Nature Precedings : doi:10.1038/npre.2010.4847.2 : Posted 16 Sep 2010

Version 4

Without magic bullets: the biological basis for public health interventions against protein folding disorders

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

September 1, 2010

Abstract

Protein folding disorders of aging like Alzheimer's and Parkinson's diseases currently present intractable medical challenges. 'Small molecule' interventions – drug treatments – often have, at best, palliative impact, failing to alter disease course. The design of individual or population level interventions will likely require a deeper understanding of protein folding and its regulation than currently provided by contemporary 'physics' or culture-bound medical magic bullet models. Here, a topological rate distortion analysis is applied to the problem of protein folding and regulation that is similar in spirit to Tlusty's (2010a) elegant exploration of the genetic code. The formalism produces large-scale, quasi-equilibrium 'resilience' states representing normal and pathological protein folding regulation under a cellular-level cognitive paradigm similar to that proposed by Atlan and Cohen (1998) for the immune system. Generalization to long times produces diffusion models of protein folding disorders in which epigenetic or life history factors determine the rate of onset of regulatory failure, in essence, a premature aging driven by familiar synergisms between disjunctions of resource allocation and need in the context of socially or physiologically toxic exposures and chronic powerlessness at individual and group scales.

Key Words: aging, amyloid, cognitive paradigm, development, endoplasmic reticulum, epigenetic, phase transition, rate distortion, stress

1 Introduction

1.1 The basic conundrum

At this writing, front page articles in major news and scientific media trumpet the intractability of protein folding disorders with headlines like "No Magic Bullet Against Alzheimer's Disease" (Kolata, 2010). Medical magic bullets are, of course, a Western, and indeed particularly American, cultural conceit. Heine (2001), describing a similar paradigm within psychology, writes

> The extreme nature of American individualism suggests that a psychology based on late 20th century American research not only stands the risk of developing models that are particular to that culture, but of developing an understanding of the self that is peculiar in the context of the world's cultures...

Henrich et al. (2010) have elaborated this point in an instantly famous critique titled "The Wierdest people in the world?". Given the fundamental biological nature of protein folding itself, a magic bullet perspective on it's disorders may be analogously wierd, and inference based on American perceptions of current research similarly suspect (e.g., Kolata, 2010). Qui et al. (2009), based in Stockholm, present a different view:

> 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.

Qui et al. conclude:

*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.

> Epidemiological research has provided sufficient evidence that vascular risk factors in middle-aged and older adults play a significant role in the develolpment and progression of dementia and [Alzheimer's disease], whereas extensive social network and active engagement in mental, social, and physical activities may postpone the onset of the dementing disorder. Multidomain community intervention trials are warranted to determine to what extent preventive strategies toward optimal control of multiple vascular factors and disorders, as well as the maintenance of an active lifestyle, are effective against dementia and [Alzheimer's disease].

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.

Here we will take what Heine and colleagues might describe as an 'East Asian' approach to protein folding disorders, and examine their embedding context from a new perspective, as opposed to perceiving them as separated from their backcloth, and hence 'naturally' amenable to magic bullets.

1.2 Implications of protein folding disorders

High rates of protein folding and aggregation diseases, in conjunction with observations of the elaborate cellular folding regulatory apparatus associated with the endoplasmic reticulum and other cellular structures that compare produced to expected protein forms (e.g., Scheuner and Kaufman, 2008; Dobson, 2003), presents a clear and powerful logical challenge to simple physical 'folding funnel' free energy models of protein folding, as compelling as these are *in vitro* or *in silico*. This suggests that a more biologically-based model is needed for understanding the life course trajectory of protein folding, a model 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 (Anfinsen, 1973; Dill et al., 2007) surely capture part of the process (but see Chou and Carlacci, 1991). The diseases range from prion illnesses like Creutzfeld-Jakob disease, in addition to amyloid-related dysfunctions like Alzheimer's, Huntington's and Parkinson's diseases, and type 2 diabetes. Misfolding disorders include emphysema and cystic fibrosis.

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.

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.

Many of 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.

We will use Tlusty's (2007a, b, 2008a, b, c, 2010a) analysis of the emergence of the genetic code as a basis for an appropriate model, and begin, from the ground up, with a reconsideration of protein symmetry from his perspective.

2 Protein symmetries

There are, it seems, numerous underlying 'protein folding codes' in the sense of characteristic segments of amino acids whose ultimate folded structures are a somewhat debatable matter of formal taxonomy. Figure 1, from Hartl and Hayer-Hartl (2009), schematically expands the spectrum of final

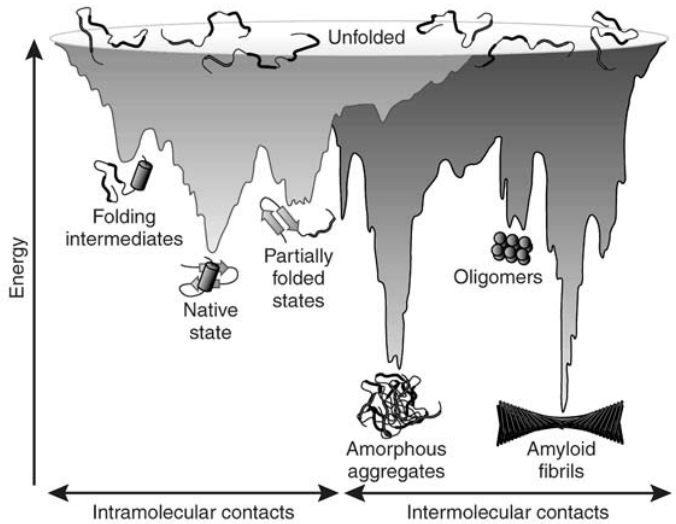

Figure 1: From Hartl and Hayer-Hartl, 2009. Energy landscape spectrum of protein folding and aggregation, parsed according to the degree of intra- vs. inter-molecular contact. Each energy valley defines an equivalence class, and the set of such classes defines the 'protein folding groupoid', in the sense of Weinstein (1996). Four basic classifications can be seen; native state, amorphous aggregates, semi-structured oligomers, and quasi-crystalline amyloid fibrils. Within the native state and the amyloid fibrils, systematic subclasses can be identified, leading to a fine structure for protein coding.

protein conformations according to an *in vivo* 'folding funnel' model dispersed across a measure of intra- vs. inter-molecular contact for hydrophobic-core proteins forming tertiary structure. Intra-molecular conformations involve three-dimensional assemblages of α-helices and β-sheets, while the most densely packed inter-molecular form is, perhaps, the ubiquitous semicrystalline amyloid fibril.

The basic spectrum of figure 1 for proteins having a hydrophobic core, in general, explains the necessity of the elaborate regulatory structures associated with the endoplasmic reticulum and its attendant spectrum of chaperone proteins (e.g., Scheuner and Kaufman, 2008), and the evolutionary pattern of protein sequences inferred by Goldschmidt et al. (2010). The inevitable corrosion of the cellular regulatory apparatus with age would then explain the subsequent onset of amyloid fibril and other aggregation disorders.

Most particularly, the spectrum of valleys in figure 1 characterizes a set of equivalence classes that defines a 'protein folding groupoid', in the sense of Weinstein (1996). As we will argue below, both the native state and amyloid fibril have structured subdivisions, internal equivalence classes, that define a nested set of groupoids. See the Mathematical Appendix for a summary of standard material on groupoids.

