'The Frozen Accident' as an Evolutionary Adaptation: A Rate Distortion Theory Perspective on the Dynamics and Symmetries of Genetic Coding Mechanisms

James F. Glazebrook
Department of Mathematics and Computer Science
Eastern Illinois University
600 Lincoln Avenue, Charleston IL 61920–3099, USA
E-mail: jfglazebrook@eiu.edu

Rodrick Wallace
Division of Epidemiology
The New York State Psychiatric Institute
Box 47, 1051 Riverside Drive, New York NY 10032, USA
E-mail: wallace@pi.cpmc.columbia.edu

Keywords: frozen accident, rate distortion function, protein folding, free energy density, spin glass, groupoid, Onsager relations, holonomy

Received: October 8, 2011

We survey some interpretations and related issues concerning 'the frozen accident' hypothesis proposed by Francis Crick and how it can be explained in terms of several natural mechanisms involving error-correction codes, spin glasses, symmetry breaking and the characteristic robustness of genetic networks. The approach to most of these questions involves using elements of Shannon's rate distortion theory incorporating a semantic system which is meaningful for the relevant alphabets and vocabulary implemented in transmission of the genetic code. We apply the fundamental homology between information source uncertainty with the free energy density of a thermodynamical system with respect to transcriptional regulators and the communication channels of sequence/structure in proteins. The collective outcome of these processes supports previous suggestions that 'the frozen accident' may in fact have been a temporal evolutionary adaptation.

Povzetek: Članek obravnava izvor genetskega kodiranja.

1 Introduction

Examining and predicting the geometric/topological structures of the genetic coding network is essential to understanding its (co)evolution as a complex communications system, employing a vocabulary of a given genetic code that determines the family of proteins encodable by the genes themselves. The architecture of this network developed from a coevolution of genes and of genetic structures that were progressively conditioned to shield against translation and replication errors. Crick's hypothesis [30, 31](surveyed in e.g. [4]), in broad terms, says that on reading the mRNA script, the coding strategy determines the amino acid sequence of the evolved proteins, as is the case for most organisms. So in a post-transitional phase any kind of alteration to the size of the code would have dire consequences owing to a global impact on proteins created by new amino acids subject to the likelihood of nonsensical messaging. Crick gave flexible rules for pairing the third base of the codon with the first base of the anticodon, to the extent that a single tRNA type would be able to recognize up to three codons. More complex protein structures arise when there is an enrichment and expansion of the vocabulary while any ambiguity in the code is minimized, so restricting the content of information. When the codon meaning is altered, the information selected would condition that codon to some advantage. In this way the 'freezing' was professed to be an outcome of such selective restrictions and this would put the brakes on further evolvability.

While over the years there has been much debate and challenge concerning these rules, and to establish a concrete mechanism for the companion 'wobble hypothesis', we outline here several scenarios from the point of view of coevolutionary rate distortion dynamics in graphs that represent 'robustness' while admitting 'meaningful' signalling paths which are susceptible to vocabulary enrichment, and furthermore, give rise to structure preserving patterns that evolve towards optimizing error-correction. These collective mechanisms can be formulated in the context of a spin-glass model (cf [12, 21, 25]), that incorporates the Onsager relations of statistical physics applied to networks of

mutating sequences and error-correction in the presence of rate distortion dynamics, then leading to phase transitions through which symmetry breaking occurs and hence causes a change in topological structure of the graph. These observations are supported by a number of relatively recent theoretical findings, and thus it seems reasonable to provide some of the necessary background material. Related are the approaches to evolutionary (population) biology employing Boltzmann statistics, Fisher and Kolmogorov diffusion equation methods, and stochastic evolution for which there is already a large amount written (see e.g. [78]).

A position often maintained is that evolution influences the emergence of the genetic code by selecting an amino acid map that is error-minimizing and the subsequent competition between organisms is determined by the overall capability of their respective codes. Following this line of thought, Tlusty [73, 74, 75, 76], implementing a topological graph-theoretic approach, has developed a model for the emergence of the genetic code as a supercritical phase transition occurring within noisy information channels as traced by maps between nucleotides and amino acids with error bounds in place. The proposed paradigm is that these processes are indeed 'cognitive' [80, 81, 82, 85] following the immunology/language perspective of Atlan and Cohen [6] (see also [26, 27]) that human and biological organizations at all scales are cognitive in so far that once patterns of threat and opportunity are perceived, these patterns are are compared with an internal image of the environment, and then a choice of responses from a vast repertoire of possibilities is initiated.

This present paper continues with this theme to establish one of several possible corollaries derived from [80, 82] by addressing the question of how coevolutionary robustness against errors, error-correction, and phase transitions modeled by the topological dynamics of graphs that can be represented by certain spin glass/error correcting structures that are susceptible to thermodynamic spontaneous symmetry breaking; these factors shed further light to explaining what exactly was the 'accident' that did occur. Such symmetry breaking of the genetic code has been considered in the context of Lie algebra representations in [10, 11, 46]. Our perspective using rate distortion dynamics, is that such a sequence of broken symmetries corresponds to phase transitions in the underlying error correcting networks through which the codon allocation to amino acids is mainly the outcome of error-correction minimization and efficiency (see [10] and references therein), a scenario that appears relevant to the approach of Ardell and Sella [4, 66, 67].

While on the mathematical-physical side of things, several explanations for 'freezing' and 'wobbling' can be given in terms of error-correction and the structural theory of Lie algebras, which we survey. A novel technique introduced here involves showing how the dynamics governing the underlying mechanisms can be represented in terms of a 'covariant differentiation' of the Shannon entropy along 'meaningful paths' embedded in a (genetic) coding graph that also includes a correlation with error-correction and folding rates. This operation over which the various 'directions' are taken [1] subsequently determines the *holonomy* of the system through an error-correction network–a broader scale geometric representation of transitional phases in which the broken symmetries may be expressed in terms of holonomy groups that collectively, via disjoint union, form a holonomy groupoid, a structure which in principle can be given explicitly.

2 'The Frozen Accident'– or Not Quite

We start by putting matters into perspective by surveying some basic observations. Recall that genes can be represented by molecular words written in terms of the nucleotide bases U(Uracil/Thymine), C(Cytosine), G(Guanine) and A(Adenine), whereas proteins are written in a language of 20 letters corresponding to the amino acids in which each of the latter is encoded by specific triplets of the basis members, known as *codons*, so connecting hereditary characteristics to vital units. In theory there are $64 = 4^3$ codons with the number of possible observables lying somewhere between 48 and 64 (see e.g. [50, 73]). However, it is claimed in [50] that the code mapping the 64 codons to the 20 amino acids is anything but random. There are at least 48 discernable codons but only 20 amino acids available (and 3 stop codons), so the code is degenerate in so far that several codons can represent the same amino acid. Entropy analysis [1, 55] reveals that the information content of a random protein structure can occupy $\log_2(20) \simeq 4.32$ bits of entropy per amino acid residue in a primary sequence.

In the presence of topological changes there would have been alterations of an excessive amount of (protein) structures, and those frequently observed tend to be the ones that have managed to remain intact as the structures became more complex. The 'wobble rules' assume that only 48 codons can be distinguished owing to the physiochemical limitations of the translational mechanism and the resulting codon graph converges to 20 amino acids. The question is: does a single sRNA molecule recognize several codons? The 'wobble' effect aside, there exist 64 distinguishable codons and the maximal number of amino acids increases to 25, which is not a dramatic amount by any means, though it has been a puzzling matter as to why evolution did freeze prior to improving the translational mechanism to single out all 64 codons. Once the meaning of a codon had changed, again, selectivity would apply that codon to a site for a new amino acid to serve to some advantage, or otherwise simply to replace it.

The traditional approach to producing more tRNAs

[1] A reader with some acquaintance with differential geometry will understand this as 'covariant differentiation over (or along) a vector field'– an operation specified by choice of 'connection'. This we implement on graphs in §6.

would have been to change the anticodons of existing ones, giving rise to a new class of amino acids proliferating across the code while systematically reshuffling a large number of codons in the process. To an extent the 'wobble hypothesis' concerns stereochemical limitations on the actual tRNA capacity to single-out codons [38]. In more basic terms, interfering with the genetic code would change the meaning of a codon, hence from our viewpoint, reducing the fidelity of information when the rate distortion estimate is violated (see §3.2).

As was recalled in the introduction, Crick's hypothesis had suggested that no new amino acids could arise without disrupting a large number of proteins, hence stalling evolution – a claim that has since been challenged from many fronts (see e.g. [4, 68]). A product of the coevolutionary dynamics gives rules for load minimization and diversification for regulating patterns of the code that were robust to both error and redundancy, the degrees of which are influenced by the code's topology that would have been alterable through sequences of stochastic fluctuations. Codons interchanged through error may subsequently be assigned to compatible amino acids so minimizing the possible detrimental effects. At the same time, an enrichment of the vocabulary provided a broader scope for the encoding of proteins [66, 67].

In [77] there is claimed a 'communality' and 'universality' to be established out of a tournament between a variety of innovative sharing protocols which may include several non-Darwinian mechanisms. Relative to time scales, the long-term reduces ambiguity, whereas in the short-term the code has to be fortified to tolerate a higher degree of ambiguity in assimilating new types of genes. More specifically [77]:

> A protein that is robust to translational errors *a fortiori* is also more tolerant to translation with a different code. Conversely, the less optimized the recipient code, the more error-tolerant its proteins, and therefore the less harmful the effect on the established genes of a code change in the direction of the donor code. This has the important consequence that in the initial stages of the genetic code evolution, when the diversification tendency of codes was strongest, HGT (horizontal gene transfer) was possible and must have been extensive despite the presence of many different codes ... Once the optimization of the genetic code is complete, there is no pressure to maintain compatibility. Therefore, the "freezing" of the universal genetic code could trigger the radiation of the underlying translational machineries...

We may reasonably assume that transmission errors eventually corrupt code patterns and those codes that can withstand and manipulate errors possess natural advantages over those that do not. In concluding differently to Crick's assertion, code-messaging evolution is perceived in [4] as producing structure preserving codes which have near optimal error-correcting properties, with the selection of mutations and translational error inducing a bias in the codon distribution to amino acids which in the long-term favors optimal error-correction patterns. Crick's claim of 'freezing' makes some sense because the errors themselves condition evolution to some sort of frozen state of an error-correcting code. Specifically, the claim is that an evolutionary constraint on messages with respect to selective pressures, may actually induce the error-correcting codes to evolve rather than to have erased them altogether. Thus, in this evolutionary context the allied and relevant mechanisms of protein synthesis, folding and mutations, provide suitable clues.

An underlying assumption proposed in [1] is that an organism's complexity reflects upon that of its genome and therefore has evolutionary consequences. So one may ask what actually is the information provided by DNA beyond a road map for the structure of an organism? The current perspective sees this as a blueprint for constructing an organism that can survive within its native environment and then pass on that information to its progeny (cf [33]). In this respect, an organism's DNA catalogs not only information concerning its structure, but to some extent information concerning its environment and the coevolution of its species as well. In keeping with this basic principle, one may propose an explanation of genomic complexity within the information-theoretic framework of Shannon's basic principles (see [1, 2] and references therein for related work). It is in this respect that the fundamental theorems of information transmission are sufficiently general to the extent that biological systems can sustain a Shannon-based coding scheme to facilitate the transmission of genomic information within a range of mechanisms, provided that semantics can be incorporated as a functional component (see §4.1 and cf [35]).

3 Encoding and Decoding

3.1 Basic genetic messaging

The transmission of genetic messaging follows a sequence starting from a source alphabet via a channel code to a target alphabet. The source messaging in the DNA alphabet is relayed to the encoding DNA alphabet to the mRNA alphabet with certain reciprocation. Leading on from mRNA messaging in the RNA alphabet is a channel to point mutation through which (genetic) noise may enter, thence a channel to decoding into which amino acylated tRNA and mischarged tRNA, with further genetic noise, enter via translation. Subsequent to decoding is the protein messaging in the target protein alphabet. This is a basic sequence of events that is schematically represented in [92, Figure 2].

At the same time, evidence suggests that primordial tRNAS along with their various companion types and the overall translation mechanism have coevolved in some de-

gree of compliance with the genetic code, rather than the reverse, and possibly the assignment of amino acids to nucleotides may have been pre-translational. If the code were to be pre-translational in nature, then how it was originally imprinted within tRNAs could be researched in the quest of the so-called 'RNA world idea' [63, 72].

3.2 The rate distortion function

For the sake of self-containment in this paper, we next briefly recall some elementary facts from the Shannon theory. As it is commonly understood, distortion arises when there is a fast relay of information through some channel which exceeds the latter's capacity. One of the guiding principles asserts that in order to reproduce a message transmitted from a source to a receiver, it is necessary to know what sort of information should be transmitted, and how. These facts along with specifying the nature of the communicating channel are essential ingredients for engineering a reliable encoding/decoding system. Following [14] we briefly recall some of the basic operations.

