Version 8

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

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

January 11, 2010

Abstract

A mathematical model based on Tlusty's topological deconstruction suggests that multiple punctuated ecosystem shifts in available metabolic free energy, broadly akin to the 'aerobic' transition, enabled a punctuated sequence of increasingly complex genetic codes and protein translators under mechanisms similar to the Serial Endosymbiosis effecting the Eukaryotic transition. These evolved until the ancestor of the present narrow spectrum of nearly maximally robust codes became locked-in by path dependence.

Keywords genetic code, information theory, metabolic free energy, phase transition, rate distortion, topology

1 Introduction

The genetic code that maps 64 codons to 20 amino acids is far from random, e.g., figure 1 of Koonin and Novozhilov (2009), and the references therein. Typically, codons that differ by only one nucleotide tend to code for either the same or two related amino acids, that is, amino acids that are 'physicochemically similar'. Koonin and Novozhilov assert that the fundamental question is how these regularities of the standard code came into being, given that there are

*Address correspondence to Rodrick Wallace, 549 W. 123 St., Apt. 16F, New York, NY, 10027 USA, email wallace@pi.cpmc.columbia.edu.

more than 10^{84} possible alternative code tables if each of the 20 amino acids and the stop signal are assigned to at leasts one codon:

> The features... seeming to need special explanation include, but are not limited to, the block structure of the code, which is thought to be a necessary condition for... robustness with respect to point mutations, translational misreading, and translational frame shifts; the link between the second codon letter and the properties of the encoded amino acid, so that codons with U in the second position correspond to hydrophobic amino acids.... the apparent minimization of the likelihood of mistranslation and point mutations; and the near optimality for allowing additional information within protein coding sequences.

Tlusty (2007) has presented a model for the emergence of the genetic code as a transition in a noisy information channel, using an approach based on the Rate Distortion Theorem. In that analysis the optimal code is described by the minimum of a 'free energy'-like functional, which leads, in his view, naturally to the possibility of describing the code's emergence as a transition akin to a phase transition in statistical physics. The basis for this is the observation that a supercritical phase transition is known to take place in noisy information channels (e.g., Rose, 1998). The noisy channel is controlled by a temperature-like parameter that determines the balance between the information rate and the distortion 'in the same way that physical temperature controls the balance between energy and entropy' in a physical system. Following Tlusty's equation (2), the 'free energy' functional has the form $D - TS$ where D is the average 'error load', equivalent to average distortion in a rate distortion problem, S is the entropy due to random drift, and T measures the strength of random drift relative to the selection force that pushes towards fitness maximization. According to Tlusty's analysis, at high T the channel is totally random and it conveys zero information. At a certain critical temperature T_c the information rate starts to increase continuously.

The average distortion D measures the average difference between the genetic 'message' send by a complicated codon 'statement' and what is actually expressed by the genetic (and epigenetic) translation machinery in terms of an amino acid sequence. We give a more complete discussion in the sections to follow.

Here we will take a different route, one in which the rate distortion function $R(D)$ between codon pattern and amino acid pattern plays the role of of a temperature-analog driving phase transitions in a corresponding free-energy analog constructed from the distribution of possible genetic codes, as measured by the source uncertainty of the information sources using them. Pettini's (2007) 'topological hypothesis' ensures topological shifts in code structure accompany these phase transitions. The question then arises as to what drives the dynamics of $R(D)$. Several models emerge.

In the simplest approach, the dynamics of the system are defined, not by the minimization of a functional, but by an 'empirical Onsager relation' to be

associated with a particular structure. The rate distortion function is the minimum channel capacity needed to keep average distortion at or below D, and information channel capacity is, following the arguments of Feynman (2000) and Bennett (1988), measured by the free energy needed to erase the sent message. Dynamics are then driven by the gradient of an entropy-analog, the rate distortion disorder S_R at distortion D, defined as

$$S_R \equiv R(D) - DdR(D)/dD. \tag{1}$$

Most simply, the dynamics will be given by a generalized empirical Onsager equation of the form

$$dD/dt = -\mu dS_R(D)/dD + F(D,t), \tag{2}$$

under the important constraint that the Rate Distortion Function $R(D)$ *is always a convex function of* D (Cover and Thomas, 1991, Lemma 13.4.1).