With regard to the disjunction between 'native' and 'amyloid' protein forms, very early on, Astbury (1935) conjectured that globular proteins could also have a linear state, based on pioneering x-ray studies. Chiti et al. (1999) argue that

> ...[P]rovided appropriate conditions are maintained over prolonged periods of time, the formation of ordered amyloid protofilaments and fibrils could be an intrinsic property of many polypeptide chains, rather than being a phenomenon limited to a very few aberrant sequences.

Wang et al. (2008), in a an elegant series of experiments on bacterial inclusion bodies, conclude that

> ..[A]myloid aggregation appears to be a common property of protein segments and consequently is observed in both eukaryotes and prokaryotes... [Thus] there must be evolved strategies against amyloid formation, which include both quality control mechanisms through molecular chaperones as well as sequence-based [evolutionary] prevention of amyloid aggregation...
>
> ...[E]ach protein may exist, not only in an unfolded or folded state, but, by containing at least one amino acid segment that is capable of participating in a sequence-specific, ordered, cross-β-sheet aggregated state, may also exist in an amyloid-like aggregate. The process of protein aggregation can thus be viewed as a primitive folding mechanism, resulting in a defined, aggregated conformation with each aggregated protein having its own distinctive properties.

Krebs et al. (2009), however, in a paper tellingly titled 'Protein aggregation: more than just fibrils', find that the amyloid fibril is not the only structure that aggregating proteins of widely different types may adopt. For example, the occurrence of spherulites, which have been found *in vivo* as well as *in vitro*, appears to be generic, although the factors that determine the equilibrium between free fibril and spherulite are not as yet clear. That is, we have not fully explained the spectrum implied by figure 1. Nevertheless, here we will use Tlusty's (2007a, b, 2008a, b, c, 2010a) arguments on the evolution of the genetic code to explore something of that spectrum.

As Kamtekar et al. (1993) point out, experimental studies of natural proteins show how their structures are remarkably tolerant to amino acid substitution, but that tolerance is limited by a need to maintain the hydrophobicity of interior side chains. Thus, while the information needed to encode a particular protein fold is highly degenerate, this degeneracy is constrained by a requirement to control the locations of polar and nonpolar residues. This is the precise protein folding analog to Tlusty's error network analysis of the genetic code, and his graph coloring arguments should thus apply, in some measure, to protein folding as well, allowing inference on the underlying structure of the 'protein folding codes' to be associated with the horizontal axis of figure 1.

Tyco (2006), likewise, argues that the amyloid fibril is a generically stable structural state of a polypeptide chain, competing thermodynamically and kinetically with globular monomeric states and unfolded monomeric states. Peptides

and proteins that are known to form amyloid fibrils have widely diverse amino acid sequences and molecular weights. He particularly finds that

> The near sequence independence of amyloid formation represents a challenge to our understanding of the physical chemistry of peptides and proteins.

Such sequence independence is, again, very precisely the degeneracy associated with Tlusty's error network approach.

Intermediate forms in figure 1 remain to be studied from this perspective.

Some of these matters have, of course, already been the subject of considerable attention. A series of elegant experiments by the Hecht group (e.g., Hecht et al., 2004), extending the work of Kamtekar et al. (1993), has focused on a basic understanding of protein folding through substitution of different polar and nonpolar amino acids in the construction of normal and fibril proteins. α-helices are found to be natural outcomes of amino acid sequences having a 3.6 residue/turn patten, i.e., a digital signal of the form 101100100110, where 1 indicates a polar, and 0 a nonpolar amino acid. The resulting three dimensional structures are formed by the propensity of the different residues to interact with an aqueous environment.

β sheets, on the other hand, emerge from a simpler period 2 code, e.g., 1010101, matching the structural repeat of the sheets. More recent work (Kim and Hecht, 2006) finds that generic hydrophobic residues of this form are sufficient to promote aggregation of the Alzheimer's Aβ42 peptide. However, while the positioning of hydrophobic residues is more important than the exact identities of the hydrophobic side chains for determining overall geometry, reaction kinetics, the rate of fibril formation, was profoundly affected by those identities. This suggests that the 'protein folding code' may be, in no small part, contextual, that is, determined as much by *in vivo* cellular regulatory machinery as by *in vitro* hydrophobic/hydrophilic physical interactions. This, we will suggest below, likely involves the operation of something like the catalytic mechanisms that Wallace and Wallace (2009) and Wallace (2010a) describe.

2.1 Large scale structure

Broadly, figure 1 embraces a four-fold classification (Wallace, 2010b):

1. The 'native state' determined, at low concentrations, entirely by the amino acid sequence in the classic sense of Anfinsen (1973).
2. Amorphous aggregates.
3. Semi-structured oligomers, as explored by Krebs et al. (2009).
4. Amyloid/amyloid-like one-dimensional fibrils.

Following the description by Tlusty, (2010b), the genetic code is a mapping of one codon to one amino acid. By contrast, the 'protein folding code' is a mapping of genes to folded amino acid chains, and the complexity gap between the two codes is very great indeed (e.g., Mirny et al., 2001). The strategy that allows adaptation of Tlusty's methods to protein folding is a coarse-graining of protein structure into a matrix of larger building blocks, e.g., α-helices and β-sheets. At this lower resolution a 'code' is a mapping between short DNA stretches, analogous to codons, and the convoluted motifs of proteins, playing the role of amino acids. As a consequence of the great tolerance to amino acid substitutions described above, as long as charge and polarity are conserved, it is possible to cluster all the sequences that encode the same structural motif. This greatly reduces the size of the resulting DNA sequence graph and thus limits the number of possible building blocks.

Generalizing Table 1 of Tlusty (2007, 2010a) according to the genus γ of the underlying graph, that is, the number of holes in the error network associated with the proposed code, we can apply Heawood's graph genus formula for the coloring number that identifies the maximal number of first excited modes of the coding graph Laplacian,

$$chr(\gamma) = Int[1/2(7 + \sqrt{1 + 48\gamma})]. \tag{1}$$

where Int is the integer value of the enclosed expression and γ itself is defined from Euler's formula (Tlusty, 2010) as

$$\gamma = 1 - \frac{1}{2}(V - E + F) \tag{2}$$

where V is the number of code network vertices, E the number of network edges, and F the number of enclosed faces.

Equation (1) produces the table

γ (# network holes)	chr(γ) (# prot. syms.)
0	4
1	7
2	8
3	9
4	10
5	11
6, 7	12
8, 9	13

In Tlusty's scheme, the second column represents the maximal possible number of product classes that can be reliably produced by error-prone codes having γ holes in the underlying coding error network.