Source encoder: We may consider some output $x(t)$ emanating from the source as projected to a finite set of preselected images; namely, the space of possible source outputs is partitioned into a set of *equivalence classes*, and the source encoder informs the channel encoder of that class containing the particular source output observed. Once the channel encoder is informed that the source output belongs to say, the m-th equivalence class, it transforms the corresponding waveform $\tilde{x}_m(t)$ across the channel. These equivalence classes as schematically represented by a graph (network), are manifestly the main computational procedures as described in this paper.

Source decoder: Within the system is a cascade of a channel encoder and a source decoder. The channel decoder receives a waveform $\tilde{y}(t)$ of a corresponding function $y(t)$ over some time interval and decides upon the nature of the message as transmitted. Then it sends its approximation m' of the message number to the source decoder which in turn creates $y_{m'}(t)$ to register the system's estimate of $x(t)$ over that time interval. Initially, we may think of $x(t)$ and $y(t)$ as 'waveforms', but in our case, we consider these as consisting of a language with its own intrinsic grammar/syntax, as well as 'meaning' – to be made more specific in §4.1. Analogous considerations apply to the channel signals $\tilde{x}(t)$ and $\tilde{y}(t)$.

One of Shannon's notable results was that a communication system can be designed such that it achieves a level of fidelity D once the *rate distortion* $R(D) \leq C$, where C denotes the channel capacity. Putting it another way, if the receiver can tolerate an average amount of distortion D, the rate distortion $R(D)$ is the effective rate at which the source can relay information with that level of tolerance, and the estimate $R(D) \leq C$ is a necessary condition for effective communication. More specifically, $R(D)$ can be defined in terms of *average mutual information* as follows. Firstly, for k, j running over a suitable alphabet, let us write a given conditional probability assignment as $Q(k|j)$ such that in the usual way, we have an associated joint distribution $P(j,k) = P(j)Q(k|j)$. We express *the average distortion* as

$$d(Q) = \sum_{j,k} P(j)Q(k|j)\, d(j,k), \tag{3.1}$$

where $d(\ ,\ )$ denotes the distortion measure. A conditional probability assignment $Q(k|j)$ is said to be *D–admissible* if and only if $d(Q) \leq D$. The set of all D–admissible conditional probability assignments we denote by

$$Q_D = \{Q(k|j) : d(Q) \leq D\}. \tag{3.2}$$

Along with an average distortion $d(Q)$, we also have an *average mutual information*

$$I(Q) = \sum_{j,k} P(j)Q(k|j) \log \big[\frac{Q(k|j)}{Q(k)}\big]. \tag{3.3}$$

Then for fixed D, the rate distortion function is defined as

$$R(D) = \min_{Q \in Q_D} I(Q). \tag{3.4}$$

The rate at which a source produces information subject to insisting upon perfect reproduction, is the *source entropy* H. Given a distortion measure such that perfect reproduction is assigned zero distortion, then we have $R(0) = H$. As D increases, $R(D)$ becomes a monotonically decreasing (convex) function which eventually is zero, typically at a maximum value for D (see [14, Ch. 1]). This is a very basic observation, and typically in rate distortion theory one seeks a reduction of H by either slowing down the emission of coding, or encoding the relevant languages at a lower rate. In view of Shannon's theorem, as long as $H < C$, there will be suitable fidelity in transmission. In the case of genetic coding considered here, conditions of *discrete memoryless information source* (DMI) and *discrete memoryless channels*(DMC) [57, 92] are usually assumed, but in any event, how well a communicating system can evolve in order to satisfy such an estimate is a common problem for communications engineering since in practice the source rate may be corrupted due to low memory and coding congestion; for protein folding and mutations; references [2, 32, 55, 73, 74, 80, 81] address such issues.

3.3 The Groupoid Free Energy Density

Recall that for a thermodynamic state of a given system at fixed temperature T with energy E and entropy S, the *free energy density* F is defined to be

$$F = E - TS. \tag{3.5}$$

In the Hamiltonian formulism one takes the volume V and the partition function $Z(K)$ derived from the system's

Hamiltonian at inverse temperature K [51, 52]. The free energy density is then defined to be

$$F[K] = \lim_{V \longrightarrow \infty} -\frac{1}{K}\frac{\log[Z(K,V)]}{V} = \lim_{V \longrightarrow \infty} \frac{\log[\widehat{Z}(K,V)]}{V}, \text{ where } \widehat{Z} = Z^{-\frac{1}{K}}. \tag{3.6}$$

At this stage we introduce the *groupoid* concept (generalizing the algebraic concept of a 'group') in relationship to *equivalence classes* which can be based upon a network with concatenation of edges, as explained in Appendix 8.1 (see also [40, 41]). Thus, consider an information source H_{G_α} over a corresponding groupoid G_α; heuristically, we can consider H as parametrized by G_α. The probability of H_{G_α} is given by:

$$P(H_{\mathsf{G}_\alpha}) = \frac{\exp[-H_{\mathsf{G}_\alpha}K]}{\sum_\beta \exp[-H_{\mathsf{G}_\beta}K]}, \tag{3.7}$$

where the normalizing sum is over all possible subgroupoids of the largest available symmetry groupoid. On setting

$$Z_{\mathsf{G}} = \sum_\alpha \exp[-H_{\mathsf{G}_\alpha}], \tag{3.8}$$

the groupoid free energy density (GFE) of the system F_{G} *at inverse normalized equivalent temperature* K is then defined as

$$F_{\mathsf{G}}[K] = -\frac{1}{K}\log[Z_{\mathsf{G}}(K)]. \tag{3.9}$$

With each such groupoid G_α we can associate a dual information source H_{G_α}. We recall the rate distortion function between the message sent by the cognitive process and the observed impact, while noting that both H_{G_α} and $R(D)$ may be considered as free energy density measures. In a sense, $R(D)$ constitutes a sort of 'thermal bath' for the process of cognition. Then the probability of the dual information source can be expressed by

$$P(H_{\mathsf{G}_\alpha}) = \frac{\exp[-H_{\mathsf{G}_\alpha}/\kappa R(D)\tau]}{\sum_\beta \exp[-H_{\mathsf{G}_\beta}/\kappa R(D)\tau]}, \tag{3.10}$$

where κ denotes a suitable dimensionless constant characteristic of the system in the context of a fixed 'machine response time' τ. Associated with (3.10) is a *free energy Morse Function*

$$F_R = -\lambda R(D)\log[\sum_{\alpha=1}^{n} \exp[-H_\alpha/\lambda R(D)]], \tag{3.11}$$

whose critical point behavior determines certain topological characteristics of an underlying manifold that can be expressed in terms of its Morse-theoretic indices [56, 58]. In each case the sum is over all possible subgroupoids of the largest available symmetry groupoid (see Appendix 8.1). Accordingly, the term $R(D)\kappa$ in (3.10) represents a rate distortion energy, in this case, a kind of temperature analog. In the context of a fixed response time τ, a decline in $R(D)$ (on increase in average distortion), acts to 'lower the machine temperature' and thus driving it to more simple, albeit less enriched signalling. Observe that if a range over all possible α is taken, the groupoids G_α and corresponding relationships such as (3.10), create an even larger picture which reveals the structure of a *groupoid atlas* [9], a concept that has been applied to several descriptive cognitive mechanisms as we have demonstrated in [40, 41, 42].

3.4 Phase transition and symmetry breaking

The relationship between phase transitions in physical systems and topological changes has become a central topic of research across a broad range of subdisciplines. One can see that phase transitions in physical systems are ubiquitous, following Landau's group symmetry shifting arguments [52, 59]. Higher temperatures enable higher system symmetries, and as temperature changes, punctuated shifts to different symmetry states occur in characteristic manners. The claim in [37] is that the standard way of studying phase transitions in a physical system is to consider how the empirical values of thermodynamic states, vary with temperature, volume, or an external field, and then to associate the experimentally observed discontinuities at a phase transition to the occurrence of a singularity. In such a case analyticity may fail in the mathematical sense, though it remains to be seen whether this is the ultimate level of an analytic understanding of such transitional phenomena, or if indeed some reduction to a more basic level is possible. It is observed that non-analyticity is the 'shadow' of a more fundamental phenomenon occurring in a given model space: *a topology change*, and that the latter is a *necessary* condition for a phase transition to occur. Such topology changes can be studied within the framework of Morse theoretic influenced topological structures such as the case, say, for certain handle-body decompositions [56], an essential observation that may be consequential for protein functions (cf [82]). Note however, that the converse of the main result of [37] does not hold, thus ruling out a one-to-one correspondence between phase transitions and topology changes. An open problem is that of *sufficiency* conditions; that is, to determine which kinds of topology changes can influence a phase transition, and how this might be achieved. There are other approaches such as demonstrated in relatively straightforward models, where as in [64], a fuzzy clustering system based of annealing through a probabilistic process leads to phase transitions with critical (non-zero) vectors for the free energy at each temperature.

Extension of such transitional arguments in terms of rate distortion and metabolic measures appear direct, particularly in the setting of the groupoids constructed by the disjoint union of the homology groups representing the different coding topologies identified in [73] (see also [80]). To clarify matters, let us recall that in many thermodynamic systems, the associated Hamiltonian may be invariant under a symmetry transformation due to certain parameter

changes, in contrast to the lowest energy state which is not. In subsequent phase transitions the overall symmetry is lost (*spontaneous symmetry breaking*) and consequently, lower temperature states will admit lower symmetries, and due to the randomization of higher temperatures, the higher states will become more accessible to the system as a result of their modified symmetries and energy levels [52]. In the informational context of error-correction, we will need to turn to the fundamental homology between the Shannon entropy and the free energy density of the system as outlined in §4.1.

This scenario becomes more apparent when we look at the symmetries of the genetic code and how these are broken (cf. [71]). For instance, in [46] it is recalled from [15] that the computation of at least 10^{71} to 10^{84} possible genetic codes entails permuting the 64 codons and distributing them over 20 amino acids. By considering those Lie algebras admitting 64 dimensional irreducible representations, [10, 11, 46] initiate a chain of sub-representations commencing from the Lie algebra $\mathfrak{sp}(6)$, and postulate a sequence of symmetry breaking in accordance with that chain:

$$\begin{aligned}\mathfrak{sp}(6) &\supset \mathfrak{sp}(4)\oplus\mathfrak{su}(2)\\ &\supset \mathfrak{su}(2)\oplus\mathfrak{su}(2)\oplus\mathfrak{su}(2)\\ &\supset \mathfrak{su}(2)\oplus\mathfrak{u}(1)\oplus\mathfrak{su}(2)\\ &\supset \mathfrak{su}(2)\oplus\mathfrak{u}(1)\oplus\mathfrak{u}(1).\end{aligned} \tag{3.12}$$

At any stage the number of representations occurring corresponds to the number of amino acids that were then incorporated into the code and those currently observed are the net outcome of broken symmetries. In this analysis, four amino acids (phenylalanine, serine, argine and cysterine) seemingly do not divide under the $\mathrm{U}(1)$(circle)-action. If they had subdivided they would have created a 'symmetry perfect code' with 26 amino acids (hence a redundancy of 6) and a stop code (see [46, Figure 1]). Such a claim may be compared with the combinatorial-geometric arguments based on the topology of codon space in [73] (see also §6.1) suggesting that further evolutionary measures may expand the code's expression from 20 to possibly 25 amino acids.

The observations of [10, 11] reflect back upon an earlier claim of [48] that the 'freezing' of the code would have been the result of partial symmetry breaking achieved by the aforementioned parameter choices in the Hamiltonian. The work of [10, 11] differs in its approach by opting for codon-anticodon pairings in place of codon-amino acid assignments and then applying combinatorial-branching techniques commencing from the Lie algebra $\mathfrak{sl}(6,1)$. Besides identifying possible 'wobble-effects' due to reshuffling through combinatorial symmetries, they investigate the structure of eukaryotic and vertebrate mitochondrial codes along branching chains and introduce a $\mathbb{Z}_2$-grading on codon space (just as there is a grading into bosonic and fermionic types in quantum mechanics) thus extending matters towards representations of super Lie algebras. Along with these codes are variants such as the metabacteria and chloroplast codes with exchange symmetries and branching rules for which such patent intricacy may eventually necessitate using groupoid techniques.

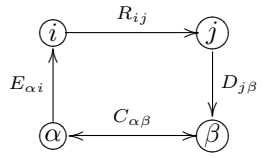
An alternative approach to Lie algebra representations due to [47] is to consider representations on hypercubes as based on Gray coding structures (for a survey of the latter in genetic error-correction, see [45]). Already some known group structures show up here for various assortments of codon doublets, and since sub-symmetries of these representations involve cubical methods, patterns of groupoid symmetries can be expected to be appear. Thus we approach increasingly complex situations involving *groupoid representations* (see e.g. [18]) and *groupoid symmetry breaking*, techniques that can be computationally highly non-trivial, since even for relatively straightforward symmetries such as those appearing in certain 'windmill patterns', constraints do apply in order to facilitate current programming capabilities [39]. Other questions may arise, such as the possibility of breaking 'mirror symmetry' states in the genetic code caused by biochemical perturbations of chiral fields at the molecular level [8].