We extend these considerations both downward and upward in scale, examining the effects of changes in $R(D)$ on internal structure within the genetic code, and studying the effect of available metabolic free energy on $R(D)$ itself, and treat more complicated models as well, recognizing that evolutionary process does not always seem to instantiate Occam's Razor.

Phase transitions in physical systems characterized by free energy are ubiquitous, following Landau's symmetry breaking arguments (Landau and Lifshitz, 2007; Pettini, 2007): Higher temperatures enable higher system symmetries, and, as temperature declines, punctuated shifts to lesser symmetry states occur in characteristic manners. Extension of this argument seems direct, particularly to groupoid structures. A full-bore mathematical treatment of these and related matters can be found in Glazebrook and Wallace (2009a, b).

1.1 The Topological Hypothesis and the topology of the genetic code

The relation between phase transitions in physical systems and topological changes has again become a central topic of research. Franzosi and Pettini (2004) and Pettini (2007), for example, argue that the standard way of studying

phase transition in physical systems is to consider how the values of thermodynamic observables, obtained in laboratory experiments, vary with temperature, volume, or an external field, and then to associate the experimentally observed discontinuities at a phase transition to the appearance of some kind of singularity entailing a loss of analyticity. However, they wonder whether this is the ultimate level of mathematical understanding of phase transition phenomena, or if some reduction to a more basic level is possible. Their theorem says that nonanalyticity is the 'shadow' of a more fundamental phenomenon occurring in configuration space: *a topology change*. Their theorem means that a topology change in a particular energy manifold is a *necessary* condition for a phase transition to take place. The topology changes are described within the framework of Morse theory through Morse-theoretic attachment handles. The converse of the Franzosi/Pettini theorem is not true. There is not a one-to-one correspondence between phase transitions and topology changes, and an open problem is that of *sufficiency* conditions, that is, to determine which kinds of topology changes can entail the appearance of a phase transition. A summary of standard material on Morse Theory is presented in a mathematical appendix to R. Wallace and R.G. Wallace, (2008).

Tlusty (2007) employs something of the kind in his analysis of the genetic code. He assumes that two codons are most likely to be confused if all their letters except for one agree. He then constructs a graph having an edge between codons if and only if they fit this confusion criterion. The resulting graph, he claims, is natural for considering the impact of translation errors or mutations because such errors almost always involve a single letter difference, hence a movement along an edge to a neighboring vertex:

> The topology of a graph is characterized by its genus γ, the minimal number of holes required for a surface to embed the graph such that no two edges cross. The more connected that a graph is the more holes are required for its minimal embedding... [T]he highly interconnected 64-codon graph is embedded in a holey, $\gamma = 41$ surface. The genus is somewhat reduced to $\gamma = 25$ if we consider only 48 effective codons.

From the perspective of Pettini (2007), a free energy construct serves as a Morse function whose critical points characterize just such a topology.

Tlusty (2007) concludes, similarly, that the topology of the code sets an upper limit to the number of low modes – critical points – of his free energy-analog functional, which 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, Tlusty points out, is precisely the well-known topological coloring problem, determined by Heawood's formula (Ringel and Young, 1968):

$$chr(\gamma) = int(\frac{1}{2}(7 + \sqrt{1 + 48\gamma})),$$

(3)

where $chr(\gamma)$ is the number of color domains of a surface with genus γ, and $int(x)$ is the integer value of x. We reproduce part of Tlusty's Table 1 that shows the topological limit to the number of amino acids for different codes.

Code	Number of Codons	Maximal # of amino acids
4-base singlets	4	4
3-base doublets	9	7
4-base doublets	16	11
16 codons	32	16
48 codons	48	20
4-base triplets	64	25

It is important to recognize that this is a fundamental topological decomposition, to which 'free energy' functionals are to be fit.

Tlusty concludes

> [This] suggests a pathway for the evolution of the present-day code from simpler codes, driven by the increasing accuracy of improving translation machinery. Early translation machinery corresponds to smaller graphs since indiscernible codons are described by the same vertex. As the accuracy improves these codons become discernible and the corresponding vertex splits. This gives rise to a larger graph that can accommodate more amino acids... [P]resent-day translation machinery with a four-letter code and 48-64 codons (no discrimination between U and C in the third position) gave rise to 20-25 amino acids. One may think of future improvement that will remove the ambiguity in the third position (64 discernible codons). This is predicted to enable stable expansion of the code up to 25 amino acids.

