: a pathway that links insulin demand with β-cell failure and daibetes, *Endocrine Reviews*, 29:317-333.

Sharma, V., V. Kaila, and A. Annila, 2009, Protein folding as an evolutionary process, *Physica A*, 388:851-862.

Tlusty, T., 2007, A model for the emergence of the genetic code as a transition in a noisy information channel, *Journal of Theoretical Biology*, 249:331-342.

Wallace, R., and M. Fullilove, 2007, *Collective Consciousness and its Discontents*, Springer, New York.

Wallace, R., and D. Wallace, 2008, Punctuated equilibrium in statistical models of generalized coevolutionary resilience: how sudden ecosystem transitions can entrain both phenotype expression and Darwinian selection, *Transactions on Computational Systems Biology IX*, LNBI 5121:23-85.

Wallace, R., and D. Wallace, 2009, Code, context, and epigenetic catalysis in gene expression, *Transactions on Computational Systems Biology XI*, LNBI 5750, 283-334.

Wallace, R., and D. Wallace, 2010, Cultural epigenetics: on the heritability of complex diseases, in press, *Transactions on Computational Systems Biology.*

Wallace, R., 2005, *Consciousness: A mathematical treatment of the global neuronal workspace model*, Springer, New York,

Wallace, R., 2009, Metabolic constraints on the eukaryotic transition, *Origins of Life and Evolution of Biospheres*, 39:165-176.

Wallace, R., 2010, Metabolic constraints on the evolution of genetic codes: Did multiple 'preaerobic' ecosystem transitions entrain richer dialects via Serial Endosymbiosis?

http://precedings.nature.com/documents/4120/version/3.

Wallach, J., and M. Rey, 2009, A socioeconomic analysis of obesity and diabetes in New York City, *Public Health Research, Practice, and Policy*, Centers for Disease Control and Prevention,

http://www.cdc.gov/pcd/issues/2009/jul/$08_0$215.*htm*.

Weinstein, A., 1996, Groupoids: unifying internal and external symmetry, *Notices of the American Mathematical Association*, 43:744-752.

Wolynes, P., 1996, Symmetry and the energy landscapes of biomolecules, *Proceedings of the National Academcy of Sciences*, 93:14249-14255.

Zhang, Q., Y. Wang, and E. Huang, Changes in racial/ethnic disparities in the prevalence of type 2 diabetes by obesity level among US adults, *Ethnicity and Health*, 14:439-457.

10 Mathematical Appendix

10.1 Basic ideas about groupoids

Following Weinstein (1996) closely, a groupoid, G, is defined by a base set A upon which some mapping – a morphism – can be defined. Note that not all possible pairs of states (a_j, a_k) in the base set A can be connected by such a morphism. Those that can define the groupoid element, a morphism $g = (a_j, a_k)$ having the natural inverse $g^{-1} = (a_k, a_j)$. Given such a pairing, it is possible to define 'natural' end-point maps $\alpha(g) = a_j, \beta(g) = a_k$ from the set of morphisms G into A, and a formally associative product in the groupoid $g_1 g_2$ provided $\alpha(g_1 g_2) = \alpha(g_1), \beta(g_1 g_2) = \beta(g_2)$, and $\beta(g_1) = \alpha(g_2)$. Then the product is defined, and associative, $(g_1 g_2)g_3 = g_1(g_2 g_3)$.

In addition, there are natural left and right identity elements λ_g, ρ_g such that $\lambda_g g = g = g\rho_g$ (Weinstein, 1996).

An orbit of the groupoid G over A is an equivalence class for the relation $a_j \sim Ga_k$ if and only if there is a groupoid element g with $\alpha(g) = a_j$ and $\beta(g) = a_k$. Following Cannas da Silva and Weinstein (1999), we note that a groupoid is called transitive if it has just one orbit. The transitive groupoids are the building blocks of groupoids in that there is a natural decomposition of the base space of a general groupoid into orbits. Over each orbit there is a transitive groupoid, and the disjoint union of these transitive groupoids is the original groupoid. Conversely, the disjoint union of groupoids is itself a groupoid.

The isotropy group of $a \in X$ consists of those g in G with $\alpha(g) = a = \beta(g)$. These groups prove fundamental to classifying groupoids.

If G is any groupoid over A, the map $(\alpha, \beta) : G \to A \times A$ is a morphism from G to the pair groupoid of A. The image of (α, β) is the orbit equivalence relation $\sim G$, and the functional kernel is the union of the isotropy groups. If $f : X \to Y$ is a function, then the kernel of f, $ker(f) = [(x_1, x_2) \in X \times X : f(x_1) = f(x_2)]$ defines an equivalence relation.

Groupoids may have additional structure. As Weinstein (1996) explains, a groupoid G is a topological groupoid over a base space X if G and X are topological spaces and α, β and multiplication are continuous maps. A criticism sometimes applied to groupoid theory is that their classification up to isomorphism is nothing other than the classification of equivalence relations via the orbit equivalence relation and groups via the isotropy groups. The imposition of a compatible topological structure produces a nontrivial interaction between the two structures. Below we will introduce a metric structure on manifolds of related information sources, producing such interaction.

In essence, a groupoid is a category in which all morphisms have an inverse, here defined in terms of connection to a base point by a meaningful path of an information source dual to a cognitive process.

As Weinstein (1996) points out, the morphism (α, β) suggests another way of looking at groupoids. A groupoid over A identifies not only which elements of A are equivalent to one another (isomorphic), but *it also parametizes the different ways (isomorphisms) in which two elements can be equivalent*, i.e., all possible information sources dual to some cognitive process. Given the information theoretic characterization of cognition presented above, this produces a full modular cognitive network in a highly natural manner.