3.5 Amino acid encoding–codon decoding and error load

In order for free energy and error load to fit into the picture, we follow part of the framework of error-correction network analysis of [73, 74] (cf [66]). We take an amino acid α to be encoded by a unique codon j represented in the encoder matrix $[E_{\alpha j}]$, satisfying $\sum_j E_{\alpha j}=1$, and similarly, the decoder matrix $[D_{j\beta}]$, satisfying $\sum_\beta D_{j\beta}=1$, means that each codon is translated into a unique amino acid β, given a number $\mathcal{N}_c$ of protein chains for c codons. Next we set

$$\begin{aligned}R_{ij} = P(&\text{the probability that codon } i \text{ may be}\\ &\text{read correctly as or misread as } j),\end{aligned} \tag{3.13}$$

and then let $[R_{ij}]$ denote *the reading matrix* and $C_{\alpha\beta}$ *the chemical distance* between the original amino acid α and the one that is read as β. As adapted from [73, Figure 2] the passage of encoding/decoding then follows as:



(3.14)

On setting $P_\alpha = P(\text{amino acid } \alpha \text{ is required})$, the *error load* H_{ED} (the average distortion in an $R(D)$ problem) of the map specified by *encoding/decoding* can be expressed in terms of paths $P_{\alpha ij\beta}$, specifically by

$$\begin{aligned}H_{\mathrm{ED}} &= \sum_{\alpha\to i\to j\to\beta} P_{\alpha ij\beta}C_{\alpha\beta}\\ &= \sum_{\alpha,i,j,\beta} P_\alpha E_{\alpha i}R_{ij}D_{j\beta}C_{\alpha\beta}.\end{aligned} \tag{3.15}$$

This leads to a 'take-over' probability given by $P_{\rm ED} \sim \exp(-H_{\rm ED}T^{-1})$ and to the *average error load* $\langle H\rangle$ as follows. If we take S to denote entropy due to random drift, and T to be inversely proportional to average error size (the strength of the random drift relative to the selection force that pushes towards maximization), then this probability can be seen to minimize a functional analogous to the Helmholtz free energy F in terms of the average error load $\langle H\rangle$ as in (3.5):

$$\begin{aligned} F &= \langle H\rangle - TS \\ &= \sum_{\rm ED} H_{\rm ED}P_{\rm ED} + T\sum_{\rm ED} P_{\rm ED}\ln P_{\rm ED}, \end{aligned} \tag{3.16}$$

which effectively averages out the difference between the genetic message relayed by a codon statement and that which is actually expressed by the genetic/epigenetic translation machinery itself.

4 Meaningful Paths, Robustness and Error Correction

4.1 Meaningful paths

We now specify our observations in a more general context. Suppose we consider a pattern of signalling input $\mathbf{S}_i$ describing the state of the protein with initial codon stream $\mathbf{S}_0$ to be mixed in an unspecified but systematic algorithmic manner with a pattern of an otherwise unspecified ongoing activity, including cellular, epigenetic and environmental signals $\mathbf{W}_i$ to create a path of combined signals $x = (a_0, a_1, \ldots, a_n, \ldots)$. Each a_k thus represents some functional composition of internal and external signals in an iterative form according to which

$$\mathbf{S}_{i+1} = f([\mathbf{S}_i, \mathbf{W}_i]) = f(a_i), \tag{4.1}$$

for some unspecified function f. Comparing this with the situation in §4.2, the above $\mathbf{S}$ would be a vector, $\mathbf{W}$ a matrix, and f a product of their function at some time stage i. This path is fed into a highly nonlinear, but otherwise similarly unspecified, decision oscillator h which generates an output $h(x)$ that is an element of one of two disjoint sets B_0 and B_1 of possible system responses, as follows. Let

$$\begin{aligned} B_0 &\equiv b_0, \ldots, b_k, \\ B_1 &\equiv b_{k+1}, \ldots, b_m. \end{aligned} \tag{4.2}$$

Then:

(1) assume a graded response, supposing that if

$$h(x) \in B_0, \tag{4.3}$$

the pattern is not recognized, and

(2) if

$$h(x) \in B_1, \tag{4.4}$$

the pattern is recognized, and some action $b_j, k+1 \leq j \leq m$, takes place.

Expecting the coding signals to filtered appropriately (cf [4]), we can further assume that B_0 and B_1 admit countable filtrations of the sort:

$$\begin{aligned} B_0 = B_0^0 \subseteq B_0^1 \subseteq B_0^2 \subseteq \cdots \\ B_1 = B_1^0 \subseteq B_1^1 \subseteq B_1^2 \subseteq \cdots \end{aligned} \tag{4.5}$$

where at level j we have set $B_0^j \equiv b_0^j, \ldots, b_k^j$, and $B_1^j \equiv b_{k+1}^j, \ldots, b_m^j$. Note that these oscillators may be influenced by 'forcing' when a signal is subjected to some impulse such that its frequency, and hence the response, adjusts accordingly with respect to an applied impulse. More familiar oscillating physical systems may react accordingly by exhibiting beats and resonance, for instance.

The principal objects of formal interest are paths x which, through information flow, trigger patterns of recognition-and-response. That is, given a fixed initial state $a_0 = [\mathbf{S}_0, \mathbf{W}_0]$, we examine all possible subsequent paths x beginning with a_0 and leading to the event $h(x) \in B_1$. Thus $h(a_0, \ldots, a_j) \in B_0$ for all $0 < j < m$, but $h(a_0, \ldots, a_m) \in B_1$. We can view B_1 then as the set of final possible states $\mathbf{S}_f \cup \{\mathbf{S}_{\rm path}\}$ that includes both the final physical states and the set of all possible pathological conformations (see [80, Figure 3]).

For each positive integer n, let $N(n)$ be the number of high probability grammatical/syntactical paths of length n which begin with some particular a_0, and further leading to the condition $h(x) \in B_1$. These are paths of combined signals as above, that are structured to some language. For short, we 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 $h(x) \in B_1$.

One critical assumption which permits an inference on the necessary conditions constrained by the asymptotic limit theorems of information theory, is that the entropy, as defined by the finite limit

$$H \equiv \lim_{n\longrightarrow\infty} \frac{\log[N(n)]}{n}, \tag{4.6}$$

both exists and is independent of the path x. The rate distortion principle applies as follows [79]: *the restriction to meaningful sequences of symbols increases the rate at which information can be transmitted with arbitrary small error, and that the grammar/syntax of the path can be associated with a dual information source*.

Besides the DMI and DMC properties introduced in §3.2, we may also assume a typical information source $\mathbf{X}$ to be 'adiabatic', 'piece-wise stationary' and 'ergodic' (APSE), and that the relevant systems engaging in a biocognitive process is describable as such. Specifically, the essence of 'adiabatic' is that given the information source is parametrized according to some appropriate scheme, then within continuous 'pieces' of that parametrization, alterations in parameter values occur slowly enough so that the information source $\mathbf{X}$ remains as close to stationary and ergodic as necessary in order to implement the specific

limit theorems. In this way, 'structure' is subsumed within the sequential grammar and syntax of the dual information source, rather than within the sets of developmental paths as considered in [85].

In view of (4.6), the Shannon entropy of $\mathbf{X}$ can be stated more specifically by (see e.g. [5, 14, 29, 49]):

$$H[\mathbf{X}] = \lim_{n \longrightarrow \infty} \frac{\log[N(n)]}{n}. \tag{4.7}$$

With respect to e.g. the robustness criteria of §4.2, the time dependent information sources $\mathbf{X}_i(t)$ are identified with the i-th component of the expressional pattern $\mathbf{S}(t)$; that is, we assign $\mathbf{X}_i(t) \mapsto \mathbf{S}_i(t)$, where as before $\mathbf{S}_i(t) = f(a_{i-1})$.

Recalling how the information source uncertainty was defined as in equation (4.6), an essential observation is a *fundamental homology* with the free energy density of a thermodynamical system such as that displayed in equation (3.6). Such a homology arises from Feynman's observations [36] reflecting in part on Bennett's work [13] where this homology is effectively an identity, at least for very simple systems. From a more general perspective, [36] postulates the information contained in a message as proportional to the amount of free energy density needed to erase it. This simply amounts to the fact that computing in any form takes work and the more complicated a coding or signalling process so measured by its source uncertainty, the greater its energy consumption. Putting it another way, the less information available to us concerning an event the higher its entropy, and information retrieved is not without a cost in expenditure (of energy), where 'cost' is interpreted as the necessary number of bits needed to encode a message (the thermodynamic minimum of energy in terms of bits of information is $k_B T \log_2 e$ erg/bit, or $= k_B T$ erg/nat). So the efficiency in an information system essentially happens when there is the minimum amount of energy expended in retrieving information. Specifically, if F is taken to denote the free energy, then setting Λ equal to the minimum number of nats/sec, the efficiency of the system is given by $\eta = k_B T F^{-1} \Lambda$ (see e.g. [14]).

4.2 Transcriptional regulators and robustness

There are certain evolutionary innovations resulting from an interplay of mutations and natural selections whereby, in a descriptive sense, a genotype corresponds to a regulatory network with a given topology and a phenotype to that of a steady state genetic pattern. This mechanism is constrained by certain conditions requiring processes to sustain a degree of robustness, meaning here a resilience towards environmental perturbations and thermodynamic effects, while at the same time admitting some 'diversity' in the process of messaging reception. Such a function of evolution and environment is to ensure that proteins can continue their catalyzing role in the presence of amino acid mutations, that the regulatory networks can continue to function in a noisy environment, and that embryos can develop normally in the presence of such perturbations. In any case, these regulatory networks, (protein) synthesis and the mutational operations can be seen as part and parcel with the question of folding (misfolding), while observing that error-minimization permits the appropriate codon allocation to amino acids through sequences of broken symmetries in terms of tRNA mutations (see [10, 11]).

Thinking back to the context of §4.1, we next turn to an analogous, but closely related sequence of N transcriptional regulators represented by their expressional patterns $\mathbf{S}(t) = (\mathbf{S}_1(t), \mathbf{S}_2(t), \ldots, \mathbf{S}_N(t))$, in network form, at some time t, that can influence expressions between themselves via cross-regulatory and auto-regulatory interactions as expressed by a matrix $W = [w_{ij}]$, where w_{ij} represents a signaled regulatory influence w_{ij} : gene $i \Rightarrow$ gene j, given the rules (1) $w_{ij} > 0$, means activating, (2) $w_{ij} < 0$, repressing, and (3) $w_{ij} = 0$, absence.

In [25] such regulatory interactions describe the expressional state of the network $\mathbf{S}(t)$ akin to a typical spin-glass model [21, 69, 91](see also Appendix 10), as specified by

$$\mathbf{S}_i(t + \tau) = \sigma\big[\sum_{j=1}^{N} w_{ij} \mathbf{S}_j(t)\big], \tag{4.8}$$

where τ is a constant and $\sigma(\ )$ is a sigmoidal function $\sigma : \mathbf{S}(t) \longrightarrow (-1, 1)$. For instance, with strong cooperation we may have $\sigma = \text{sgn}$, giving $\mathbf{S}_i = \pm 1$. Here $\mathbf{S}(t)$ can be taken as an *incoming input*, mixed in a systematic way relative to $W = [w_{ij}]$, to create a path of combined signals $x = (a_0, a_1, \ldots, a_n, \ldots)$ as to be seen in §4.1, homologous to the sequence $\mathbf{S}(t + \Delta t)$, with $n = t(\Delta t)^{-1}$, where on recalling expression (4.1), we set $\mathbf{S}_{i+1} = f([\mathbf{S}_i, \mathbf{W}_i]) = f(a_i)$. Accordingly, the structure becomes as much of a function of the sequential grammar and syntax of the dual information source as it is for the cross-sectional intervals of the space of the $W = [w_{ij}]$ (see [87]). Typically, one would denote by $\mathbf{S}(0)$ an initial state and by $\mathbf{S}_\infty$ a stable equilibrium state, with a distance measure $\mathcal{D}$ for graph topologies W, W' taken to be

$$\mathcal{D}(W, W') = \frac{1}{2M_+} \sum_{i,j} |\text{sgn}(w_{ij}) - \text{sgn}(w'_{ij})|, \tag{4.9}$$

where M_+ denotes the maximum number of regulatory interactions.

In essence this construction reveals that genotype space, for instance, can be traversed in small increments without changing the phenotype which has evolutionary significance for genetic patterns: randomly selected pairs of networks of the same phenotype may have very different structure and may be subject to varying selective pressures. One may imagine that a large overall 'diameter' of the network may be a critical feature for diversity of phenotype, and because some lengthy travel across the graph may be necessary to find all new phenotypes [25], a distance measure of two phenotypes $\mathbf{S}_\infty, \mathbf{S}'_\infty$ is given by the Hamming

distance d_H in the form

$$d_H = d_H(\mathbf{S}_\infty(j), \mathbf{S}'_\infty(j)) \\ = 1 - \sum_j \frac{\delta}{N}[\mathbf{S}_\infty(j) - \mathbf{S}'_\infty(j)],\ 0 \le d_H \le 1, \tag{4.10}$$

where Kronecker $\delta = 1$ should both arguments be equal, and $\delta = 0$ otherwise. Note that for such Hamming codes it is a basic fact that decoding all patterns of length $\le k$ is equivalent to $(d_H)_{\min} \ge 2k + 1$ (see e.g. [57, 92]).