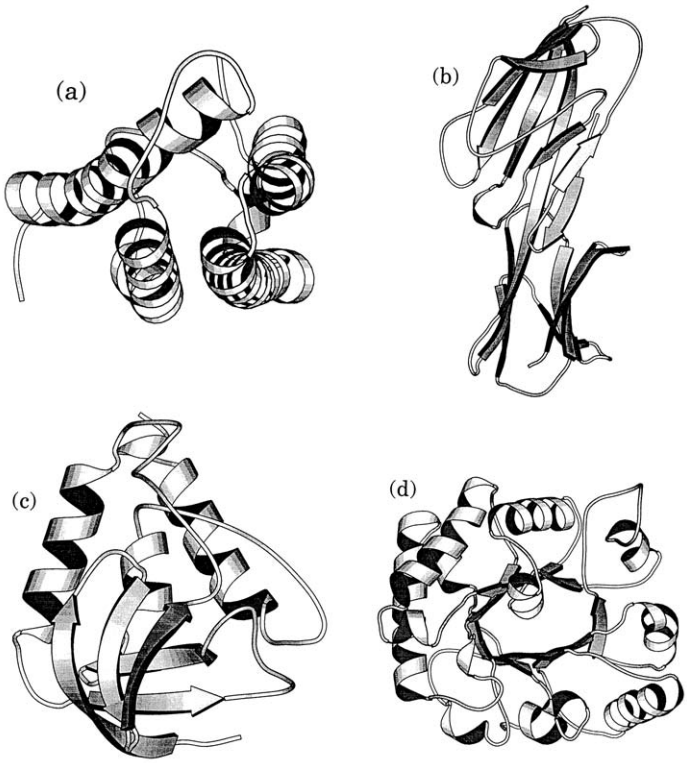

Figure 2: From Chou and Zhang, 1995. Standard equivalence classes for inexact protein symmetries according to Levitt and Chothia, 1976: (a) All-α helices. (b) All-β sheets. (c) $\alpha + \beta$. (d) α/β. More recent work identifies a minimum of seven, and possibly as many as ten, such classes (Chou and Maggiora, 1998).

From Tlusty's perspective, then, our four-fold classification for figure 1 produces a the simplest possible large-scale 'protein folding code', a sphere limited by the four-color problem, and the simplest cognitive cellular regulatory system would thus be constrained to pass/fail on four basic flavors, as it were, of folded proteins.

Within the funnel leading to the native state, however, chaperone processes would face far more difficult choices.

This suggests a possible two-fold cellular regulatory structure, and next we consider the two most fully characterized geometric structures in more detail, the normal and amyloid forms.

2.2 Normal globular proteins

Normal irregular protein symmetries were first classified by Levitt and Chothia (1976), following a visual study of polypeptide chain topologies in a limited dataset of globular proteins. Four major classes emerged; all α-helices; all β-sheets; α/β; and $\alpha + \beta$, as illustrated in figure 2.

While this scheme strongly dominates observed irregular protein forms, Chou and Maggiora (1998), using a much larger data set, recognize three more 'minor' symmetry equivalence classes; μ (multi-domain); σ (small protein); and ρ (peptide), and a possible three more 'subminor' groupings.

We infer that the normal globular 'protein folding code error network' is, essentially, a large connected 'sphere' – producing the four dominant structural modes of figure 2 – having one minor, and possibly as many as three more 'subminor' attachment handles, in the Morse Theory sense (Matsumoto, 2002), a matter opening up other analytic approaches.

2.3 Amyloid fibrils

As described above, Kim and Hecht (2006) suggest that overall amyloid fibril geometry is very much driven by the underlying β-sheet coding 1010101, although the rate of fibril formation may be determined by exact chemical constitution. Work by Sawaya et al. (2007) parses some of those subtleties: They identify an eight-fold 'steric zipper' symmetry necessarily associated with the linear amyloid fibrils that characterize a vast spectrum of protein folding disorders. Figure 3, adapted from their work, shows those symmetries. In essence, two identical sheets can be classified by the orientation of their faces (face-to-face/face-to-back), the orientation of their strands (with both sheets having the same edge of the strand up or one up and the other down), and whether the strands within the sheets are parallel or anti parallel. Five of the eight symmetry possibilities have been observed. This suggests, from the text table above, that the 'amyloid folding code error network' is a double donut, that is, has two, different sized, interior holes, resembling, perhaps, a toroid with a smaller attachment handle.

2.4 Amyloid self-replication

Maury (2009) has recently proposed an 'amyloid world' model for the emergence of prebiotic informational entities, based on the extraordinary stability of amyloid structures in the face of the harsh conditions of the prebiotic world. From this perspective, the synthesis of RNA, and the evolution of the RNA-protein world, were later, but necessary events for further biomolucular evolution. Maury further argues that, in the contemporary DNA$\Leftrightarrow$RNA$\Rightarrow$protein world, the primordial β-conformation-based information system is preserved in the form of a cytoplasmic epigenetic memory.

Falsig et al. (2008) examine the many different strains of prions, finding that differences in kinetics of the elementary steps of prion growth underlie the differential proliferation of prion strains, based on differential frangibility of prion fibrils. They argue that an important factor is the size of the stabilizing cross-β amyloid core that appears to define the physical properties of the resulting structures, including their propensity to fragment, with small core sizes leading to enhanced frangibility. In terms of the protein folding funnel approach, they find that intrinsic frustration implies that several distinct arrangements favoring a certain subset of globally incompatible interactions are possible, reflecting the observed strain-dependent differences in the parts of the sequence incorporated into the fibril core.

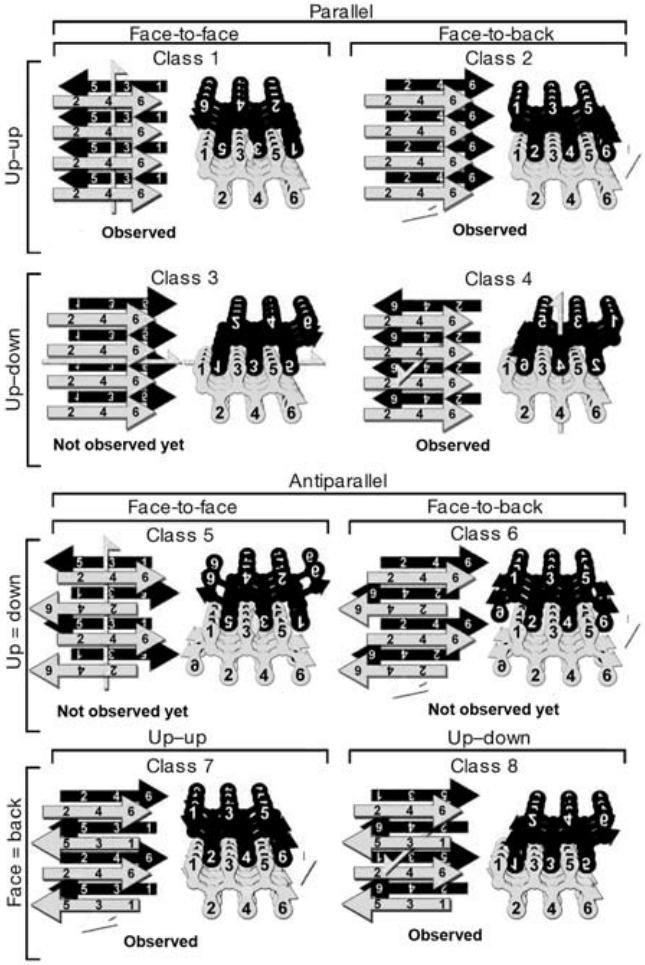

Figure 3: From Sawaya et al., 2007. The eight possible steric zipper symmetry classifications for amyloid fibrils.

In addition, they argue, there are unexplored similarities between Alzheimer's and prion diseases, that is, the analogies between prion and Aβ aggregates could be broader than initially suspected.