Brown (1987) describes the fundamental structure as follows:

> A groupoid should be thought of as a group with many objects, or with many identities... A groupoid with one object is essentially just a group. So the notion of groupoid is an extension of that of groups. It gives an additional convenience, flexibility and range of applications...
>
> EXAMPLE 1. A disjoint union [of groups] $G = \cup_\lambda G_\lambda, \lambda \in \Lambda$, is a groupoid: the product ab is defined if and only if a, b belong to the same G_λ, and ab is then just the product in the group G_λ. There is an identity 1_λ for each $\lambda \in \Lambda$. The maps α, β coincide and map G_λ to λ, $\lambda \in \Lambda$.
>
> EXAMPLE 2. An equivalence relation R on [a set] X becomes a groupoid with $\alpha, \beta : R \to X$ the two projections, and product $(x, y)(y, z) = (x, z)$ whenever $(x, y), (y, z) \in R$. There is an identity, namely (x, x), for each $x \in X$...

Weinstein (1996) makes the following fundamental point:

> Almost every interesting equivalence relation on a space B arises in a natural way as the orbit equivalence relation of some groupoid G over B. Instead of dealing directly with the orbit space B/G as an object in the category S_{map} of sets and mappings, one should consider instead the groupoid G itself as an object in the category G_{htp} of groupoids and homotopy classes of morphisms.

The groupoid approach has become quite popular in the study of networks of coupled dynamical systems which can be defined by differential equation models, (e.g., Golubitsky and Stewart 2006).

10.2 Global and local symmetry groupoids

Here we follow Weinstein (1996) fairly closely, using his example of a finite tiling.

Consider a tiling of the euclidean plane R^2 by identical 2 by 1 rectangles, specified by the set X (one dimensional) where the grout between tiles is $X = H \cup V$, having $H = R \times Z$ and $V = 2Z \times R$, where R is the set of real numbers and Z the integers. Call each connected component of $R^2 \backslash X$, that is, the complement of the two dimensional real plane intersecting X, a tile.

Let Γ be the group of those rigid motions of R^2 which leave X invariant, i.e., the normal subgroup of translations by elements of the lattice $\Lambda = H \cap V = 2Z \times Z$ (corresponding to corner points of the tiles), together with reflections through each of the points $1/2\Lambda = Z \times 1/2Z$, and across the horizontal and vertical lines through those points. As noted by Weinstein (1996), much is lost in this coarse-graining, in particular the same symmetry group would arise if we replaced X entirely by the lattice Λ of corner points. Γ retains no information about the local structure of the tiled plane. In the case of a real tiling, restricted to the finite set $B = [0, 2m] \times [0, n]$ the symmetry group shrinks drastically: The subgroup leaving $X \cap B$ invariant contains just four elements even though a repetitive pattern is clearly visible. A two-stage groupoid approach recovers the lost structure.

We define the transformation groupoid of the action of Γ on R^2 to be the set

$$G(\Gamma, R^2) = \{(x, \gamma, y | x \in R^2, y \in R^2, \gamma \in \Gamma, x = \gamma y\},$$

with the partially defined binary operation

$$(x, \gamma, y)(y, \nu, z) = (x, \gamma\nu, z).$$

Here $\alpha(x, \gamma, y) = x$, and $\beta(x, \gamma, y) = y$, and the inverses are natural.

We can form the restriction of G to B (or any other subset of R^2) by defining

$$G(\Gamma, R^2)|_B = \{g \in G(\Gamma, R^2) | \alpha(g), \beta(g) \in B\}$$

[1]. An orbit of the groupoid G over B is an equivalence class for the relation

$x \sim_G y$ if and only if there is a groupoid element g with $\alpha(g) = x$ and $\beta(g) = y$.

Two points are in the same orbit if they are similarly placed within their tiles or within the grout pattern.

[2]. The isotropy group of $x \in B$ consists of those g in G with $\alpha(g) = x = \beta(g)$. It is trivial for every point except those in $1/2\Lambda \cap B$, for which it is $Z_2 \times Z_2$, the direct product of integers modulo two with itself.

By contrast, embedding the tiled structure within a larger context permits definition of a much richer structure, i.e., the identification of local symmetries.

We construct a second groupoid as follows. Consider the plane R^2 as being decomposed as the disjoint union of $P_1 = B \cap X$ (the grout), $P_2 = B \backslash P_1$ (the complement of P_1 in B, which is the tiles), and $P_3 = R^2 \backslash B$ (the exterior of the tiled room). Let E be the group of all euclidean motions of the plane, and define the local symmetry groupoid G_{loc} as the set of triples (x, γ, y) in $B \times E \times B$ for which $x = \gamma y$, and for which y has a neighborhood $\mathcal{U}$ in R^2 such that $\gamma(\mathcal{U} \cap P_i) \subseteq P_i$ for $i = 1, 2, 3$. The composition is given by the same formula as for $G(\Gamma, R^2)$.

For this groupoid-in-context there are only a finite number of orbits:

$\mathcal{O}_1$ = interior points of the tiles.

$\mathcal{O}_2$ = interior edges of the tiles.

$\mathcal{O}_3$ = interior crossing points of the grout.

$\mathcal{O}_4$ = exterior boundary edge points of the tile grout.

$\mathcal{O}_5$ = boundary 'T' points.

$\mathcal{O}_6$ = boundary corner points.

The isotropy group structure is, however, now very rich indeed:

The isotropy group of a point in $\mathcal{O}_1$ is now isomorphic to the entire rotation group O_2.

It is $Z_2 \times Z_2$ for $\mathcal{O}_2$.

For $\mathcal{O}_3$ it is the eight-element dihedral group D_4.

For $\mathcal{O}_4, \mathcal{O}_5$ and $\mathcal{O}_6$ it is simply Z_2.

These are the 'local symmetries' of the tile-in-context.