Related is how, in the statistical mechanics formulation, genetic algorithms based on spin glass models can reveal optimal selectivity as increasing with evolution. In [61] it is shown how selecting those solutions that are at a higher level of fitness, can be paired (through a crossover operation say) and then tested. This is performed iteratively through an algorithm up to the point where there is no further improvement in the examined population. Using spin glass states, [61] apply a chain as represented by vectors of the spins $\sigma^{(\alpha)}$ (where $\alpha = 1, \ldots, P$) indexed by different members of the population; this spin vector is then implemented in the genetic algorithm. In such a case new spins $\tau_i^\alpha = \sigma_i^\alpha \sigma_{i+1}^\alpha$ are created. Selectivity on the basis of mutation and crossover follows from the energy levels of the Ising spin glass (which is described later in Appendix 10).

5 Rate Distortion Coevolutionary Dynamics

5.1 The basic equations

Understanding the time dynamics of cognitive systems away from phase transition critical points thus requires a phenomenology similar to the thermodynamic Onsager relations. If the dual source uncertainty of a cognitive process is parametrized by some vector of quantities $\mathbf{K} \equiv (K_1, \ldots, K_m)$, then in view of the analogy with nonequilibrium thermodynamics, the gradients in the K_j of the *disorder*, defined as

$$S \equiv H(\mathbf{K}) - \sum_{j=1}^{m} K_j\, \partial H/\partial K_j, \tag{5.1}$$

are of central interest. Note that equation (5.1) is analogous to the definition of entropy in terms of the free energy density of a physical system, as suggested by the homology between the latter and the information source uncertainty. Pursuing the homology further, the generalized Onsager relations defining temporal dynamics become

$$dK_j/dt = \sum_i L_{ji}\, \partial S/\partial K_i, \tag{5.2}$$

where the kinetic coefficients L_{ji} are, in first order, constants interpreted as reflecting the nature of the underlying cognitive phenomena (without requirement of the symmetry condition $L_{ij} = L_{ji}$). The partial derivatives $\partial S/\partial K$ are analogous to thermodynamic forces in a chemical system, and may be subject to override by external physiological driving mechanisms as shown in [79, 88] along with further extensions of these dynamical procedures.

Induced by the fundamental homology between the Shannon entropy and free energy density, the rate distortion $R(D)$ follows a homologous path relation to the latter, thus suggesting that the dynamics of any bio-cognitive module interacting in characteristic real–time τ, will be constrained by the system as described in terms of $R(D)$. This can be seen more generally [85, 86] by producing a vector–valued function $R(\mathbf{Q})$ where in the vector $\mathbf{Q} = (Q_1, \ldots, Q_k)$, the first component is defined to be the average distortion, and then (cf (5.1)), we have

$$S_R \equiv R(\mathbf{Q}) - \sum_{i=1}^{m} Q_i\, \partial R/\partial Q_i, \tag{5.3}$$

which leads to the deterministic and stochastic systems of equations analogous to the Onsager relations of nonequilibrium thermodynamics

$$dQ_j/dt = \sum_i L_{ji}\, \partial S_R/\partial Q_i, \tag{5.4}$$

together with

$$dQ_t^j = L^j(Q_1, \ldots, Q_k, t)\, dt \\ + \sum_i \sigma^{ji}(Q_1, \ldots, Q_k, t)\, dB_t^i, \tag{5.5}$$

where the dB_t^i represents often highly structured stochastic noise whose properties may be described in terms of Brownian motion and quadratic variation (see e.g. [60]).

5.2 The phenomenological Onsager relations

Here we turn to different developmental subprocesses of gene expression characterized by information sources H_m interacting via chemical or other types of signals, and assume that different processes become each other's principal environments. This is a working hypothesis within a broad coevolutionary context that underscores the cognitive element. Let

$$H_m = H_m(K_1, \ldots, K_s, \ldots, H_j, \ldots), \tag{5.6}$$

where the K_s represent other relevant parameters, and $j \ne m$. We regard the dynamics of this system as driven by a recursive network of stochastic differential equations. Letting the K_j and H_m all be represented as parameters Q_j (with the caveat that H_m does not depend on itself), we follow the generalized Onsager formulation of [85] in terms of the equation

$$S^m = H_m - \sum_i Q_i\, \partial H_m/\partial Q_i, \tag{5.7}$$

to obtain a recursive system of *phenomenological Onsager relations*, in terms of a system of stochastic differential equations

$$dQ_t^j = \sum_i [L_{ji}(t, \ldots, \partial S^m/\partial Q^i, \ldots)\, dt + \sigma_{ji}(t, \ldots, \partial S^m/\partial Q^i, \ldots)\, dB_t^i], \tag{5.8}$$

in which, for ease of notation, both the terms H_j and the external K_j's are expressed by the same symbol Q_j. As m ranges over the H_m, we could allow different kinds of 'noise' dB_t^i having particular forms of quadratic variation which may represent a projection of environmental factors within the scope of what may be viewed as *a rate distortion manifold* [41]. The noise factor is significant in view of the findings of [7] where it was observed that perturbations of the network parameters inducing stochastic fluctuations in the molecular patterns, may in turn influence regulatory mechanisms, and in a similar way to how the presence of stochastic resonance may amplify certain signals, noise-spectral measurements may then uncover further mechanisms which could be potentially beneficial to the code's evolution.

We remark that equation (5.8) can be generalized somewhat [85] with respect to crosstalk, its distortion, the inherent time constants of the various bio-cognitive modules, and in particular, the overall available free energy density. As shown in [42], analysis of the rate distortion dynamics on a case-by-case basis, motivates integration to a multidimensional Itô process as given by

$$Q_t^\alpha = Q_0^\alpha + \sum_{\beta=\{ij\}} [\int_0^t L_\beta(s, \ldots, \partial S_R^\beta/\partial Q^\alpha, \ldots)\, ds + \int_0^t \sigma_\beta(s, \ldots, \partial S_R^\beta/\partial Q^\alpha, \ldots)\, dB_s^\beta], \tag{5.9}$$

and this in turn leads to a stochastic flow on a suitable topological manifold which in this present context could serve as a more general model for the codon space. In fact, such a flow property had already been observed in [73], namely, that the standard genetic code and it variants evolve as a flow within the codon space. However, given that 'freezing' of some sort is likely to re-occur in the quest for optimal error-correction, we expect such a flow to be stalled at certain time intervals, thus creating singularities in the flow in a dynamical systems sense (an analytic technicality to be finessed here).

5.3 A metric on a space of languages

Let us note that equations (5.1) and (5.2) can be derived in a simple parameter-free covariant manner which relies on the underlying topology of the information source space that is implicit to the processes as envisaged. Different bio-cognitive phenomena have, according to our development, dual information sources, and we are interested in the local properties of the system near a particular reference state. We impose a topology on the system, so that near to a particular language A dual to an underlying bio-cognitive process, there is an open set U of closely similar languages $\hat{A}$, such that A and $\hat{A}$ are subsets of U.

Since the information sources dual to the processes are similar, for all pairs of languages $A, \hat{A}$ in U within a given embedding alphabet, we define a metric on the latter by

$$\mathcal{M}(A, \hat{A}) = |\lim \frac{\int_{A,\hat{A}} d(Ax, \hat{A}x)}{\int_{A,A} d(Ax, A\hat{x})} - 1|, \tag{5.10}$$

with respect to a distortion measure $d(Ax, \hat{A}x)$, and apply standard integration arguments over the high probability paths, where the usual metric properties apply, as in e.g.[22]. In the context of [4], we may see such a metric as derived from an informational driven physico-chemical distance function with respect to the analogous A and $\hat{A}$ coding. Also, since H and $\mathcal{M}$ are both scalars, a covariant derivative can be defined directly as

$$dH/d\mathcal{M} = \lim_{\hat{A} \longrightarrow A} \frac{H(A) - H(\hat{A})}{\mathcal{M}(A, \hat{A})}, \tag{5.11}$$

where $H(A)$ is the source uncertainty of language A.

A relatively straightforward case is the following. Suppose the system is set in some reference configuration A_0. To obtain the unperturbed dynamics of that state, impose a Legendre transform using this derivative, defining another scalar

$$S \equiv H - \mathcal{M} dH/d\mathcal{M}. \tag{5.12}$$

The simplest possible Onsager relation – here seen as an empirical, fitted, equation like a regression model, becomes

$$d\mathcal{M}/dt = L dS/d\mathcal{M}, \tag{5.13}$$

where t is the time and $dS/d\mathcal{M}$ represents an analog to the thermodynamic force in a chemical system (cf [14, §6.4]).

5.4 Mutations: mutual entropy between sequence-structure

As analogous to the expressional patterns of §4.2, the previous techniques are applied to the following case of mutations which are themselves functions of evolution, and together with selection and translational error, can influence the distribution of codons to the extent that the latter favor patterns of error-correction that drift to some optimal level and can ameliorate mutation effects [4, 66, 67]. For instance, let us consider as in [55] a series of amino acid sequences

$$\{\ldots, \mathrm{Seq}_{t-1}, \mathrm{Seq}_t, \mathrm{Seq}_{t+1}, \ldots\} = \{\mathrm{Seq}_t\}_{t\in\mathbb{Z}}, \tag{5.14}$$

where each Seq_t applies to one protein chain, ordered by a discrete temporal order $t \in \mathbb{Z}$ of corresponding tertiary structures

$$\{\ldots, \mathrm{Str}_{t-1}, \mathrm{Str}_t, \mathrm{Str}_{t+1}, \ldots\} = \{\mathrm{Str}_t\}_{t\in\mathbb{Z}}. \tag{5.15}$$

Such a chain can be represented as a noisy digital communication channel with an output probability of at least $\sim 30\%$, and with a Shannon limit at 10^{-2} bits/amino acid, where at each level t of sequence-structure we have the coding sequence

$$\begin{aligned} \mathrm{Seq}_t \Rightarrow\ & \text{Encoder} \Rightarrow \text{Folding channel} \\ & \Rightarrow \text{Decoder} \Rightarrow \mathrm{Str}_t \end{aligned} \tag{5.16}$$

as depicted in [55, Figure 1].

In [4] it is claimed that codes evolving with messages that mutate under such a process, tend to freeze with redundancy. This situation can be reduced to analyzing three different possibilities: the coevolution of genetic codes with:

(1) transitional-biased message mutation and no translation misreading;

(2) translational misreading and no transition bias in mutation;

(3) transition-biased message mutation and translational misreading.

An example in [55] considers concatenated primary sequences $\left\{\mathrm{Seq}_t\right\}_{t\in\mathbb{Z}}$ resulting in a stream of letters from the amino acid alphabet A with (alphabetical) size $|A| = 20$. The encoder is a map that uses a block code of fixed length n, say, to encode the source through the code book; in other words, a map for every sequence

$$\mathrm{Seq}_t \longrightarrow (\text{single code word}) X^n(\mathrm{Seq}_t), \tag{5.17}$$

represented by an n-vector $(X_1, \dots X_n)$ of integers. The code word in turn belongs to the book of 20 possible structure symbols $A^* = \{a_1^*, \dots, a_{20}^*\}$, the finite set of all code words corresponding to the 20 amino acid symbols $\{A, G, \dots\}$, where $a_j^* \in A^*$ are contact vectors determining the amino acid sequence. The message input term $X^n(\mathrm{Seq}_t)$ from (5.17) is relayed over a noisy channel which then outputs an n-vector $\Upsilon^n(\mathrm{Str}_t) = (Y_1, \dots, Y_n)$ representing the folded protein chain Str_t, following which a single use of the channels is the transmission of a single amino acid sequence subject to the *channel capacity*

$$C = \max_{p(A)} I(A, A^*). \tag{5.18}$$

In view of §5.3, we modify the role of $\hat{A}$ via the assignment $\hat{A} \mapsto A^*$, and for times stages t, t', take as above the metric $\mathcal{M}(\mathrm{Str}_t, \mathrm{Str}_{t'})$. At each side of the communication channel we have for the symbol sequences $|S_A| = 7702314$ amino acid symbols and $|S_{A^*}| = 31609$ corresponding structural symbols [55].