From the perspective of Glazebrook and Wallace (2009b), similar results can probably be obtained by using ideas of network holonomy applied to simplicial complexes and triangulations, that is, network phase transitions on graphs via connections and groupoids.

Here we will reconsider the evolutionary trajectories of codes in the context of available metabolic free energy, taking the perspective of Wallace (2009) and

Wallace and Wallace (2008) that punctuated ecosystem resilience transitions (e.g., Holling, 1973; Gunderson, 2000) can entrain evolutionary process, and that the availability of metabolic free energy is central to the evolution of complex phenomena of biological communication.

We begin with a restatement of a few central ideas from information theory, in particular the essential details of the Rate Distortion Theorem.

2 Some facts from information theory

The existence of a code implies the existence of an information source using that code, and the behavior of such sources is constrained by the asymptotic limit theorems of information theory. That is, the interaction between biological subsystems associated with a code can be formally restated in communication theory terms. Wallace and Wallace (2008, 2009) use an elaborate cognitive paradigm for gene expression to infer such information sources, i.e., cognition implies 'language', in a large sense, but our focus here on codes neatly finesses the argument.

Here we think of the machinery listing a sequence of codons as communicating with machinery that produces amino acids, and suppose we can compare what we is actually produced with what should have been produced, perhaps by a simple survival of the fittest selection mechanism, perhaps via some more sophisticated error-correcting systems.

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

$$b^n \rightarrow \hat{y}^n = \hat{y}_1, \hat{y}_2, ...$$

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

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

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

For continuous variates the squared error distortion is just

$$d(y, \hat{y}) = (y - \hat{y})^2.$$

There are many such possibilities. The distortion between *paths* y^n and $\hat{y}^n$ is defined as

$$d(y^n, \hat{y}^n) \equiv \frac{1}{n} \sum_{j=1}^{n} d(y_j, \hat{y}_j).$$

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

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

$$p(y^n), p(\hat{y}^n), p(y^n, \hat{y}^n), p(y^n|\hat{y}^n).$$

The average distortion is defined as

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

(4)

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

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

(5)

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

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

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

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

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

(6)

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

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

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

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

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

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

$$R(D) = 0, D > \sigma^2.$$

(7)

For the 'natural' channel that seems to describe compression of real images (e.g., Sarshar and Wu, 2007)

$$R(D) = \frac{\beta}{D^\alpha}$$

(8)

with $\alpha \approx 1$.

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

$$H[X] \leq C,$$

(9)

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

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

(10)

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

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

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

$$\begin{aligned} H[X] &= \lim_{n\to\infty} \frac{\log[N(n)]}{n} \\ &= \lim_{n\to\infty} H(X_n|X_0, ..., X_{n-1}) \\ &= \lim_{n\to\infty} \frac{H(X_0, ..., X_n)}{n+1}, \end{aligned}$$

(11)

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

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

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

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

with $\hat{Z} = Z^{-1/\beta}$. The latter expression is formally similar to the first part of equation (11), a circumstance having deep implications: Feynman (2000) describes in great detail how information and free energy have an inherent duality. Feynman, in fact, defines information precisely as the free energy needed to erase a message. The argument is surprisingly direct (e.g., Bennett, 1988), and for very simple systems it is easy to design a small (idealized) machine that turns the information within a message directly into usable work – free energy. Information is a form of free energy and the construction and transmission of information within living things consumes metabolic free energy, with nearly inevitable losses via the second law of thermodynamics. If there are limits on available metabolic free energy there will necessarily be limits on the ability of living things to process information.

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

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

In sum, if a biological system characterized by H_1 has a richer and more complicated internal communication structure than one characterized by H_2, then necessarily $H_1 > H_2$ and system 1 represents a more energetic process than system 2.

3 Internal structure of the genetic code

Ash's comment leads directly to a model in which the average distortion between codon stream and amino acid stream becomes a dominant force. This is via the relation between the Rate Distortion Function and free energy. The simplest model finds codons generated by a black box information source whose source uncertainty is constrained by the richness of the coding scheme of Tlusty's Table 1, summarized above. In general we may expect more complex codes to be associated with higher information source uncertainties, i.e., the ability to 'say' more in less time, using a more complicated coding scheme. Suppose there are n possible coding schemes. The simplest approach is to assume, that for a given rate distortion function and distortion measure, $R(D)$ serves much as an external temperature bath for the possible distribution of information sources, the set $\{H_1, ..., H_n\}$. That is, low distortion, represented by a high rate of transmission of information between codon machine and amino acid machine, permits more complicated coding schemes according to the classic formula

$$Pr[H_j] = \frac{\exp[-H_j/\kappa R(D)]}{\sum_{i=1}^{n} \exp[-H_i/\kappa R(D)]},$$

(12)

where $Pr[H_j]$ is the probability of coding scheme j having information source uncertainty H_j.

