Biochemical 'Cambrian events': on the evolution of biological codes

Rodrick Wallace

Tlusty's topological analysis of the genetic code suggests ecosystem changes in available metabolic free energy that predated the aerobic transition enabled a punctuated sequence of increasingly complex genetic codes and protein translators, protein folding codes, and monosaccharide cell-surface codes. These coevolved via various 'Cambrian explosions' until, very early on, the ancestors of the present narrow spectrum of such biological machineries outcompeted other codings and became evolutionarily locked in at surprisingly modest levels of fitness likely reflecting a modest embedding metabolic free energy ecology. Thus biochemical 'Cambrian singularities' must have occurred at different scales and levels of organization on Earth, with competition or chance-selected outcomes frozen at a far earlier period than the physical bauplan Cambrian explosion. Beyond codes, other examples might include explosive variations in mechanisms of photosynthesis, and the subsequent manifold oxygen metabolisms. Intermediate between Cambrian bauplan and genetic code, variants remain today, even after evolutionary pruning, often protected in specialized ecological niches. It is even possible to interpret the most basic biological 'coding' from this perspective, i.e., homochirality. This suggests that, under other astrobiological ecologies, different spectra of biochemical codes and other mechanisms may survive in appropriate niches.

PeerJ PrePrints | http://dx.doi.org/10.7287/peerj.preprints.308v1 | CC-BY 4.0 Open Access | received: 28 Mar 2014, published: 28 Mar 2014

Version 11

Biochemical 'Cambrian events': on the evolution of biological codes

Rodrick Wallace
Division of Epidemiology
The New York State Psychiatric Institute *

March 27, 2014

Abstract

Tlusty's topological analysis of the genetic code suggests ecosystem changes in available metabolic free energy that predated the aerobic transition enabled a punctuated sequence of increasingly complex genetic codes and protein translators, protein folding codes, and monosaccharide cell-surface codes. These coevolved via various 'Cambrian explosions' until, very early on, the ancestors of the present narrow spectrum of such biological machineries outcompeted other codings and became evolutionarily locked in at surprisingly modest levels of fitness likely reflecting a modest embedding metabolic free energy ecology. Thus biochemical 'Cambrian singularities' must have occurred at different scales and levels of organization on Earth, with competition or chance-selected outcomes frozen at a far earlier period than the physical bauplan Cambrian explosion. Beyond codes, other examples might include explosive variations in mechanisms of photosynthesis, and the subsequent manifold oxygen metabolisms. Intermediate between Cambrian bauplan and genetic code, variants remain today, even after evolutionary pruning, often protected in specialized ecological niches. It is even possible to interpret the most basic biological 'coding' from this perspective, i.e., homochirality. This suggests that, under other astrobiological ecologies, different spectra of biochemical codes and other mechanisms may survive in appropriate niches.

Key Words: chemical evolution; evolutionary dynamics; information theory; punctuated equilibrium

1 Introduction

Wallace (2014a) argues that 'Cambrian explosions' are standard features of blind evolutionary process, representing outliers in the ongoing routine of evolutionary punctuated equilibrium. Most such explosions, however, will be severely pruned by selection and chance extinction. That work suggested, in passing, that the evolution of the genetic code, involving the transmission of information between codon machinery and amino acid machinery, is likely to have undergone just such an 'explosion' as significant levels of chemical free energy became available to metabolic process. Here, we make that argument explicit, and extend it to other biological codes.