As for the code rate, we have $R(D) = H(A)/n$, where $H(A)$ is interpreted as the Shannon entropy of the amino acid sequence, where n is the code block length implemented by the encoder. Assuming the code rate $R(D)$ and channel capacity C are known, then in accordance with the Rate Distortion Theorem, we have $R(D) < C$, leading to, for every block size, $n > n_{\min} = H(A)/C$, and the codes exist, and no such code when $R(D) \geq C$. The Shannon entropy $H(A) = 3.90$ bits for the amino acid alphabet A, and $H(A^*) = 3.76$ bits for the structural code words in A^* [55]. Further, the mutual entropy between structure and sequence following [2] is given by

$$I(\mathrm{Seq}_t : \mathrm{Str}_t) = H(\mathrm{Seq}_t) - H(\mathrm{Seq}_t|\mathrm{Str}_t), \tag{5.19}$$

and should the environment directly influence the structure, then we would have

$$H(\mathrm{Str}_t|\mathrm{Seq}_t) \simeq H(\mathrm{Seq}_t|\mathrm{Env}_t). \tag{5.20}$$

When taking $H(\mathrm{Str}_t|\mathrm{Seq}_t) = 0$, we can re-formulate (5.19) as

$$\begin{aligned} I(\mathrm{Seq}_t : \mathrm{Env}_t) &\simeq I(\mathrm{Seq}_t : \mathrm{Str}_t) \\ &= H(\mathrm{Str}_t) - H(\mathrm{Str}_t|\mathrm{Seq}_t) \\ &= H(\mathrm{Str}_t), \end{aligned} \tag{5.21}$$

which in view of the mutual entropy between sequence and structure, expresses to what extent the thermodynamical entropy of possible protein structures can be constrained by information about the environment as it is coded by the sequence. For instance, excessive noise and random inputs of symbols in S_{A^*} would most probably corrupt a corresponding code in A^*, and once again the Shannon estimate serves as a threshold should errors exceed a critical bound. Empirically, the Protein Data Bank (PBD) provides sequence-structure data giving $H(A) = 3.90$ bits, with block length $n = 400$, with transmission rate $R(D) = 0.010$ bits per amino acid symbol followed, with channel capacity estimated at $C = 0.016$ bits (per amino acid symbol). When restricted to $\mathcal{N}_{25} = 2372$ protein chains with mutual sequence identity of < 0.25, the estimated $C_{(25)} = 0.016$ bits, was attained (see [55, Figure 4]).

6 The Topological Hypothesis and Phase Transitions

6.1 The codon space as a graph

The carrier for the dynamics surveyed here is modeled on a rate distortion manifold which has wide-scale overlap with those codon spaces structured in such a way that evolution can be influenced by mapping out those regions which can accommodate load minimization and diversification so that site type, coding fitness, targets, etc. can be correlated as in [4]. One expects the rate distortion manifold to have (in an analytic sense) some degree of differentiability, though here we will finesse this technical issue and elect to consider the underlying combinatorial structure. Specifically, we let $\Gamma = (V, E)$ denote a graph with V denoting a finite vertex set, E an edge set with an oriented edge $e = (u, v)$ (accordingly, $e^{-1} = (v, u)$) such that $u = i(e)$ is the initial vertex and $v = t(e)$ is the terminal vertex, and let F be

the number of enclosed faces. As seen in [73, 74] there is a formulation of the code that emerges at the phase transition appears in the form of a mode $e_{\alpha i}$ that minimizes the free energy F. The codon space can be described as such a graph Γ whose vertices are the codons and two codons i, j are linked by an edge if (see §3.5) there exists an associated $R_{ij}(\neq 0)$ in the reading matrix, under the following conditions/observations:

(1) The vertex set V consists of codons whereby two codons are linked by an edge in the likelihood they may be confused by misreading.

(2) Two codons are most likely to be confused if all their letters, except for one, agree and then they are connected by an edge. The resulting graph Γ is natural for considering the impact of translation errors on mutations because such errors almost always involve a single letter difference, that is, a movement along an edge of the graph to a neighboring vertex.

(3) The native state of the protein has the lowest available free energy induced by the interaction of the amino acid sequence with the embedding environment.

(4) Recall that there is an embedding $\Gamma \longrightarrow S$ into a surface S, and the topology of Γ is characterized by its genus $\gamma(S)$ which is the minimal number of holes required for Γ to be embedded in S such that no two edges cross. For the underlying network we have the well-known combinatorial formula $\gamma = 1 - \frac{1}{2}(V - E - F)$.

Thus the greater the number of connected components in the graph, the higher the genus becomes for a minimal embedding. In [73] the interconnected 64-codon graph can be embedded in a surface with genus $\gamma(S) = 41$. If only 48 effective codons are considered, then the genus is reduced to $\gamma(S) = 25$.

In light of these observations, it is claimed that the evolution of the code is determined by the underlying topology of its graph and in a transitional phase, it is only those modes with the least error-bound that can emerge and are subjected to alteration by the topology. From the perspective of [59], a free energy argument serves as a Morse function whose critical points characterize just such a topology. More specifically, [73] considers the topology of the code as imposing an upper limit to the number of low modes – critical points – of the corresponding free energy-analog functional, and this is also the number of amino acids. The low modes define a partition of the codon surface into domains, and in each domain a single amino acid is encoded. The partition optimizes the average distortion by minimizing the boundaries between the domains as well as the dissimilarity between neighboring amino acids. This bound on the number of low nodes (and thus as claimed, the number of amino acids) arises as an application of the well-known *chromatic number* as given by Heawood's formula [62]:

$$\mathrm{chr}(\gamma(S)) = \mathrm{int}[\frac{1}{2}(7 + \sqrt{1 + 48\gamma(S)})], \qquad (6.1)$$

where $\mathrm{chr}(\gamma(S))$ is the number of color domains of a surface S with genus $\gamma(S)$, and $\mathrm{int}[x]$ denotes the integer value of x. Recall also that the Euler characteristic $\chi(S) = 2 - 2\gamma(S)$. In particular, in [73, 75] it is the genus that represents the number of holes in the protein folding error network associated with the code and the chromatic number $\mathrm{chr}(\gamma(S))$ is a measure of the number of protein symmetries (see Tables 1 and 2.)

Example 6.1. Several topological configurations for doublet and triplet codes of 3-letter alphabets drawn from the mRNA alphabet $\{\mathrm{U}, \mathrm{C}, \mathrm{G}, \mathrm{A}\}$ are exhibited in [73, Fig. 3] and are enumerated by (6.1). The topological limit to the number of amino acids (AA's) for different codes as given by the chromatic number $\mathrm{chr}(g(S))$ is also given. For instance, a code of 48 codons gives rise to $g = g(S) = 25$ and $\mathrm{chr}(g(S)) = 20$, the maximal number of amino acids. Other cases are listed in [73, Table 1]. Further calculations for pairs $(g(S), \mathrm{chr}(g(S))$ are presented in [82] where the chromatic number $\mathrm{chr}(g(S))$ gives the number of protein symmetries: $(0, 4), (1, 7), (2, 8), (3, 9), (5, 10), (6, 11), (7, 11),$ $(8, 12), (9, 12)$.

More generally, for a topological manifold M having a Morse function F, $\chi(M)$ can be expressed as the alternating sum of the function's Morse indices μ_i $(i = 0, 1, ..., m)$ of F on M, defined as the number of critical points $(dF(x_c) = 0)$ of index i, that is, the number of negative eigenvalues of the matrix $H_{i,j} = \partial F^2/\partial x_i \partial x_j$. Then by the Poincaré-Hopf theorem,

$$\chi(M) = \sum_{i=0}^{m} (-1)^i \mu_i, \qquad (6.2)$$

which holds true for any Morse function on M (see e.g. [56] and Appendix 9.2 here).

Remark 6.1. Applying a spontaneous symmetry breaking argument to F_R generates topological transitions in the codon graph structure as the 'temperature' $R(D)$ increases; that is, as the average distortion D declines, via the inherent convexity of the rate distortion function. In other words, as the channel capacity connecting codon machines with amino acid machines increases, the more complex coding schemes become possible. In this respect, we recall that for the surface S, the Euler characteristic $\chi(S) = 2 - 2\gamma(S)$ as in (9.4) can be expressed in terms of the cohomology structure of S (e.g. [53, Theorem 13.38]) where by the Poincaré Duality Theorem, the homology groups of a manifold are related to the cohomology groups in the complementary dimension (e.g. [19, p.348]) and thus points to the 'fundamental homology' described earlier. One can then envisage the (co)homology groupoid to be taken as the disjoint union of the (co)homology groups of the embedding manifold.

6.2 Spectrum of the graph Laplacian

Next we consider the Laplacian Δ of Γ. If a pair of vertices $(i,j) \in E$ are adjacent, then in terms of e.g. the reading matrix $[R_{ij}]$ (with $R_{ij} > 0$), we have

$$\Delta_{ij} = \Delta_{ji} = -R_{ij} < 0, \tag{6.3}$$

otherwise $\Delta_{ij} = 0$, and $\Delta_{ii} = -\sum_{i \neq j} \Delta_{ij}$ (see Appendix §9.3). For instance, if Γ is taken to be the error graph of §6.1, then Δ is the operator that measures the effect of errors and so regulates any phase transition.

Corresponding to the n-th eigenvalue λ_n, the eigenfunction u_n admits at most n weak sign graphs; in particular, for $n = 2$, the eigenfunction u_2 divides Γ into precisely two weak sign graphs (see §9.3). Thus it is of interest to determine the dimension of the corresponding eigenspace and multiplicity m of λ_2. The quantity m is a measure of the first energy excitation being the primal mode for types of continuous (or second order) phase transitions. The chromatic number $\mathrm{chr}(\gamma(S))$ of (6.1) identifies the maximal number of first excited modes of the Δ.

Letting $\bar{m}(S)$ denote the supremum of m over all possible Δ on S, there is the estimate of Colin de Verdiére stating that $\bar{m}(S) \geq \mathrm{chr}(\gamma(S)) - 1$ (see e.g. [74]). In the case of functions, the graph Γ is a reliable 'spectral' model for S in the sense that from [34, Theorem 5.7], the eigenvalues of all orders of Δ on Γ converge to those of the *continuous* Laplacian on functions as defined on S (see Appendix 9.3).

6.3 Phase transitions and holonomy

Given the graph $\Gamma = (V, E)$, the *star of a vertex* $\mathsf{st}(v)$ is the set of edges emanating from v, that is

$$\mathsf{st}(v) = \{e : i(e) = v\}. \tag{6.4}$$

The various components of the graph may be thought of a comprising a cell network in which the coupling and equivalence of cells leads to a natural groupoid structure having a system of specific equivalence classes $[v]_V$ and $[e]_E$, for vertices and edges, respectively (see Appendix 8.1). With the inclusion of this extra structure we then append Γ to $\Gamma = (V, E, \sim_v, \sim_e)$. Here the vertices (nodes) of the network are representative of certain cells where the synchrony of the system depends on groupoid symmetries that in a sense is broken by an impinging rapid crosstalk internal to the system while the latter attempts to manage a slower external crosstalk.

Next, we implement some general procedures based upon the idea of a *connection* ∇ on Γ, relative to *the stars (*st*) of vertices* which following [17], is explained with some details in Appendix 9.1 as the combinatorial analog of covariant differentiation (a principle familiar to students of calculus). We take vertices $(e_1, e_2, \ldots, e_{k+1})$ interpreted as $k+1$ information sources $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{k+1})$ in accordance with the APSE condition of §4.1, where the $\mathbf{X}_i$ act with the set of tuning parameters. A connection ∇ is considered as an operation

$$\nabla(\mathbf{X}_i, \mathbf{X}_j) : \mathsf{st}(\mathbf{X}_i) \longrightarrow \mathsf{st}(\mathbf{X}_j), \tag{6.5}$$

for $1 \leq i, j \leq k+1$, satisfying certain properties (see Appendix 9.1). With respect to the metric $\mathcal{M} = \mathcal{M}(\mathbf{X}_i, \mathbf{X}_j)$ applied to these information sources, the above connection in (6.5) implements on the underlying network, the covariant differentiation along the path $\mathbf{X}_i \longrightarrow \mathbf{X}_i$, just as in (5.11):

$$dH/d\mathcal{M} = \lim_{\mathbf{X}_j \longrightarrow \mathbf{X}_i} \frac{H(\mathbf{X}_j) - H(\mathbf{X}_i)}{\mathcal{M}(\mathbf{X}_i, \mathbf{X}_j)}. \tag{6.6}$$

Corresponding to each $\mathbf{X}_i$, a maximized channel capacity $\mathbf{C}_i$ is assigned, in accordance with the Shannon estimate $H(\mathbf{X}_i) \leq \mathbf{C}_i$, for $1 \leq i \leq k+1$, thus respecting the Rate Distortion Theorem along paths $\mathbf{X}_j \longrightarrow \mathbf{X}_i$. If necessary, we can view $(\mathbf{X}_1, \mathbf{X}_2, \ldots, \mathbf{X}_{k+1})$ as comprising a closed geodesic, and as explained in Appendix 9.1, the set of these in a given graph will thus specify ∇. Once we have a handle on ∇ it is then possible to apply to Γ certain operations analogous to the more familiar differential-geometric setting in order explore the structural geometry of the various graphs as described (cf [40]).