The free energy Morse Function associated with this probability is

$$F_R = -\kappa R(D) \log[\sum_{i=1}^{n} \exp[-H_i/\kappa R(D)]].$$

(13)

Applying Pettini's topological hypothesis to the Morse Function F_R generates topological transitions in codon graph structure as the 'temperature' $R(D)$ increases, i.e., as the average distortion D decreases, via the inherent convexity of the Rate Distortion Function. That is, as channel capacity connecting codon machines with amino acid machines increases, more complex coding schemes can emerge in a punctuated manner.

What, then, drives $R(D)$, as this drives, in turn, punctuated changes in the genetic code? We present a series of three increasingly complicated perspectives.

4 Rate Distortion Dynamics

4.1 A crude model

The direct approach is to impose the nonequilibrium generalized Onsager model of the Introduction: Living biological structures are nonequilibrium systems, and the flow of metabolic free energy is required for physiological function. Since $R(D)$ is a free energy analog, then, defining $S_R = R(D) - DdR(D)/dD$ as an entropy-analog leads to

$$dD/dt = -\mu dS_R/dD + F(D,t).$$

Taking $F(D,t)$ as proportional to the available metabolic energy density M, for a Gaussian channel gives

$$dD/dt = \mu/2D - \kappa M,$$

(14)

having the equilibrium value, at $dD/dt = 0$,

$$D_{eq} = \frac{\mu}{2\kappa M}.$$

That is,

$$R \propto \log[M]$$

Thus, in this model, the internal temperature affecting codon topology – the Rate Distortion Function – grows as the log of the available metabolic free energy density.

For the 'natural' channel, taking $\alpha = 1$,

$$R \propto \sqrt{M}.$$

4.2 A less crude model

A more subtle analysis is to suppose there may be a spectrum of possible evolutionary or other machineries imposing penalties for high distortion in the translation of codons to amino acids, or, inversely, that there may be a distribution of possible channel capacities enabling accurate translation. We iterate the model of equation (13), supposing that the probability of a particular Rate Distortion value is determined by something like equation (12), so that we may define a Rate Distortion partition function as

$$\mathcal{Z}_R[M] \equiv \int_{D_{min}}^{D_{max}} \exp[-R(D)/\kappa M] dD,$$

(15)

where M is the available metabolic free energy density, a function of the embedding environment. Then the 'Rate Distortion Free energy' becomes

$$\mathcal{F}_R[M] = -\kappa M \log[\mathcal{Z}_R[M]],$$

(16)

and we can apply Pettini's topological hypothesis for punctuated changes in the topology underlying Morse Function $\mathcal{F}_R$. These entrain punctuated changes in codon topology through a cascade mechanism, a double phase change, driven by punctuated ecosystem resilience transitions in available metabolic energy density. That is, sudden shifts in ecosystem availability of metabolic energy trigger, first, transitions in $\mathcal{F}_R$, and then, through changes in R, punctuated changes in codon topology.

While Occam's Razor may seem to favor the simple model of Section 4.1, one need only remember the more recondite parasite life cycles for a class of evolutionary counterexamples. Fitness is contingent and context-driven, and path-dependent evolutionary process need not conform to our cultural aesthetics.

Indeed, the image of complicated parasite life cycle dynamics leads to a more detailed examination of the transition mechanism.

4.3 Coevolution

R. Wallace and R.G. Wallace (2008) used information theory methods to reconsider Eigen's paradox, and in particular the interaction between two prebiotic vesicles under mutual recursion similar to Van Valen's (1973) famous Red Queen. Application to the dynamics of the Rate Distortion function driving punctuated changes in the genetic code is surprisingly direct.