Given the eight-fold symmetry of the amyloid fiber, say versions A $\rightarrow$ H, then the simplest 'frangibility code' is the set of identical pairings: $\{AA, BB, ..., GG, HH\}$, producing eight different possible structures and their reproduction by fragmentation. More complex prion symmetries, or the possibility of combinatorial recombination, would allow a much richer structure, producing quasi-species, in the sense of Collinge and Clarke (2007). Permitting different sequence lengths or explicitly identifying different sequence orders would vastly enlarge what Collinge has characterized as a 'cloud' of possibilities, in the case of prion diseases. Indeed, classic studies by Bruce and Dickinson (1987) found 15 or more different prion strains in a mouse model.

Recent work on prions appears to support something of Maury's hypothesis. Li et al. (2010) find that infectious prions, mainly what is called PrP^{Sc}, a spectrum of β sheet-rich conformers of the normal host protein PrP^{C}, undergo Darwinian evolution in cell culture. In that work, prions show the evolutionary hallmarks: they are subject to mutation, as evidenced by heritable changes of their phenotypes, and to selective amplification, as found by the emergence of distinct populations in different environments. Figure 4, from Li et al. (2010), shows a prion energy landscape similar to figure 1. This suggests the possibility of characterizing the underlying topology of a 'prion reproduction code', in the sense of the sections above.

One might speculate that prions and prion diseases represent fossilized remains of Maury's prebiotic amyloid world.

3 Spontaneous symmetry breaking

We begin the theoretical analysis of protein folding dynamics with a classic conceptual context:

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

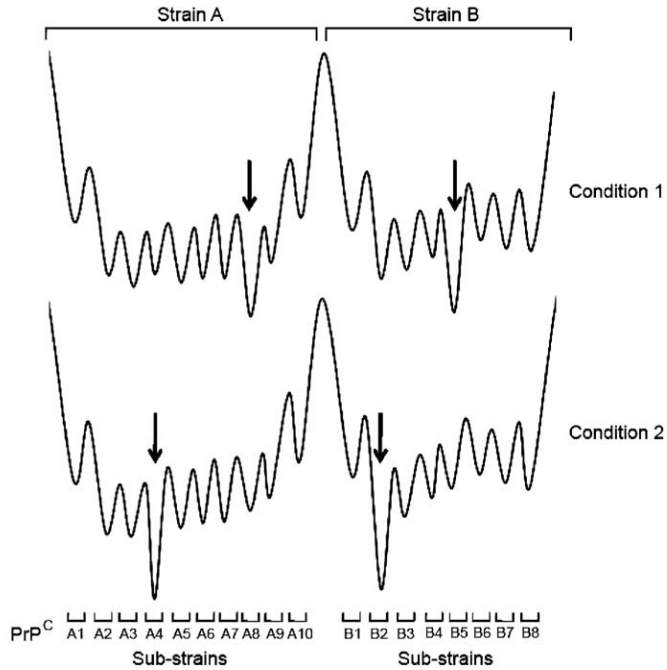

Figure 4: From Li et al. 2010, figure S10. Schematic energy landscape for prion strains and substrains. The energy landscape diagram suggests that substrains are distinguishable collectives of prions that interconvert reproducibly and readily because they are separated by low activation energy barriers. The properties of a strain may vary depending on the environment in which it replicates, as the proportions of component substrains may change to favor that replicating most rapidly, indicated by the arrows. Comparison with figure 1, and the subsequent argument, suggests an underlying topological structure for a 'prion reproduction code'.

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, where symmetry is measured in terms of classic group structures.

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 and Fullilove, 2008). 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 (Chou and Zhang, 1995).

4 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 a particular real-world physiological context, and produces the final symmetric protein. We then suppose it possible to compare what is actually produced with what should have been produced – what the codon stream proposes, taking Anfinsen's (1973) perspective – and what the protein production machinery disposes – comparing observed folded proteins with their idealized image, that can now be well described using 'physics' models like Rosetta. Such comparison is entirely an empirical matter, and can be done by human experimenters, but is most often done in real time by internal cellular machinery: the endoplasmic reticulum and friends.

The average distortion between what is sent by the codon stream and what is observed in the cell or tissue is an essential parameter of the transmission channel, and the relation between the minimum channel capacity needed for some average distortion measure is a fundamental empirical characteristic of an information channel, characterized by the Rate Distortion Theorem, one of the basic asymptotic limit theorems of probability theory. The rate distortion function, $R(D)$, that measures the minimimum channel capacity ensuring an average distortion D, by some measure is, in essence, a different way of looking at the protein folding funnel.

Onuchic and Wolynes (2004) have put something of 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 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 – that is, 'decoded' in engineering jargon – back into a reproduction of the original biological signal, $b^n \rightarrow \hat{y}^n = \hat{y}_1, \hat{y}_2, ...$ To reiterate, such decoding can be done by human experimenters, taking Anfinsen's viewpoint that the codon stream characterizes the intended protein form as a message that is distorted by transmission along the DNA $\rightarrow$ RNA $\rightarrow$ Protein channel.

Define a distortion measure $d(y, \hat{y})$ that compares the original to the retranslated/decoded 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

$$D \equiv \sum_{y^n} p(y^n) d(y^n, \hat{y}^n). \quad (3)$$

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

$$R(D) = \min_{p(y,\hat{y}); \sum_{(y,\hat{y})} p(y)p(y|\hat{y})d(y,\hat{y}) \leq D} I(Y, \hat{Y}). \quad (4)$$

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 that 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),

$$R(D) = 1/2 \log[\sigma^2/D], 0 \leq D \leq \sigma^2$$
$$R(D) = 0, D > \sigma^2. \quad (5)$$

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

$$H[X] \leq C, \tag{6}$$

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

$$C \equiv \max_{P(X)} I(X|Y), \tag{7}$$

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),

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

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

$$= \lim_{n\to\infty} \frac{H(X_0, ..., X_n)}{n+1}, \tag{8}$$

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 (8), 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 5 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 6 of Wallace (2010a) gives a fairly elementary derivation of such a relation in terms of rate distortion theory. Figure 5 suggests that H may indeed be a good, if highly nonlinear, index of large-scale free energy dynamics.

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 $H_1 > H_2$

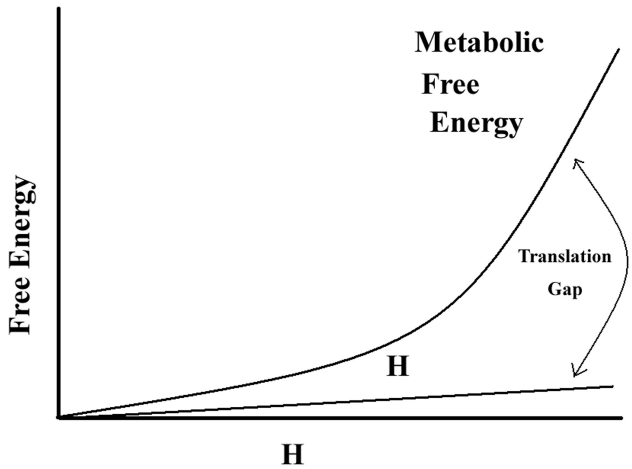

Figure 5: Nonlinear increase in metabolic free energy needed to maintain and generate linear increase in the information source uncertainty of a complex physiological process. H is seen to 'leverage' metabolic expenditures, parameterizing a more complicated nonequilibrium thermodynamics. See Wallace (2010a) for an explicit calculation.

and system 1 represents a more energetic process than system 2, and by the arguments of figure 5, may trigger even greater metabolic free energy dynamics.