This technique of the network geometry can be applied to the entropy rates occurring in the various cases we have considered so far. For the sequence-structure-environment in the noisy communication channels along with the data of §5.4, we assign Str_t (at time t) to a corresponding sensory input $\mathbf{S}_t$, further combined with environmental signals W_t, and combined signals a_t just as in (4.1):

$$\begin{cases} \mathrm{Str}_{t+1} & = f([\mathrm{Str}_t, W_t]) = f(a_t) \\ I(\mathrm{Seq}_t : \mathrm{Env}_t) & = H(\mathrm{Str}_t) \end{cases} \tag{6.7}$$

(here we have made replacements $i \mapsto t$ and $j \mapsto t'$), where we make a straightforward assignment from the vertex information source, at time t:

$$\mathbf{X}_t \mapsto \mathrm{Str}_t, \tag{6.8}$$

(and likewise for Seq_t). Using the principle of (6.5) applied to the mutual information

$$I(\mathrm{Seq}_t : \mathrm{Env}_t) = H(\mathrm{Str}_t), \tag{6.9}$$

in (5.21), leads to considering the covariant derivative

$$dH/d\mathcal{M} = \lim_{\mathrm{Str}_t \longrightarrow \mathrm{Str}_{t'}} \frac{H(\mathrm{Str}_t) - H(\mathrm{Str}_{t'})}{\mathcal{M}(\mathrm{Str}_t, \mathrm{Str}_{t'})}, \tag{6.10}$$

as implementing the graph connection

$$\nabla(\mathrm{Str}_t, \mathrm{Str}_{t'}) : \mathsf{st}(\mathrm{Str}_t) \longrightarrow \mathsf{st}(\mathrm{Str}_{t'}), \tag{6.11}$$

where again at each time stage t, the Shannon estimate $H(\mathrm{Str}_t) \leq C_t$ is observed. Likewise, the error load H_{ED} of §3.5 expressed in terms of paths $P_{\alpha ij \beta}$ in (3.15) and their concatenation, now become the meaningful paths of §4.1.

In this present graph formulism these paths are considered as determined by edges $e_\nu \in E$, where each $\nu = \nu(\alpha ij \beta)$ is a multi-index of the path subscripts.

A property of the connection ∇ in (6.5) is its *holonomy* which can be best described by considering how, in the traditional differential-geometric sense, a smooth connection implements the parallel translation of vectors around closed paths, and the induced representation of the space of the latter into a group of global symmetries is essentially the holonomy (of the connection). The classic example is the *Poincaré first–return map* of a dynamical system that incorporates typical phase transitions. In the combinatorial setting of [17] the holonomy of ∇ can be described formally in terms of permuting the 'stars' of vertices towards *a spatiotemporal reorientation*, as follows. Let $\mathcal{C} = \{e_1, \ldots, e_n\}$ be any cycle in the graph Γ, for which the terminal and initial vertices satisfy $t(e_\alpha) = i(e_{\alpha+1})$ modulo n. Then the connection around $\mathcal{C}$ leads to a permutation

$$\nabla_{\mathcal{C}} = \nabla_{e_n} \circ \cdots \circ \nabla_{e_1} \circ \nabla_{e_0}, \tag{6.12}$$

of the star set $\mathtt{st}(u)$. *The holonomy group* $\mathrm{Hol}(\Gamma, \nabla)_u$ at a vertex u of Γ, is the subgroup of the *permutation group of* $\mathtt{st}(u)$ generated by the permutations $\nabla_{\mathcal{C}}$ over all such cycles $\mathcal{C}$ that pass through the vertex u. A phase transition may then be represented by a permutation through vertices in Γ, and such a 'geometric phase' accounts for how the various bio-cognitive modules shift gear and create a reorientation of the system.

Now let us return to equivalence classes and the role of groupoids. This implements the above permutation groups of $\mathtt{st}(u)$. A holonomy groupoid is obtained via the disjoint union

$$\mathrm{Hol}(\Gamma, \nabla) = \bigvee_{u \in \Gamma} \mathrm{Hol}(\Gamma, \nabla)_u, \tag{6.13}$$

which pieces together the local operations, and at the same time produces an equivalence class representation of the phase transition and its internal amplitudes. We summarize this as follows: *the holonomy groupoid represents a globalization of the local dynamic iterates by providing what is essentially a representation of the graph's path components onto some prevailing group of symmetries*. In the presence of symmetry breaking, it would be reasonable to consider the groups $\mathrm{Hol}(\Gamma, \nabla)_u$ as commensurable to some degree with, for instance, the corresponding Lie groups featuring in the $\mathfrak{sp}(6)$ chain in (3.12), or that of the $\mathfrak{sl}(6, 1)$ chain as enumerated [10, 11].

7 Discussion and Conclusions

The code's development passed through 'accidental phases' created by probabilistic events that could be both regulated and manipulated by an evolving error-correction mechanism. Here we have viewed the latter within the framework of Shannon entropy and the context of the fundamental homology relative to the free energy density of a thermodynamical system. A common thread to this and other works suggests that increased selection forces may have been significantly enhanced by rate distortion dynamics in regard to the critical behavior of the free energy Morse function and varying topology, a function of which would have induced an order of redundancy so mandated by coevolution. Thermodynamic parameter changes in turn induced spontaneous symmetry breaking, which we have shown can captured by several techniques of representation theory. One can then invert Landau's arguments and apply them to the (co)homology groupoid in terms of the rising 'temperature' $R(D)$, to obtain a punctuated shift to increasingly complex genetic codes with increasing channel capacity. Our development here realizes mappings **codon space⟶amino acid space** quite explicitly in the context of rate distortion manifolds.

Such arguments can be supported by the known mechanisms occurring in the case of protein folding. The latter originating from an amino acid string is not an entirely random process, but may be the consequence of an evolved structured statement by an information source's uncertainty, and the occurrence of mutations which may not have been all random but were subject to environmental forces. Thus our present survey, besides regarding the functioning of gene expression as a cognitive process, has a link to the theme of the thermodynamic free energy landscape picture as a function of information sequences [3, 91](cf [54]), evolution as a problem in non-equilibrium statistical physics, and the self-referential character of evolutionary processes at large [43] (cf [83, 84]). We certainly acknowledge (though details are beyond the scope of this survey) that the evolution of organisms has evolved through environmentally sensitive biochemical processes. The phylogenetic analysis of sequence data and branching events suggests that amino acid sequences alter at almost a constant rate which is purported to depend on the functional nature of each class of protein. Thus the changing mechanism has been hypothesized in terms of an evolutionary, stochastic 'molecular clock' whereby minor fluctuations can alter the evolutionary rate of certain protein classes [90]. At the same time we have seen in the cognitive paradigm that some organisms may increase their rates of potentially deleterious mutation in response to environmental stress, and such occurrences afford a parallel interpretation in terms of rate distortion analysis as was previously surveyed.

Returning to the redundancy issue, the corresponding evolutionary processes may be capable of extending the code's expression from 20 to 25 amino acids with the possibility of there being many other protein folding codes [73] (cf [10, 11, 46]). Having said this, we add that there remain a number of open questions concerning the role of the rate distortion function $R(D)$, since this in turn drives punctuated changes in the genetic code and further exploration will be necessary. But what seems to follow from the collective processes we have described in explaining 'the frozen accident', is that certain *adaptation effects* are

in play (just as one finds in various neurocognitive and biosociological phenomena), and in this respect it seems fitting to quote from [4]:

> ... Our work has been motivated by the belief that the patterns of the standard genetic code may be explicable as adaptations of a system of information processing. If this turns out to be plausible and correct, we may say that adaptations have reduced the deleterious consequences of genetic and physiological error at a very fundamental level of biological organization ...

So the 'frozen accident' by any reasonable account, may have arisen as an evolutionary 'adaptation' against a temporary unreadiness (or an enforced over-robustness) to assimilate a barrage of highly complex genetic messaging, in a noisy and not so user-friendly biological environment, during which time error-correction patterns strived to crystallize and to evolve accordingly in order to withstand ongoing selective pressures. It is perhaps from this point of view that advocates of the 'RNA world idea' are likely to view a given adaptation at one stage as simply providing a pre-adaptation at another [63, 72].

We point out that holonomy and symmetry breaking are essentially geometric concepts that arise from the iterates of local-to-global procedures, and one such product of this is indeed the holonomy groupoid, a concept that has been introduced in this paper for the purpose of analyzing genetic networks in a novel setting. Further, the question of groupoid representations may uncover deeper conceptual issues in view of representation spaces that are spaces of operators ('fields of Hilbert or Banach spaces' as in e.g. [18]), a setting that may be compared the 'supersymmetric' model of [10, 11], but one that is likely to be highly nontrivial and costly in a computational sense. Thus in view of the various methods we have brought to the forefront, we cannot fail to acknowledge the remarkable insight of Erwin Schrödinger who claimed that classical physics was insufficient for understanding fundamental life processes. In particular, Schrödinger [65] had envisaged the potential importance of information theory in evolutionary genetics, how living systems can be alterable under thermodynamic effects that are often the results of adverse biological contagion and that quantum mechanical effects might catalyze potential mutations, revealing the organization and evolutionary drive of the genetic code all the more extraordinary.

Acknowledgements We wish to thank the reviewers for their various comments and the editors for their management of this paper. JFG wishes to thank Dr. Patrick Coulton for discussions concerning Heawood's formula. We are also grateful to Tracy Grauman for some production assistance.

8 Appendix: Groupoids and Their Atlases

8.1 Concept of a groupoid

Many bio-cognitive processes are naturally dynamical systems (see e.g. [40]). One aim in these systems is to unify the internal and external symmetries, and to be able to reduce vast myriad–like network configurations into manageable schemes involving the corresponding equivalence classes analogous to those already mentioned in source encoding/decoding, etc. in §3.2 (see also §5.3 below). A precise way of doing this lies within the categorical concept known as a *groupoid* (see e.g. [20, 28, 89]). In essence, a groupoid G consists of both a set of objects X and a set of morphisms, or 'arrows', each of which project to an object in X, and all such morphisms are invertible.

Remark 8.1. The most familiar example of a groupoid, as known to students of algebra, is that of a 'group' where there is a single object ('the identity'). Hence groupoids can be viewed as extensions of the 'group' concept to sets of *multiple identities* thus providing a wide scope of applications to the dynamics of neurocognitive, socio–bioinformatic and cellular networks (see e.g. [40, 71]).

A groupoid can be depicted by

$$\alpha, \beta \ : \ \mathsf{G} \underset{\beta}{\overset{\alpha}{\rightrightarrows}} X \tag{8.1}$$

where the groupoid morphisms (α, β) onto objects, are called the *range* and *source maps*, respectively. Informally, the groupoid represents a feature of built in reciprocity between its algebraic structures, internalizing and externalizing the prevailing symmetries. The morphisms α, β satisfy certain algebraic relations of associativity, existence of two-sided identities, etc. (for details, see [20, 28, 89]). A groupoid can here be understood in relationship to a linkage by a meaningful path of an information source dual to a cognitive process for which the underlying principle is that: *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 point, and the tuning across the various possible ways in which that can happen – the different cognitive languages – parametrizes the set of equivalence relations and creates the groupoid.*

Example 8.1. Since we have already mentioned equivalence classes in the context of source encoding/decoding, it seems appropriate to see how an equivalence relation $\mathcal{R}$ defined on (a set) X takes shape as a groupoid. Here we have the two projections $\alpha, \beta : \mathcal{R} \longrightarrow X$, and a product $(x, y)(y, z) = (x, z)$ whenever $(x, y), (y, z) \in \mathcal{R}$ together with an identity, namely (x, x), for each $x \in X$. Moreover, the essential equivalence relations and equivalence classes derived from a systems space (network) arise from the orbit equivalence relation of some groupoid G acting on that space (see e.g. [89]). In the context of con-

nected (sub)networks/graphs with path concatenation, representable in terms of equivalence classes, natural groupoid structures arise in accordance with equivalence classes of relations $\mathcal{R}(xy)$, as above, that is simply interpreted as there exists an edge linking node x to node y (thus $x\mathcal{R}y$). Conversely, a groupoid (of equivalence relations) admits an underlying graph structure via its implicit scheme of objects and morphisms between objects (for details, see e.g. [20, 44]). Thus we have the two-way associations whereby 'objects' can be identified with 'nodes', and 'morphisms' identified with 'edges' in groupoids (of equivalence relations) and networks, respectively:

$$\text{Network} \overset{\text{equivalence relation}}{\Longrightarrow} \text{Groupoid}$$

$$\text{Network} \overset{\text{underlying graph}}{\Longleftarrow} \text{Groupoid}$$

9 Appendix: Some Geometry of the Network Architecture: Geodesics and Phase Transitions

9.1 Connections on graphs and geodesics

Firstly, for graph–theoretic models there are certain combinatorial notions which can be used to replicate a 'differential' structure as realized on a standard differentiable manifold (such as a sphere or a torus). Let $\Gamma = (V, E)$ be a graph with V denoting a finite vertex set, E an edge set with an oriented edge $e = (u, v)$ (accordingly, $e^{-1} = (v, u)$) such that $u = i(e)$ is the initial vertex and $v = t(e)$ is the terminal vertex. The *star of a vertex* $\mathsf{st}(v)$ is the set of edges emanating from v, that is

$$\mathsf{st}(v) = \{e : i(e) = v\}. \tag{9.1}$$

In principle, we would like a handle on both the groupoid and geometric dynamics of a given network. One point is that the star of a vertex may be viewed as the combinatorial version of the tangent space to a manifold at a point, rather similar to how the latter may be regarded as an equivalence class of curves through that point. In [17] there is defined the notion of a *connection* ∇ on a graph Γ expressed in terms of a set of one–to–one functions $\nabla(u, v)$, one for each oriented edge $e = (u, v)$ of Γ satisfying the following relationships:

(1) $\nabla(u, v) : \mathsf{st}(u) \longrightarrow \mathsf{st}(v)$

(2) $\nabla(u, v)(u, v) = (v, u)$

(3) $\nabla(v, u) = (\nabla(u, v))^{-1}$

Given a graph Γ admits a connection ∇, [17] define the notion of a *3–geodesic* as a sequence of four vertices (u, v, w, z) with edges $\{u, v\}$, $\{v, w\}$ and $\{w, z\}$ for which

$$\nabla(v, w)(v, u) = (w, z). \tag{9.2}$$

Remark 9.1. In differential calculus, a 'connection' is simply a generalized gradient implementing covariant differentiation. We have already encountered a form of this in (5.11). The notion of a graph/network connection introduced here is a more manageable concept, particularly for bio-cognitive modules, and does not involve applying the advanced techniques of calculus.