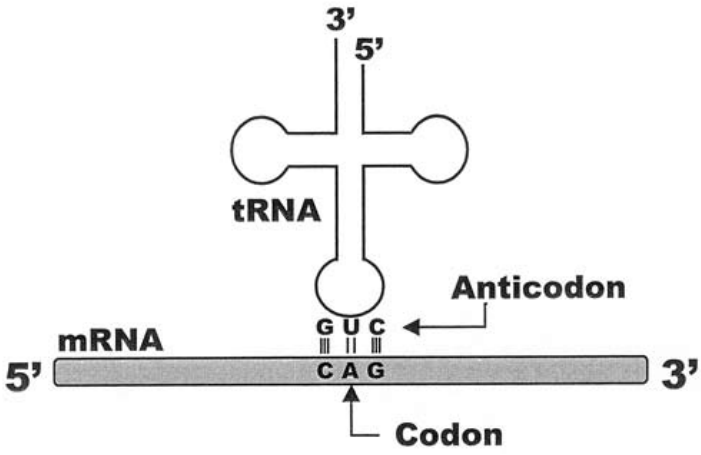

Figure 1: Adapted from fig. 1.8 of Shmulevich and Dougherty (2007). Modern protein synthesis; the anticodon at one end of a tRNA molecule binds to its complementary codon in mRNA derived directly from the genome. Sequence-to-sequence translation is not highly parallel, in this model, and the process can be characterized in terms of the Shannon uncertainty in the transmission of information between codon machinery and amino acid machinery.

Figure 1, adapted from Shmulevich and Dougherty (2007), gives a schematic of the present highly evolved system relating code to protein component. 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. Sequence-to-sequence translation is not highly parallel, in this model, and the process can be characterized in terms of the Shannon uncertainty in the transmission of information between codon machinery and amino acid machinery.

To paraphrase Tlusty (2007, 2008), the genetic code

*Correspondence: R. Wallace, Box 47, NYSPI, 1051 Riverside Dr., NY, NY, 10032. Wallace@nyspi.columbia.edu, rodrick.wallace@gmail.com

emerges as a transition in a noisy information channel, using a Rate Distortion Theorem argument: the optimal code is described by the minimum of a 'free energy'-like functional, which leads naturally to the possibility of describing the code's emergence as a transition akin to a phase transition in statistical physics (Rose, 1998). This is essentially a Morse function argument, in the sense of Pettini (2007) and Matsumoto (2002).

More specifically, as Tlusty (2007) puts it,

> To discuss the topology of errors we portray the codon space as a graph whose verticies are the codons... Two codons... are linked by an edge if they are likely to be confused by misreading... We assume that two codons are most likely to be confused if all their letters except for one agree and therefore draw an edge between them. The resulting graph 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.
>
> 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.

Tlusty further concludes that the topology of the code sets an upper limit to the number of low modes – critical points – of his 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.

Tlusty states:

> The maximum [of the functional] determines a single contiguous domain where a certain amino acid is encoded... Thus every mode corresponds to an amino acid and the number of modes is the number of amino acids. This compact organization is advantageous because misreading of one codon as another codon within the same domain has no deleterious impact. For example, if the code has two amino acids, it is evident that the error-load of an arrangement where there are two large contiguous regions, each coding for a different amino acid, is much smaller than a 'checkerboard' arrangement of the amino acids.

This, Tlusty (2010) points out, is analogous to the well-known topological coloring problem: "in the coding problem one desires maximal similarity in the colors of neighboring 'countries', while in the coloring problem one must color neighboring countries by different colors". After some development (Tlusty, 2008), the number of possible amino acids in this scheme is determined by Heawood's formula (Ringel and Young, 1968):

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

where $chr(\gamma)$ is the number of color domains of a surface with genus γ, and $Int(x)$ is the integer value of x. The genus is, roughly speaking, the number of holes in an orientable manifold.

However, from Morse Theory (Matsumoto, 2002):

$$\gamma = 1 - \frac{1}{2}\chi \quad (2)$$

where χ is the Euler characteristic of the underlying topological manifold. For Tlusty's system, $\chi = V - E + F$ where V is the number of code network vertices, E the number of network edges, and F the number of enclosed faces.

We reproduce part of Table 1 of Tlusty (2007), showing the topological limit to the number of amino acids for different codes:

Code	**# Codons**	**Max. # AA's**
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

This is the fundamental topological decomposition – representing an increasing gradient in overall symmetries driven by 'holes' in the underlying topological structure – to which Morse-theoretic 'free energy' functionals are to be fit.

Note, however, that, while the scheme limits the basic code bauplan, for the current coding a simple combinatorial argument shows there are 10^{84} possible alternative code tables if each of the 20 amino acids and the stop signal are assigned at least one codon. Smaller, but still astronomical, numbers can be associated with the less complicated codes, permitting a later statistical mechanics-like model driven by available metabolic free energy.

Tlusty (2007) 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.