By equations (6-8), the Rate Distortion Function, $R(D)$ is likewise a free energy measure, constrained by the availability of metabolic free energy.

5 What 'decodes the codon'?

A number of commentators have raised the question of what actually observes the rate distortion function, i.e., what decodes the codon and makes a comparison between what is expected from the sequence of codons and the protein that is actually produced? The question is, in a different idiom, a fundamental conundrum in cellular biology, as it appears to violate the central dogma of molecular biology that information flows from DNA to RNA to protein. There is, in fact, an elaborate cellular level apparatus that does the essential decoding. Hebert and Molinari (2007) put the matter thus:

> Understanding the mechanisms regulating degradation of folding-defective polypeptides expressed in the [endoplasmic reticulum, ER] is one of the central issues of cell biology. Rapid disposal of folding-incompetent polypeptides produced in the ER lumen is instrumental to maintain ER homeostasis. The degradation machinery is easily saturated. Defective adaptation of the cellular degradation capacity to the ER load may result in accumulation of aberrant polypeptides that eventually impairs the ER capacity to assist maturation of newly synthesized secretory proteins.

Thus cells invoke a complicated, highly evolved, internal regulatory process, incorporating the elaborate machinery of the endoplasmic reticulum and chaperone proteins, that decodes the codon, i.e., compares a produced protein with an internal (inherited or learned) pattern, and then chooses one among several possible actions based on the comparison: pass the protein on to the next stage, attempt to repair a damaged protein, attempt to recycle or eliminate a protein that cannot be repaired. After further theoretical development, Section 8 below will rephrase protein folding and regulation in terms of a cognitive paradigm that formalizes this decision process at cellular and higher – as opposed to the molecular – levels of analysis, thus finessing the apparent violation of the central dogma.

From a broader perspective, however, there is another, quite relentless if crude, mechanism for decoding the codon, and that is the continued life of the cell, tissue, or organism, i.e., Darwin's survival of the fittest, writ small: protein misfolding kills.

But the basic question – what decodes the codon – is, in fact, a misundersdtanding of the regularities inherent in all information transmission. The rate distortion theorem is, to any form of information transmission, as basic as the central limit theorem is to sums of stochastic variates, and it is no surprise that many mechanisms exist in nature to decode/retranslate the codon. The essential point is that $R(D)$ is a fundamental empirical characteristic of any information transmission channel, and it can be measured by internal cellular, or external human, agencies. That is, a clever experimenter could study a cellular process, infer from the codon stream what protein should be produced, and then compare the actual with the intended product, calculating a distortion measure, and determining the minimum channel capacities needed to map out $R(D)$. This would be as arduous as, but no more arduous than, measuring the protein folding funnel, and indeed, the folding funnel and the rate distortion function appear to be different images of the same phenomenon.

The endoplasmic reticulum and friends appears to do such measurement in situ and almost in real time, triggering heat shock protein corrective mechanisms as needed. This is no small evolutionary accomplishment.

6 The energy picture

Ash's comment above 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 most direct model is parameterized by the average 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

$$Pr[\mathcal{H}_j] = \frac{\exp[-\mathcal{H}_j/\lambda D]}{\sum_{i=1}^{n} \exp[-\mathcal{H}_i/\lambda D]}, \tag{9}$$

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

$$F_R = -\lambda D \log[\sum_{i=1}^{n} \exp[-\mathcal{H}_i/\lambda D]]. \tag{10}$$

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'.

7 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 evolutionarily-sculpted 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

$$H = \lim_{n\to\infty} \log[N(n)]/n \tag{11}$$

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.

7.1 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 the Appendix. 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.

7.1.1 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

$$P[H_\beta] = \frac{\exp[-H_\beta/\mu R]}{\sum_\alpha \exp[-H_\alpha/\mu R]}, \tag{12}$$

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.

7.1.2 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 (Weinstein, 1996), as described in the Appendix. 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

$$P[H_{G_i}] = \frac{\exp[-H_{G_i}/\kappa R]}{\sum_j \exp[-H_{G_j}/\kappa R]}, \tag{13}$$

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

$$Z_G = \sum_j \exp[-H_{G_j}/\kappa R]. \tag{14}$$

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

$$F_G[R] = -\frac{1}{\kappa R}\log[Z_G[R]]. \tag{15}$$

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 $\mathbf{S}_f$.

7.2 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 rate distortion 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 (12), 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

$$P[H_\beta] = \frac{\exp[-H_\beta/\mu R]}{\sum_\alpha \exp[-H_\alpha/\mu R]} \tag{16}$$

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.

In particular, if R is fixed, then the log of the folding rate will be given as

$$\log[P[H_\beta]] = \log[\frac{\exp[-H_\beta/\mu R]}{\sum_\alpha \exp[-H_\alpha/\mu R]}] = C(R) - H_\beta/\mu R, \quad (17)$$

where $C(R)$ is positive. β indexes increasing topological complexity, using some appropriate measure.

For simplicity, assume $H_\beta \propto \beta$. Then, taking an integral approximation,

$$P[\beta] \approx \frac{\exp[-m\beta/\mu R]}{\int_{\alpha=0}^{\infty} \exp[-m\alpha/\mu R] d\alpha} = (m/\mu R) \exp[-m\beta/\mu R], \quad (18)$$

and

$$\log[P[\beta]] \approx \log[m/\mu R] - m\beta/\mu R. \quad (19)$$

Thus one expects, at a fixed R defining a maximum channel capacity, that

$$\log[FoldingRate] \approx C - k\beta, \quad (20)$$

C, k constant and all values positive.

A standard index of protein complexity is the absolute contact order (Plaxco et al, 1998):

$$ACO = 1/N \sum^N \Delta L_{i,j} \quad (21)$$

where N is the number of contacts within 6 Angstroms between nonhydrogen atoms in the protein, and $\Delta L_{i,j}$ is the number of residues separating the interacting pair of nonhydrogen atoms.

Adjacent residues are assumed to be separated by one residue.

Figure 6, adapted from Gruebele (2005), shows the correlation of the log of the folding rate with fold complexity, measured by the ACO. The upper line estimates folding speed limited only by fold complexity, following Yang and Gruebele (2004), and seems clearly to represent a maximum possible rate distortion function/channel capacity, according to equation (20). The molecular species along the lower curve are assumed to be 'frustrated' by an irregular folding funnel, and follow a narrow spectrum of relations like equation (20), necessarily below the line defined by maximum channel capacity, and necessarily somewhat scattered, according to the variation in R.