Here the two 'vesicles' are the machinery of the genetic code and that of the epigenetic regulatory machinery that translates the gene to protein. Figure 1, adapted from figure 1.8 of Smulevich and Dougherty (2007), shows the two systems intersecting, at the point where messenger RNA carries a gene codon to

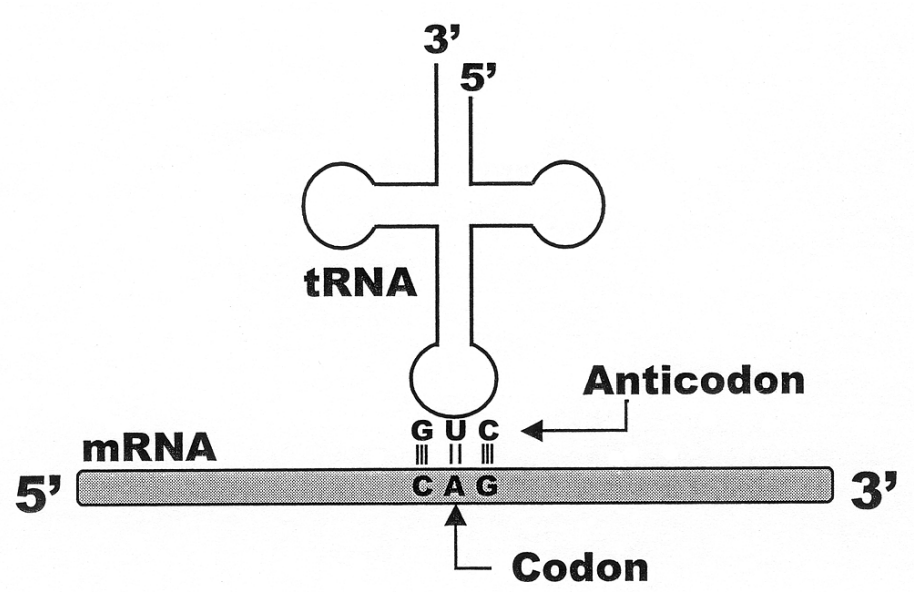

Figure 1: Where the DNA World meets the RNA World in modern protein synthesis: The anticodon at one end of a tRNA molecule binds to its complementary codon in mRNA derived directly from the genome. Adapted from fig. 1.8 of Shmulevich and Dougherty, (2007).

meet the anticodon of translational RNA, that carries, on its back, the precursor to the appropriate amino acid.

Recall that the Rate Distortion Function is defined by the minimization of a mutual information

$$I(Y, \hat{Y}) = H(Y) + H(\hat{Y}) - H(Y, \hat{Y}) \equiv H_Y + H_{\hat{Y}} - H_{Y,\hat{Y}}.$$

We are particularly interested in the magnitudes of $H(Y)$ and $H(\hat{Y})$, supposing that increases in both will, generally, lead to increases in $R(D)$, in spite of the negative joint uncertainty term.

The essential point is to view the genetic and translational (and, in another context, epigenetic) machinery as being each other's principal environments, similar to, or, indeed, taken as, prebiotic interacting vesicles. Then we write

$$H_Y = H_Y[K], K = 1/H_{\hat{Y}},$$

(17)

and similarly for $H_{\hat{Y}}$. That is, both H_Y and $H_{\hat{Y}}$ are parameterized by the other's inverse. That is, increase or decline in the source uncertainty of one system leads to increase or decline in the source uncertainty of the other. The richness of the two information sources is closely linked.

Start at the right of the lower graph for H in figure 2, the source uncertainty of one system, but to the left of the critical point K_C that indicates collapse of the interaction between the 'vesicles' analogous to Eigen's error catastrophe, following the model of R. Wallace and R.G. Wallace (2008). Assume $H_{\hat{Y}}$ increases, so K declines., and thus H_Y increases, walking up the lower curve of figure 2 from the right, so that the richness of the first vesicle's internal language increases.

Increase in H_Y leads, in turn, to a decline in $K_Y = 1/H_Y$, and triggers an increase of $H_{\hat{Y}}$, whose increase leads to a further increase of H_Y, and vice versa. This is the Red Queen, taking the system from the right of figure 2 to the left, up the lower curve as the two vesicles interact.

Now recognize the possibility of a reverse dynamic as well, driven by the gradient of the disorder $S = H - KdH/dK$ that, in the absence of a Red Queen, would simply drive the system toward the minimum energy critical point for this system.