A *k–geodesic* is defined inductively across a sequence of $(k + 1)$ vertices. The three consecutive edges $\{d, e, f\}$ of a 3–geodesic is referred to as *an edge chain*. A *closed geodesic* can then be specified as a sequence of edges $e_1, \ldots, e_n$ such that each consecutive triple $(e_\alpha, e_{\alpha+1}, e_{\alpha+2})$ is an edge chain for each $1 \leq \alpha \leq n$, modulo n . The geodesic returns to the same pair of edges in the same order. Thus one finds a unique closed geodesic through each pair of edges in the star of the vertex, and as pointed out in [17], the set of all closed geodesics completely determines the connection on the graph.

In terms of the geometric evolution of our networks, the family $(\mathsf{G}_\mathcal{A}, \nabla_\mathcal{A})$ of local groupoids with connection satisfies:

(1) Once $\nabla_\mathcal{A}$ is given, then the graph geodesics can be derived iteratively from (9.2).

(2) Conversely, given the underlying graph of each $\mathsf{G}_\mathcal{A}$, the connection $\nabla_\mathcal{A}$ is determined by the set of all closed geodesics as specified.

We also have the following useful characterization [17]: given (Γ, ∇), a subgraph $\Gamma_0 = (V_0, E_0) \subset \Gamma$ is said to be *totally geodesic* if all geodesics commencing at E_0 remain within E_0. In other words, for every two adjacent vertices u, v in Γ_0, we have

$$\nabla(u, v)(\mathsf{st}(u) \cap E_0) \subseteq E_0. \tag{9.3}$$

Note that the above concepts have been formulated graph-theoretically, and as mentioned in Remark 9.1, they do not require the usual manipulations of advanced differential calculus.

9.2 The graph Betti numbers

By analogy with finding the dimensions of the homology groups of a topological manifold, [17] specify the notion of *Betti numbers* associated with Γ. This involves the using certain concepts such as an *axial function* φ and *generic direction* ξ. Thus we regard (Γ, ∇) as having an axial function φ and write this as (Γ, φ) when ∇ is understood. In which case the *index* of a vertex $u \in V$ is the number of edges $e \in \mathsf{st}(u)$ such that the product $\varphi(e) \cdot \xi < 0$. Let $\beta_i(\xi)$ denote the number of vertices u such that the index at u is exactly i. When these values do not depend on the choice of direction ξ, they are called the *Betti numbers of* (Γ, φ), and satisfy a combinatorial duality condition $\beta_i(\Gamma, \varphi) = \beta_{d-i}(\Gamma, \varphi)$, for $1 \leq i \leq d$. In certain cases, they can shown to be similar to the indices of a standard

Morse function (see [17, 58, 56]) such as F_R in (3.11). Thus on the underlying graph of the groupoid on which F_R is defined, we identify F_R with a Morse function compatible with a generic direction on (Γ, φ) whose index is essentially a measure of the homology of information relay within the graph, where at level i, we have $\mu_i = \beta_i(\Gamma, \varphi)$.

In fact, to clarify the role of the topological invariants of Γ to those of the surface S, we need the following description. Firstly, S taken to be a compact surface permits seeing S also as a (connected) compact, one-dimensional complex manifold (viz. a Riemann surface) on which a certain analytic group action takes place. The standard way of representing Γ (see e.g. [17, §4]) is to identify V as the (finite) fixed point set, and E as the (finite) set of one-dimensional orbits of this action. Consequently, the $\beta_i(\Gamma, \varphi)$ coincide with the usual Betti numbers $\beta_i(S)$ of S, and by the Poincaré-Hopf Theorem we have

$$\chi(S) = \sum_i (-1)^i \beta_i(\Gamma, \varphi). \tag{9.4}$$

9.3 The graph Laplacian

Suppose now that $\Gamma = (V, E)$ is an undirected loop-free graph. If the vertices(nodes) are indexed $1 \leq i \leq N$, then the *graph Laplacian* Δ can be viewed as a symmetric $N \times N$ matrix defined as follows (see e.g. [16, 74]):

(1) If vertices $(i, j) \in E$ are adjacent, then the corresponding entry in the matrix $\Delta_{ij} = \Delta_{ji} < 0$.

(2) Otherwise, $\Delta_{ij} = 0$, and the diagonal terms imply that the sum over rows and columns vanishes, leading to $\Delta_{ii} = -\sum_{i \neq j} \Delta_{ij}$.

Note the term 'weighted Laplacian' is sometimes used for the operator Δ, whereas in other cases 'Laplacian' is used for when the negative entries are all $\Delta_{ij} = -1$. Specifically, if $f : V \longrightarrow \mathbb{R}$ is a vector function induced by the vertices of Γ, and $x \sim y$ denotes there is an edge linking x and y, then from [16]:

$$(\Delta f)(x) = -\sum_{x \sim y} [f(x) - f(y)]. \tag{9.5}$$

Of particular interest are the eigenvalues of Δ ordered as $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_N$, obtainable through the spectrum of an associated operator L, for which

$$\langle f, Lf \rangle = \sum_{x,y \in V} L_{xy} f(x) f(y) = \sum_{xy \in E} [f(x) - f(y)]^2. \tag{9.6}$$

Also, we have the *Rayleigh Quotients* [16], given by

$$\begin{cases} \mathcal{R}_\Delta & = \frac{\langle f, \Delta f \rangle}{\langle f, f \rangle} \\ \mathcal{R}_L(f) & = \frac{\sum_{xy \in E} [f(x) - f(y)]^2}{\sum_{x \in V} f(x)^2}. \end{cases} \tag{9.7}$$

In [34, Theorem 5.7] estimates on (9.7) lead to showing that, in the case of functions, the eigenvalues of the graph Laplacian converge to those of the continuous Laplacian. Further, in [34] it is shown that the zeta functions of the former converge to those of the latter, where

$$\zeta^{(n)}(s) = \sum_{\lambda_k^n \neq 0} (\lambda_k^n)^{-s}. \tag{9.8}$$

In the continuous case, sets that are the zero-level sets of the eigenfunctions are called *nodal sets*, and *nodal domains* are those sets in which a corresponding eigenfunction takes on one sign and they are separated by nodal sets. Courant's nodal line theorem (see e.g. [24]) states that if the eigenfunctions of a continuous Laplacian on a domain are ordered according to increasing eigenvalues, then the nodes of the n-th eigenfunction divide the domain into no more than n nodal domains. In the combinatorial case, for the graph Laplacian, the nodal domains become *sign-graphs*: maximal connected subgraphs on which an eigenfunction carries the same sign. On *weak sign-graphs* the eigenfunction is either ≥ 0 or ≤ 0, while on *strong sign-graphs*, the sign of the eigenfunction is either > 0 or < 0. This leads to an analogue of Courant's nodal line theorem in the combinatorial case [16]: *On a connected graph Γ, the n-th eigenfunction u_n of the Laplacian Δ admits at most n weak sign graphs*. The case $n = 2$ is significant because the corresponding eigenfunction u_2 then splits Γ into exactly two weak sign graphs and λ_2 is significant for Brownian motion on the graph and to its first excited energy level.

10 Spin Glasses in Brief

Spin glass models, as discrete structures, may be based on combinatorial decompositions of surfaces usually in some square lattice configuration which can be modified (e.g. from square to triangular). The basic idea leading to the prototypical *2-dimensional Ising model* goes as follows (we follow [23, 69]). Firstly, consider a sequence of symbols $a_i = 0, 1$ and a signal v_i transmitted across some time interval. Set $v_i = v$ if $a_i = 1$, and $v_i = -v$ if $a_i = 0$. Then let $a(i, j)$ (for $1 \leq i, j \leq m$) denote the m^2 bits of information transmitted. These are subject to redundancy relations

$$\begin{aligned} a(i, m+1) &= \sum_{j=1}^{m} a(i, j), \\ a(m+1, j) &= \sum_{i=1}^{m} a(i, j), \end{aligned} \tag{10.1}$$

with addition mod 2. The quantity $m^2/(m+1)^2$ is the rate of the code and measures the redundancy. With noise terms $y(i, j)$ included, the modified signal is then taken to be $u(i, j) = v(i, j) + y(i, j)$. This leads to a simple error-correcting code that is of the Hamming type [57]. Further, the correspondence $u(i, j) = \frac{1}{2}(\sigma(i, j) + 1)$ between information bits and Ising spins or qubits $\sigma(i, j)$ in mod 2 addition and spin multiplication respectively, are equivalent.

More specifically, let a qubit $\sigma(i,j)$ be attached to each edge of some lattice which is to be viewed as a configuration space $P = \{\pm\}^{\mathbb{Z}^2}$. On taking J_1 (horizontal) and J_2 (vertical) to be interaction constants, the Hamiltonian $\mathcal{H}(\sigma)$ is given by

$$\begin{aligned}\mathcal{H}(\sigma) = -\sum J_1\sigma(i,j)\,\sigma(i+1,j) \\ + J_2\sigma(i,j)\,\sigma(i,j+1),\end{aligned} \tag{10.2}$$

for the appropriate ranges of summation. Suppose we consider $H(\sigma)$ over a finite lattice given by $\Lambda_{LM} = \{(i,j) : |i| \leq M\ ,\ |j| > L\}$, and then take the thermodynamic limit. If $J_1, J_2 > 0$, there are interactions in which the energy is minimized on alignment of all of the spins. Then either:

i) all are $\uparrow$ or $\downarrow$, or,

ii) $\sigma(i,j) \equiv 1$, or $\sigma(i,j) \equiv -1$, respectively.

For absolute temperature T, the equilibrium state is that which minimizes [internal energy] $- T \cdot$ [entropy]. Within the model two competing forces can be realized by the following:

1. One minimizes the internal energy by attempting to align the signs either $\uparrow$ or $\downarrow$ to create order: it wins if T is small.

2. The other, on maximizing entropy, attempts to produce as much chaos as possible: it wins if T is large.

At finite critical temperature T_c, chaos wins if $T \geq T_c$, and order wins if $T < T_c$.

References

[1] Adami, C., Ofria, C., and Collier, T., 2000, Evolution of biological complexity, *Proc. Natl. Acad. Sci. USA* **97**, 4463–4468.

[2] Adami, C., 2004, Information theory in molecular biology, *Physics of Life Reviews* **1**, 3-22.

[3] Anfinsen, C.B., 1973, Principles that govern the folding of protein chains, *Science* **181** (96), 223–230.

[4] Ardell, D.H., and Sella, G., 2002, No accident: genetic codes freeze in error-correcting patterns of the standard genetic code, *Phil. Trans. R. Soc. Lond. B* DOI 10.1098/rstb.2002.1071

[5] Ash, R., 1990, *Information Theory*, Dover Publications, New York.

[6] Atlan, H., and Cohen, I., 1998, Immune information, self-organization and meaning *Int. Immunology*, **10**, 711-717.

[7] Austin, D.W., Allen, M.S., McCollum, J.M., Dar, R.D., Wilgus, J.R., Sayler, G.S., Samatova, N.F., Cox, C.D., and Simpson, M.L., 2006, Gene network shaping of inherent noise spectra, *Nature* **439**, doi:10.1038/nature04194

[8] Avetisov, V., and Goldanskii, V., 1996, Mirror symmetry breaking at the molecular level, *Proc. Natl. Acad. Sci. USA* **93**, 11435–11442.