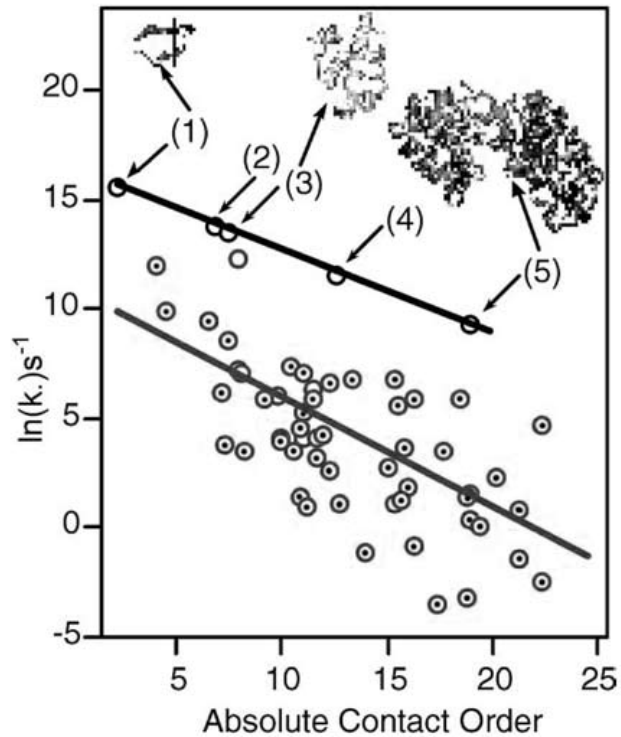

Figure 6: From Gruebele, 2005. Relation of the log of the protein folding rate to fold complexity. The upper folding speeds are limited only by fold complexity, without the 'frustration' effects of a rough folding funnel. Frustration, in this model, is equivalent to increasing noise that constrains channel capacity, and drives R irregularly lower than the value implied by the relation for the fastest folders. Equations (19) and (20) reproduce something of these results.

It is possible to reproduce something like figure 6 by describing 'smooth' and 'rough' folding funnels in terms of a Gaussian channel, that is, one in which the signal transmission from initial to final protein state is perturbed by Gaussian noise having a squared-error distortion, so that the rate distortion function has the standard form of equation (5), $R(D) = (1/2)\log[\sigma^2/D]$. Again, $R(D)$ is the rate distortion function at average distortion D, and σ^2 represents the amplitude of the imposed random noise. A smooth folding funnel would have little noise.

Plugging equation (5) into equation (19) gives, over an appropriate range of parameters, the spectrum of linear relations for log folding rate shown in figure 7. D, m, and μ are fixed, and β and σ^2 increase, as indicated.

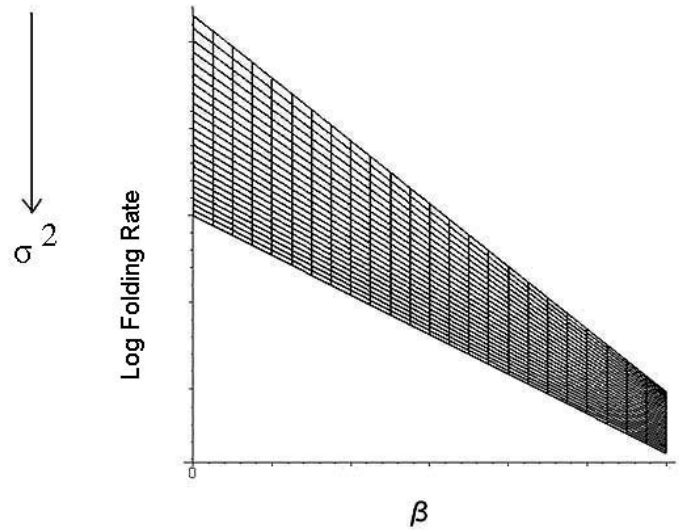

Figure 7: Spectrum of linear relations between log folding rate and increasing topological complexity for increasing 'roughness' of the folding funnel, as measured by noise σ^2 for a Gaussian channel. β increases to the right and σ^2 increases downward, analogous to an increasingly irregular folding funnel.

These matters lead to the next central question: can folding rates be modulated by other means than noise in the folding funnel? Can the effects of noise be 'reversed'? This will lead toward our cognitive model for protein folding.

7.3 Catalysis of protein folding

Incorporating the influence of embedding contexts – epigenetic or cellular regulatory chaperone effects, or the impact 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}.

7.4 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 8 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 7.3. The topological equivalence classes define a groupoid on the developmental system.

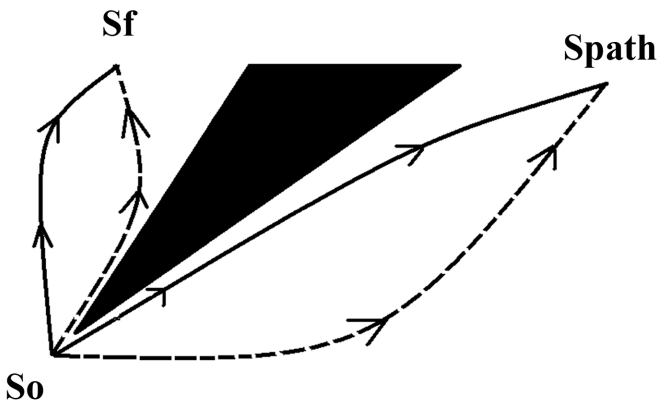

Figure 8: Given an initial state $\mathbf{S}_0$ and a critical period casting a path-dependent developmental shadow, there are two different directed homotopy equivalence classes of deformable paths leading, respectively, to the normal folded protein state $\mathbf{S}_f$ and the pathological state – e.g., amyloid – $\mathbf{S}_{path}$. These sets of paths form equivalence classes defining a topological groupoid.

8 Toward a cognitive paradigm for protein folding disorders

We now take the developmental perspective as the foundation for generating an empirically-based statistical model – effectively a cognitive paradigm for normal and pathological protein folding – that incorporates 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 $\mathbf{S}_0$ – 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

$$\mathbf{S}_{i+1} = f([\mathbf{S}_i, \mathbf{W}_i]) = f(a_i) \quad (22)$$

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 8 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

$$H = \lim_{n\to\infty} \frac{\log[N(n)}{n} \quad (23)$$

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 (8).

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 8 and the added catalysis arguments of Section 7.3.

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 6 and 7 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 the endoplasmic reticulum, chaperones, and other corrective mechanisms, involves 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

$$H_k = H_k(K_1, ..., K_s, ..., H_j, ...) \tag{24}$$

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

$$S_k = H_k - \sum_i Q_i \partial H_k / \partial Q_i, \tag{25}$$

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:

$$dQ_t^j = \sum_i [L_{j,i}(t, ..., Q_k, ...)dt + \sigma_{j,i}(t, ..., Q_k, ...)dB_t^i] \tag{26}$$

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).

There are several obvious possible dynamic patterns for the system of equation (26):

1. Setting equation (26) equal to zero and solving for stationary points gives attractor states since the noise terms preclude unstable equilibria.

2. This system may converge to limit cycle or pseudorandom 'strange attractor' behaviors in which the system seems to chase its tail endlessly within a limited venue – a traditional coevolutionary 'Red Queen' (Wallace and Wallace, 2009).

3. What is converged to in both cases is not a simple state or limit cycle of states. Rather it is an equivalence class, or set of them, of highly dynamic information sources coupled by

mutual interaction through crosstalk. Thus 'stability' in this structure represents particular patterns of ongoing dynamics rather than some identifiable static configuration.

Here we become deeply enmeshed in a system of highly recursive phenomenological stochastic differential equations, but in a dynamic rather than static manner. The objects of this dynamical system are equivalence classes of information sources, rather than simple 'stationary states' of a dynamical or reactive chemical system. Imposition of necessary conditions from the asymptotic limit theorems of communication theory has beaten the mathematical thicket back one full layer.