Thus the system has two quasi-stable limit points, a low energy solution near the error limit phase transition point, analogous to Eigen's error threshold, and a high energy solution near to, but never at, the zero error limit, depending on the availability of sufficient metabolic energy to the system. Absent a relatively energetic metabolic source, low error rate translation of the genetic code would

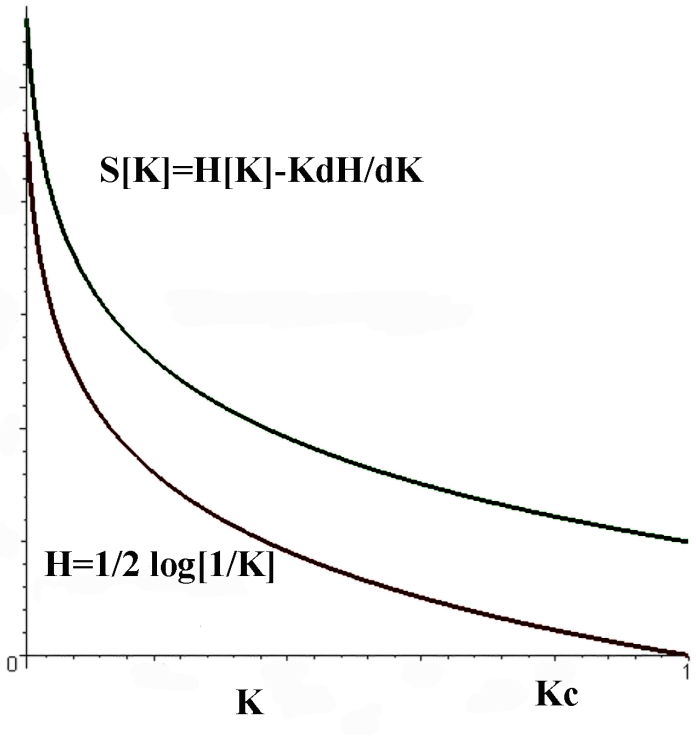

Figure 2: A general curve for source uncertainty $H[Y]$ – measuring genetic code language richness – as a function of an inverse temperature parameter $K = 1/H[\hat{Y}]$. To the right of the critical point K_C the system collapses in an analog to Eigen's error catastrophe. Since K is an inverse source uncertainty for the epigenetic information source $\hat{Y}$, a Red Queen dynamic can become enabled, driving the system strongly to the left. No intermediate points are asymptotically stable. To the right of the critical point K_C the system fails catastrophically. Thus there are two quasi-stable points, a low energy solution near the error limit phase transition point, and a high energy state nearer to, but never at, the zero error limit, determined by the availability of metabolic free energy.

be impossible, according to the model. Adapting the arguments of R. Wallace and R.G. Wallace (2008), this suggests that some major, large scale, ecosystem transformation in metabolic energy availability was a necessary condition for low error rate genetic code translation to protein structures.

5 Discussion and conclusions

Several points emerge from this analysis. Not only have the codons undergone evolutionary process, but so too has the translational machinery, as logically implied by the complementary anticodon structure: Recently Sun and Caetano-Anolles (2008) claimed evidence for deep evolutionary patterns embedded in tRNA phylogenies, calculated from trees reconstructed from analyses of data from several hundred tRNA molecules. They argue that an observed lack of correlation between ancestries of amino acid charging and encoding indicates the separate discoveries of these functions reflects independent histories of recruitment. These histories were, in their view, probably curbed by co-options and important take-overs during early diversification of the living world. That is, disjoint evolutionary patterns were associated with evolution of amino acid specificity and codon identity, indicating that co-options and take-overs embedded perhaps in horizontal gene transfer affected differently the amino acid charging and codon identity functions. These results, they claim, support a strand symmetric ancient world in which tRNA had both a genetic and a functional role (Rodin and Rodin, 2008).

Clearly, 'co-options' and 'take-overs' are, perhaps, most easily explained as products of a prebiotic serial endosymbiosis, in our model instantiated by a Red Queen between significantly, perhaps radically, different precursor chemical systems.

Thus our coevolution argument in this context is not new, although the particular mathematical approach is innovative.