[9] Bak, A., Brown, R., Minian, G., and Porter, T., 2006, Global actions, groupoid atlases and related topics, *J. Homotopy and Related Structures* **1**, 1-54.

[10] Bashford, J.D., Tsohantjis, I., and Jarvis, P.D., 1998, A supersymmetric model for the evolution of the genetic code, *Proc. Natl. Acad. Sci. USA* **95**, 987–992.

[11] Bashford, J.D., and Jarvis, P.D., 2008, Spectroscopy of the genetic code, in (Abbott, D. et al. eds) *Quantum Aspects of Life*, Imperial College Press, London.

[12] Belongie, M.L., 1994, Spin glasses and error-correcting codes, *TDA Progress Report 42–118*, 26–36.

[13] Bennett, C.H., 1982, The thermodynamics of computation: a Review, *Internat. J. Theor. Phys.* **21** (12), 905–940.

[14] Berger, T., 1971, *Rate Distortion Theory: A mathematical basis for data compression*, Prentice–Hall Inc., Englewood Cliffs, NJ.

[15] Bertman, M.O., and Jungck, J.R., 1978, Some unresolved mathematical problems in genetic coding, *Notices Amer. Math. Soc.* **25** A-174.

[16] Biyikoğlu, T., Leydold, J., and Stadler, P.F., 2007, *Laplacian Eigenvectors of Graphs*, Lecture Notes in Math. **1915**, Springer-Verlag, Berlin Heidelberg.

[17] Bolker, E.D., Guillemin, V.W., and Holm, T.S., 2006, How is a graph like a manifold?, to appear. http://arxiv:math.CO/0206103

[18] Bos, R., 2011, Continuous representations of groupoids, *Houston J. Math.* **37**(3), 807–844.

[19] Bredon, G., 1993, *Topology and Geometry*, Springer, New York.

[20] Brown, R., 2006, *Topology and Groupoids* (3rd Ed.), BookSurge LLC, Charleston, S. Carolina.

[21] Bryngelson, J.D., and Wolynes, P.G., 1987, Spin glasses and the statistical mechanics of protein folding, *Proc. Natl. Acad. Sci. USA* **84**, 7524–7528.

[22] Burago, D., Burago, Y., and Ivanov, S., 2001, *A Course in Metric Geometry*, American Mathematical Society, Providence, RI.

[23] Carey, A., and Evans, D. E., 1988, The operator algebras of the two–dimensional Ising model, in *'Braids' (Santa Cruz, CA, 1986)*, 117–165, *Contemp. Math.* **78**, Amer. Math. Soc., Providence, RI.

[24] Chavel, I., 1984, *Eigenvalues in Riemannian Geometry*, Academic Press, New York.

[25] Ciliberti, S. Martin, O., and Wagner, A., 2007, Innovation and robustness in complex regulatory genetic networks, *Proc. Natl. Acad. Sci. USA* **104**, 13591–13596.

[26] Cohen, I., 2000, *Tending Adam's Garden: Evolving the Cognitive Immune Self*, Academic Press, New York.

[27] Cohen, I., and Harel, D., 2007, Explaining a complex living system: dynamics, multiscaling and emergence, *J. Royal Soc. Interface* **4**, 175–182.

[28] Connes, A., 1994, *Noncommutative Geometry*, Academic Press, San Diego, CA.

[29] Cover, T., and Thomas, J., 1991, *Elements of Information Theory*, John Wiley and Sons, New York.

[30] Crick, F., 1966, Codon-anticodon pairing:the wobble hypothesis, *J. Mol. Biol.* **19**, 548–553.

[31] Crick, F., 1968, The origin of the genetic code, *J. Mol. Biol.* **38**, 367–379.

[32] Crooks, G.E. and Brenner, S., 2004, Protein secondary structure:entropy, correlations and prediction, *Bioinformatics* **20**(10), 1603–1611.

[33] Dawkins, R., 1976, *The Selfish Gene*, Oxford Univ. Press, London.

[34] Dodziuk, J., 1976, Finite-difference approach to the Hodge theory of harmonic forms, *Amer. J. Math.* **98**(1), 79–104.

[35] Dretske, F., 1981, *Knowledge and the Flow of Information*, MIT Press, Cambridge, MA.

[36] Feynman, R., 1996, *Feynman Lectures on Computation*, Addison-Wesley, Reading, MA.

[37] Franzosi, R., and Pettini, M., 2004, Theorem on the origin of phase transitions, *Phys. Rev. Lett.* **92**:060601.

[38] Freeland, S.J., Wu, T., and Keulmann, N., 2003, The case for an error minimizing standard genetic code, *Origins of Life and Evolution of the Biosphere* **33**, 457–477.

[39] Gent, I.P., Kelsey, T., Linton, S., Pearson, J., and Roney-Dougal, C.M., 2010, Groupoids and conditional symmetry, preprint, Univ. St. Andrews, UK.

[40] Glazebrook, J.F., and Wallace, R., 2009, Small Worlds and Red Queens in the Global Workspace: an information–theoretic approach, *Cognitive Systems Research* **10**, 333–365.

[41] Glazebrook, J.F., and Wallace, R., 2009, Rate distortion manifolds as model spaces for cognitive information, *Informatica* **33** (2009), 309–345.

[42] Glazebrook, J.F., and Wallace, R., 2010, Rate distortion coevolutionary dynamics and the flow nature of cognitive epigenetic systems, arXiv:1101.4984v1 [q-bio.OT]

[43] Goldenfeld, N., and Woese, C., 2011, Life is physics: evolution as a collective phenomenon far from equilibrium, *Annu. Rev. Condens. Matter Phys.* **2**, 375-399.

[44] Golubitsky, M., and Stewart, I., 2006, Nonlinear dynamics and networks: the groupoid formalism, *Bull. Amer. Math. Soc.* **43**, 305-364.

[45] Gupta, M.K., 2006, The quest for error correction in biology, *IEEE Engineering in Medicine and Biology Magazine*, 46–53.

[46] Hornos, J.E., and Hornos, Y.M., 1994, A search for symmetries in the genetic code, *J. Bio. Phys.* **20**, 289–294.

[47] Jiménez-Montaño, M.A., de la Mora-Basáñez, C.R., and Pöschel, T., 1996, The hypercube structure of the genetic code explains non-conservative aminoacid substitutions *in vivo* and *in vitro*, *BioSystems* **39**, 117–125.

[48] Jukes, T.H., 1983, Evolution of the amino acid code, pp. 191–207 in (Nei, M. et al. eds.) *Evolution of Genes and Protein*, Sinauer, Sunderland, MA.

[49] Khinchin A., 1957, *The Mathematical Foundations of Information Theory*, Dover Publications, New York.

[50] Koonin, E., and Novozhilov, A., 2009, Origin and evolution of the genetic code: the universal enigma, *Life* **61**, 99–111.

[51] Kurzynski, M., 2006, *The Thermodynamic Machinery of Life*, Springer–Verlag, Berlin Heidelberg New York.

[52] Landau, L., and Lifshitz E., 2007, *Statistical Physics* (I) (3rd Ed.), Elsevier, New York.

[53] Lee, J., 2000, *Introduction to Topological Manifolds*, Springer, New York.

[54] Levinthal, L., 1968, Are there pathways for protein folding?, *J. Chim. Phys. PCB* **65**, 44-45.

[55] Lisewski, A. M., 2008, Random amino acid mutations and protein misfolding lead to Shannon limit in sequence-structure communication, PloS ONE, **3**(9), e3110 doi:10.371/journal.pone.0003110

[56] Matsumoto, Y., 2001, *An Introduction to Morse Theory*, Translations Amer. Math. Soc. **208**, Providence, RI.

[57] McEliece, R.J., 2004, *The Theory of Information and Coding*, Encyclopedia of Mathematics and its Applications, Vol. 86, Cambridge University Press.

[58] Milnor, J., 1963, *Morse Theory*, Princeton University Press, Princeton, NJ.

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

[60] Protter, P., 1995, *Stochastic Integration and Differential Equations:A New Approach*, Springer, New York.

[61] Prügel-Bennett, A. and Shapiro, J. L., 1994, Analysis of genetic algorithms using statistical mechanics, *Phys. Rev. Lett.* **72**(9), 1305–1309.

[62] Ringel, G., and Youngs J., 1968, Solutions of the Heawood map-coloring problem, *Proc. Natl. Acad. Sci. USA* **60**, 438-445.

[63] Rodin, A.S., Szathmáry, E., and Rodin, S.N., 2011, On origin of genetic code and tRNA before translation, *Biology Direct* **6**:14, 1–24.

[64] Rose, K., Gurewitz, E., and Fox, G.C., 1990, Statistical mechanics and phase transitions in clustering, *Phys. Rev. Lett.* **65** No. 6, 945–948.

[65] Schrödinger, E., 1967, *What is Life?*, Cambridge University Press.

[66] Sella, G. and Ardell, D.H., 2002, The impact of message mutation on the fitness of the genetic code, *J. Mol. Evol.* **54**, 638–651.

[67] Sella, G. and Ardell, D.H., 2006, The coevolution of genes and genetic codes:Crick's frozen accident revisited, *J. Mol. Evol.* **63**, 297–313.

[68] Söll, D., and RajBhandary, U.L., 2006, The genetic code–Thawing the 'frozen accident', *J. Biosci.* **31**(4), 459–463.

[69] Sourlas, N., 1989, Spin-glass models as error-correcting codes, *Nature* **339**, 693–695. doi:10.1038/339693a0

[70] Stewart, I., Golubitsky, M., and Pivato, M., 2003, Symmetry groupoids and patterns of synchrony in coupled cell networks, *SIAM J. Appl. Dynam. Sys.***2**, 609–646.

[71] Stewart, I., 1994, Broken symmetry in the genetic code?, *New Scientist* **1915**, 16.

[72] Szathmáry, E., 1999, The origin of the genetic code, *Trends in Genetics* **15**(6), 223–229.

[73] Tlusty, T., 2007, A model for the emergence of the genetic code as a transition in a noisy information channel, *J. Theor. Bio.* **249**, 331–342.

[74] Tlusty, T., 2008, Rate-distortion scenario for the emergence and evolution of noisy molecular codes, *Phys. Rev. Lett.* **100**, 048101 (1-4).

[75] Tlusty, T., 2007, 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, *Elect. J. Linear Algebra* **16**, 315–324.

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

[77] Vetsigian, K., Woese, C., and Goldenfeld, N., 2006, Collective evolution and the genetic code, *Proc. Natl. Acad. Sci. USA* **103**, 10696-10701.

[78] de Vladar, H.P., and Barton, N.H., 2011, The contribution of statistical physics to evolutionary biology, *Trends in Ecology and Evolution* **26**(8), 424–432.

[79] Wallace, R., 2005, *Consciousness:A Mathematical Treatment of the Global Neuronal Workspace Model*, Springer, New York.

[80] Wallace, R., 2010, A rate distortion approach to protein symmetry, *BioSystems* **101**, 97–108.

[81] Wallace, R., 2011, Metabolic constraints on the evolution of genetic codes: Did multiple 'preaerobic' ecosystem transitions entrain richer dialects via Serial Endosymbiosis, http://precedings.nature.com/documents /4120/version/4.

[82] Wallace, R., 2011, Structure and dynamics of the 'protein folding code' inferred using Tlusty's topological rate distortion approach, *Biosystems* **103**, 18–26.

[83] Wallace, R., and Wallace R.G., 1998, Information theory, scaling laws, and the thermodynamics of evolution, *J. Theor. Bio.* **192**, 545-559.

[84] Wallace, R., and Wallace R.G., 1999, Organisms, organizations and interactions:an information theory approach to biocultural evolution, *BioSystems* **51**, 101-119.

[85] Wallace, R., and Wallace, D., 2009, *Gene Expression and its Discontents: The social production of pandemic chronic disease*, Springer, New York.

[86] Wallace, R., and Wallace, D., 2008, Punctuated equilibrium in statistical models of generalized coevolutionary resilience: how sudden ecosystem transitions can entrain both phenotype expression and Darwinian selection, *Trans. Comp. Systems Biology IX*, LNBI 5121, 23 –85.

[87] Wallace, R., and Wallace, D., 2011, Cultural epigenetics: on the heritability of complex diseases, *Trans. Comp. Systems Biology XIII*, LNBI 6575, 131–170.

[88] Wallace, R., and Fullilove M., 2008, *Collective Consciousness and Its Discontents: Institutional distributed cognition, racial policy, and public health in the United States*, Springer, New York.

[89] Weinstein, A., 1996, Groupoids: unifying internal and external symmetry, *Notices Amer. Math. Soc.* **43**, 744-752.

[90] Wilson, A.C., Carlson, S.S., and White, T.J., 1977, Biochemical Evolution, *Ann. Rev. Biochem.* **46**, 573–639.

[91] Wolynes, P.G., 1996, Symmetry and the energy landscapes of biomolecules, *Proc. Nat. Acad. Sci. USA* **93**, 14249–14255.

[92] Yockey, H.P., 2005, *Information Theory, Evolution and the Origin of Life*, Cambridge University Press.