These results are essentially similar to those of Diekmann and Law (1996), who invoke evolutionary game dynamics to obtain a first order canonical equation for coevolutionary systems having the form

$$ds_i/dt = K_i(s)\partial W_i(s'_i, s)|_{s'_i=s_i}. \quad (27)$$

The s_i, with $i = 1, ..., N$ denote adaptive trait values in a community comprising N species. The $W_i(s'_i, s)$ are measures of fitness of individuals with trait values s'_i in the environment determined by the resident trait values s, and the $K_i(s)$ are non-negative coefficients, possibly distinct for each species, that scale the rate of evolutionary change. Adaptive dynamics of this kind have frequently been postulated, based either on the notion of a hill-climbing process on an adaptive landscape or some other sort of plausibility argument.

When this equation is set equal to zero, so there is no time dependence, one obtains what are characterized as 'evolutionary singularities' or stationary points.

Equation (26) above is similar, although focused on information sources representing protein folding regulation, allowing elaborate patterns of phase transition punctuation in a natural manner.

Champagnat et al. (2006), in fact, derive a higher order canonical approximation extending equation (27) that is closer to equation (26), i.e., a stochastic differential equation describing coevolutionary dynamics. Champagnat et al. extend the argument, using a large deviations approach to analyze dynamical coevolutionary paths, not merely quasi-stable singularities. They contend that in general, the issue of evolutionary dynamics drifting away from trajectories predicted by the canonical equation can be investigated by considering the asymptotic of the probability of 'rare events' for the sample paths of the diffusion.

By 'rare events' they mean diffusion paths drifting far away from the canonical equation. The probability of such rare events is governed by a large deviation principle: when a critical parameter (designated ϵ) goes to zero, the probability that the sample path of the diffusion is close to a given rare path ϕ decreases exponentially to 0 with rate $I(\phi)$, where the 'rate function' I can be expressed in terms of the parameters of the diffusion. This allows study of long-time behavior of the diffusion process when there are multiple attractive singularities. Under proper conditions the most likely path followed by the diffusion when exiting a basin of attraction is the one minimizing the rate function I over all the appropriate trajectories. The time needed to exit the basin is of the order $\exp(H/\epsilon)$ where H is a quasi-potential representing the minimum of the rate function I over all possible trajectories.

An essential fact of large deviations theory is that the rate function I which Champagnat et al. invoke can almost always be expressed as a kind of entropy, that is, in the form $I = -\sum_j P_j \log(P_j)$ for some probability distribution. This result goes under a number of names; Sanov's Theorem, Cramer's Theorem, the Gartner-Ellis Theorem, the Shannon-McMillan Theorem, and so forth (e.g., Dembo and Zeitouni, 1998). A detailed example is given in R. Wallace and R.G. Wallace (2008).

These considerations lead very much in the direction of equation (26), seen as subject to internally-driven large deviations *that are themselves described as information sources*, providing H-parameters that can trigger punctuated shifts between quasi-stable modes, in addition to resilience transitions driven by external catalytic events.

Indeed, the direct inclusion of large deviations regularities within the context of the statistical model of equation (26) suggests that other factors that can be characterized in terms of information sources may be directly included within the formalism. Section 6.1 of Wallace et al. (2009), for example, explores the impact of culture, taken as a generalized language, on the evolution of human pathogens.

The basic statistical model is illustrated by figure 9. Here, two quasi-equilibria – one normal, one pathological – are characterized by diffusive drift about their singularities in a two dimensional system, but are coupled by a highly structured large deviation connecting them. That large deviation excursion is by means of an information source having a 'grammar' and a 'syntax', rather than representing a random event. Understanding that grammar and syntax would, in this model, represent understading the etiology of a protein folding disorder.

9 Aging and protein folding: extending the time scale

9.1 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 (4-8) suggest that the rate distortion function,

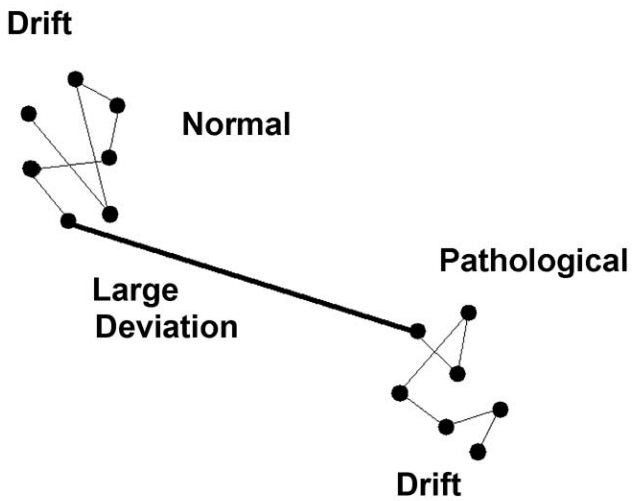

Figure 9: Adapted from Wallace (2010c). Dynamic behavior of the system obtained by setting equation (26) to zero. Diffusive drift about a 'normal' protein folding quasi-equilibrium is interrupted by a highly structured large deviation leading to a pathological quasi-equilibrium.

$R(D)$, 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 8, 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,

$$S_R \equiv R(D) - DdR(D)/dD \tag{28}$$

such that, over a life-history timeline,

$$dD/dt = f(dS_R/dD) \tag{29}$$

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

$$dD/dt = -\mu dS_R dD = \mu/2D \tag{30}$$

with the explicit solution

$$D = \sqrt{\mu t}. \tag{31}$$

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,

$$dD/dt = -\mu dS_R/dD - \kappa M = \mu/2D - \kappa M \tag{32}$$

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

$$D_{equlib} = \mu/2\kappa M. \tag{33}$$

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 (9), 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).

9.2 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 (9) and (12):

$$Pr[R(D), \kappa M] = \frac{\exp[-R(D)/\kappa M]}{\int_{D_{min}}^{D_{max}} \exp[-R(D)/\kappa M] dD} \quad (34)$$

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

$$< R >= \int_{D_{min}}^{D_{max}} R(D) Pr[R(D), \kappa M] dD. \quad (35)$$

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

$$< R >= \kappa M/(1 + 2\kappa M). \quad (36)$$

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 M in terms of R produces a 'metabolic singularity' much like that proposed in figure 5.

Note also that taking the nonequilibrium Onsager relation

$$dD/dt = -\mu dS_R/dD - \frac{\mu}{2\sigma^2} \exp[\frac{2\kappa M}{1+2\kappa M}] \quad (37)$$

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

$$R_{eq} = \kappa M/(1 + 2\kappa M), \quad (38)$$

so that the two approaches are indeed dual.

10 Concluding remarks

The fidelity of the translation between genome and final protein conformation, characterized 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 equivalence class groupoids like those figure 1.

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

In essence, this work extends Tlusty's (2010a) 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 *in vivo* 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 common misfolding and aggregation diseases more clearly to both cellular and epigenetic processes and environmental stressors (e.g., Schnabel, 2010). This would be, in the current reductionist sandstorm (e.g., Kolata, 2010), 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.

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.

We have, in the sense of Heine (2001) and Wallace (2007), focused on the broad physiological context of protein folding and its disorders, a context that includes epigenetic and life history stress factors that can act as catalysts to induce highly structured 'large deviations' that accelerate the deterioration of protein folding regulation. And we have done this from the ground up, as it were, providing a 'basic biological' model of

what Qui et al. (2009) and others have observed. A narrow focus on medical magic bullets (Kolata, 2010) is not consonant with the broad scale of protein folding regulation and its dysfunctions, and a successful search for effective interventions will necessarily involve far broader perspectives than seem comfortable to the strongly culture-bound majority of senior American researchers.