Indeed, Witzany (2009) also takes a broadly similar 'language' perspective. In that paper he reviews a massive literature, arguing that not only rRNA, but also tRNA and the processing of the primary transcript into the pre-mRNA and the mature mRNA seem to be remnants of viral infection events that did not kill their host, but transferred phenotypic competences to their host and changed both the genetic identity of the host organism and the identity of the former infectious viral swarms. His 'biocommunication' viewpoint investigates both communication within and among cells, tissues, organs and organisms as sign-mediated interactions, and nucleotide sequences as code, that is, language-like text. Thus editing genetic text sequences requires, similar to the signaling codes between cells, tissues, and organs, biotic agents that are competent in correct sign use. Otherwise, neither communication processes nor nucleotide sequence generation or recombination can function. From his perspective, DNA is not only an information storing archive, but a life habitat for nucleic acid language-using RNA agents of viral or subviral descent able to carry out almost error-free editing of nucleotide sequences according to systematic rules of grammar and

syntax.

Here we have outlined a formal modeling strategy for this process, using the asymptotic limit theorems of information theory.

In sum, the punctuated topology of the genetic code, as examined by Tlusty (2007), implies, in turn, the possibility of a number of punctuated shifts in the availability of metabolic free energy that may have been as fundamental as the transition from anaerobic to aerobic metabolism (e.g., Wallace, 2009). Each such transition would have enabled higher channel capacities in the communication between interacting biological vesicles, in a large sense, and each would probably have initiated new rounds of serial endosymbiosis (e.g., Villarreal and Witzany, 2009; Witzany, 2009), among other things. Tables 1 and 2 of Canfield et al. (2006) display a considerable range of feasible electron donors and receptors available to early anaerobic metabolisms, and the ecosystems that could have been based on them. Other possibilities include the development of systems for the storage of energy to be released during reproduction: think 'seeds'. In any event, transitions to higher energy metabolic systems would have, according to our model, been associated with punctuated transitions to more complex genetic codes. By the time of the aerobic transition, however, the code may, after reaching a near-maximal error robustness, have become nearly locked-in by path dependence, although some 20 subsequent slight variants have been recognized (Koonin and Novozhilov, 2009, and references therein).

The search for evidence of such a sequence of 'preaerobic' metabolic free energy transitions is, of course, fraught with difficulties.

6 References

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

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

Canfield, D., M. Rosing, and C. Bjerrum, 2006, Early anaerobic metabolisms, *Philosophical Transactions of the Royal Society, B*, 351:1819-1836.

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

Dembo, A., and O. Zeitouni, 1998, *Large Deviations and Applications*, 2nd. edn., Springer, NY.

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

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

Franzosi, R., and M. Pettini, 2004, Theorem on the origin of phase transitions, *Physical Review Letters*, 92:060601.

Glazebrook, J.F., and R. Wallace, 2009a, Small worlds and Red Queens in the Global Workspace: An information-theoretic approach, *Cognitive Systems Research*, 10:333-365.

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

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

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

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

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

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

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

Ringel, G., and J. Youngs, 1968, Solutions of the Heawood map-coloring problem, *Proceedings of the National Academy of Sciences*, 60:438-445.

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

Rodin, S., and A. Rodin, 2008, On the origin of the genetic code: signatures of its primordial complementarity in tRNAs and aminoacyl-tRNA synthetases, *Heredity*, 100:341-355.

Rose, K., 1998, Deterministic annealing for clustering, compression, classification, regression, and related optimization problems, *Proceedings of the IEEE*, 86:2210-2239.

Sarshar, N., and X. Wu, 2007, On rate-distortion models for natural images and wavelet coding performance, *IEEE Transactions on Image Processing*, 3:87-93.

Sun, F., and G. Ceataeno-Anolles, 2008, Evolutionary patterns in the sequence and structure of transfer RNA: a window into early translation and the genetic code, *PLOSone*, 3(7):32799.

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

Van Valen, L., 1973, A new evolutionary law, *Evolutionary Theory*, 1:1-30.

Villarreal, L., and G. Witzany, 2009, Viruses are essential agents within the roots and stem of the tree of life, *Journal of Theoretical Biology*,
doi:10.1016/j.jtbi.2009.10.014.

Wallace, R., and R. G. Wallace, 2008, On the spectrum of prebiotic chemical systems: an information-theoretic treatment of Eigen's Paradox, *Origins of Life and Evolution of Bioshperes*, 38:419-455.

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

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

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

Witzany, G., 2009, Noncoding RNAs: persistent viral agents as modular tools for cellular needs, *Annals of the New York Academy of Sciences*, 1178:244-267.