11 Acknowledgments

The author thanks D. Eisenberg, M. Grubele, M. Hecht, and D.N. Wallace for useful discussions.

12 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.

Astbury, W., 1935, The x-ray interpretation of the denaturation and the structure of the seed gobulins, *Biochemistry*, 29:2351-2360.

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

Beck, C., and F. Schlogl, 1995, *Thermodynamics of Chaotic Systems*, Cambridge University Press, New York.

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.

Bruce, M., and A. Dickinson, 1987, Biological evidence that scrapie agent has an independent genome, *Journal of General Virology*, 68:79-89.

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.

Chiti, F., P. Webster, N. Taddei, A. Clark, M. Stefani, G. Ramponi, and C. Dobson, 1999, Designing conditions for *in vitro* formation of amyloid protofilaments and fibrils, *Proceedings of the National Academy of Sciences of America*, 96:3590-3594.

Chou, K.C., and L. Carlacci, 1991, Energetic approach to the folding of α/β barrels, *Proteins: Structure, Function and Genetics*, 9:280-295.

Chou, K.C., and C.T. Zhang, 1995, Prediction of protein structural classes, *Reviews in Biochemistry and Molecular Biology*, 30:275-349.

Chou, K.C., and G. Maggiora, 1998, Domain structural class prediction, *Protein engineering*, 11:523-528.

Chou, K., and Y.D. Cai, 2004, Predicting protein structural class by functional domain composition, *Biochemical an Biophysical Research Communications*, 321:1007-1009.

Collinge, J., and A. Clarke, 2007, A general model of prion strains and their pathogencity, *Science*, 318:930-936.

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.

Falsig, J., K. Nilsson, T. Knowles, and A. Aguzzi, 2008, Chemical and biophysical insights into the propagation of prion strains. *HFSP Journal*, 2:332-341.

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.

Golubitsky M., and I. Stewart, 2006, Nonlinear dynamics and networks: the groupoid formalism, *Bulletin of the American Mathematical Sociey*, 43:305-364.

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

Grubele, M., 2005, Downhill protein folding: evolution meets physics, *Comptes Rendus Biologies*, 328:701-712.

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.

Hartl, F., and M. Hayer-Hartl, 2009, Converging concepts of protein folding *in vitro* and *in vivo*, *Nature Structural and Molecular Biology*, 16:574-581.

Hebert, D., and M. Molinari, 2007, Protein folding, quality control, degradation, and related human diseases, *Physiologal Reviews*, 87: 1377-1408.

Hecht, M., A. Das, A. Go, L. B. Aradley, and Y. Wei, 2004, *Protein Science*, 13:1711-1723.

Heine, S., 2001, Self as cultural product: an examination of East Asian and North American selves, *Journal of Personality*, 69:881-906.

Henrich, J., S. Heine, and A. Norenzayan, 2010, The Weirdest people in the world, *Behavioral and Brain Sciences*, 33:61-135.

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

Ivankov, D., S. Garbuzynsky, E. Alm, K. Plaxco, D. Baker, and A. Finkelstein, 2003, Contact order revisited: influence of protein size on the folding rate, *Protein Science*, 12:2057-2062.

Kamtekar, S., J. Schiffer, H. Xiong, J. Babik, and M. Hechtg, 1993, Protein deisgn by patterning of polar and nonpolar amino acids. *Science*, 262:1680-1685.

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

Kim, W., and M. Hecht, 2006, Generic hydrophobic residues are sufficient to promote aggregation of the Alzheimer's Aβ42 peptide, *Proceedings of the National Academy of Sciences USA*, 103:552-557.

Kolata, G., Years later, no magic bullet against Alzheimer's disease, *New York Times*, 8/28/2010:1.

Krebs, M., K. Domike, and A. Donald, 2009, Protein aggregation: more than just fibrils, *Biochemical Society Transactions*, 37(part 4):682-686.

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.

Lei, J., S. Browning, S. Mahal, A. Oelschlegel, and C. Weissman, 2010, Darwinian evolution of prions in cell culture, *Science*, 327:869-872.

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.

Levitt, M., and C. Chothia, 1976, Structural patterns in globular proteins, *Nature*, 261:552-557.

Matusmoto, Y., 2001, *An Introduction to Morse Theory*, Translations of the American Mathematical Society 208, Providence, RI.

Maury, C., 2009, Self-propagating β-sheet polypeptide structures as prebiotic informational molecular entities: the amyloid world, *Origins of Life and Evolution of Biospheres*, 39:141-150.

Mirny, L., E. Shakhnovich, 2001, Protein folding theory: from lattice to all-atom models, *Annual Reviews of Biophysics and Biomolecular Structure*, 30:361-396.

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.

Sawaya, M., S. Sambashivan, R. Nelson, M. Ivanova et al., 2007, Atomic structures of amyloid corss-β splines reveal varied steric zippers, *Nature*, 447:453-457.

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.

Schnabel, J., 2010, Secrets of the shaking palsy, *Nature*, 466:August 26, s2-s5.

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

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

Tlusty, T., 2007b, A relation between the multiplicity of the second eigenvalue of a graph Laplacian, Courant's nodal line theorem and the substantial dimension of tight polyhedral surfaces, *Electrical Journal of Linear Algebra*, 16:315-324.

Tlusty, T., 2008a, Rate-distortion scenario for the emergence and evolution of noisy molecular codes, *Physical Review Letters*, 100:048101-048104.

Tlusty, T., 2008b, A simple model for the evolution of molecular codes driven by the interplay of accuracy, diversity and cost, *Physical Biology*, 5:016001.

Tlusty, T., 2008c, Casting polymer nets to optimize noisy molecular codes, *Proceedings of the National Academy of Sciences of America*, 105:8238-8243.

Tlusty, T., 2010a, A colorful origin for the genetic code: information theory, statistical mechanics and the emergence of molecular codes, *Physics of Life Reviews*, 2010;doi:10.1016/j.plrev.2010.06.002.

Tlusty, T., 2010b, Reply to comments, *Physics of Life Reviews*, 7:381-384.

Tycko, R., 2006, Molecular structure of amyloid fibrils: insights from solid-state NMR, *Quarterly Reviews of Biophysics*, 39:1-55.

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

Wallace, R., and R.G. Wallace, 2008, On the spectrum of prebiotic chemical systems: an information-theoretic treat-

ment of Eigen's paradox, *Origins of Life and Evolution of Biospheres*, 38:419-455.

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., 2007, Culture and inattentional blindness, *Journal of Theoretical Biology*, 245:378-390.

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

Wallace, R., 2010a, A rate distortion approach to protein symmetry, *BioSystems*, 101:97-108.

Wallace, R., 2010b, A scientific open season, *Physics of Life Reviews*, 7:377-378.

Wallace, R., 2010c, Extending the modern synthesis. In press, *Comptes Rendus Biologies.*

Wang, L., S. Maji, M. Sawaya, D. Eisenberg, and R. Reik, 2008, Bacterial inclusion bodies contain amyloid-like structures, *PLOSBiology*, 6:e195.

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/$08_0215.htm$.

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.

Yang, W., and M. Grubele, 2004, Folding λ-repressor at its speed limit, *Biophysical Journal*, 87:596-608.

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.

13 Appendix: 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).