The underlying picture is that of phase transitions in physical systems. Following Landau's group symmetry shifting arguments (Landau and Lifshitz, 2007; Pettini, 2007), higher temperatures enable higher system symmetries, and, as temperature changes, punctuated shifts to different symmetry states that occur in characteristic manners. Extension of this argument in terms of information transmission between codons and proteins, in the context of metabolic energy measures, seems direct, particularly involving the groupoids constructed by the disjoint union of the homology groups representing the different coding topologies that Tlusty identifies. A more complete mathematical treatment of some of these and related matters can be found in Glazebrook and Wallace (2009a, b). Here, the genus γ can be seen as a kind of inverse order parameter, representing the increase in symmetry, in a large sense.

Indeed, one can obtain, for deterministic-but-for error codes, something much like Landau's group theoretic construct by noting that the fundamental group of a closed, orientable surface of genus γ is the quotient of the free group on the 2γ generators $a_1, ..., a_\gamma, b_1, ..., b_\gamma$ by the normal subgroup generated by the product of the commutators

$$a_1 b_1 a_1^{-1} b_1^{-1} ... a_\gamma b_\gamma a_\gamma^{-1} b_\gamma^{-1}$$

This is a standard construction (e.g., Lee 2000).

2 Extending the model

Tlusty's method can also be applied to other biochemical phenomena, for example large-scale globular protein folding and information transmission at the cell surface (Wallace, 2010, 2012b). Equation (1), most simply, produces the table

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

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

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

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

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

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

Wallace (2012d) generalizes these methods to intrinsically disordered proteins using nonrigid molecule symmetry arguments, via semidirect and wreath products over appropriate sets of finite and/or compact groups.

The underlying idea seems also applicable to the generation of the 12 elementary monosaccharides that are the basis of the 'glycan code' for information transmission at the mammalian cell surface (Wallace, 2012b), involving a code network with six or seven topological holes. However, these numbers vary considerably across eukaryotes, prokaryotes and archaea, but seem to broadly fit Heawood's classification, although the actual business of signaling involves complex 'kelp fronds' built from these 12 base elements, resulting in a composite 'code' having a literally astronomical symmetry structure that must be treated using other methods.

That is, while genome and protein folding machineries, and perhaps the simplest glycan monosaccharide code, are template driven, i.e., deterministic-but-for-errors, glycome/lectin interactions, carrying much more information, require deeper study. Glycome signaling structures at the cellular surface are finger-like projections of a characteristic number of basic monosaccharides and sidechains whose expression is contingent, not deterministic, and dependent on a shifting, tunable, cascade of processes. Cummings (2009) estimates there are from 5,000 to 7,500 glycan determinants that are the fundamental 'amino acids' of cell surface structure. Applying Tlusty's analysis suggests a code network manifold having between 2 and 5 million holes. This is truly a different world that must be regulated by an added layer of highly sophisticated cognitive machinery (Wallace, 2012b), a matter to which we will return.

We will reconsider, in some generality, the evolutionary trajectories of biological codes in the context of intensive measures of available metabolic free energy, taking the perspective that the availability of metabolic free energy is central to the evolution of complex phenomena of biological communication. That is, the 'temperature' analog is the chemical free energy available to early anaerobic metabolisms, in the

sense of Canfield et al. (2006). For example, the proposed hydrogen/sulfur reaction

$$H_2 + \frac{1}{2}S_2 \rightarrow H_2S$$

produces something like $M = 21$ KJ/mol, while the aerobic reaction

$$H_2 + \frac{1}{2}O_2 \rightarrow H_2O$$

produces about $M = 241$ KJ/mol, more than an order of magnitude greater. The genetic code, however, was locked in by evolutionary path dependence well before oxygen became widely available for metabolic process.

Canfield et al. (2006) examine a considerable range of possible electron donors and acceptors available to early anaerobic metabolisms on earth.

We begin with a summary of ideas from Wallace (2014a).

3 Information theory

A genetic code, translating codons into proteins, or other such biological 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. Thus, 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 the focus here on codes condenses the argument because a code directly implies existence of an information source using it.

Here we think of the machinery listing a sequence of codons as communicating with machinery that produces amino acids, allowing definition of an information source embedded in an environment whose regularities themselves imply the existence of an information source.

Following Wallace (2014a), assume there are n possible code 'species' interacting with an embedding environment represented by an information source Z. The processes associated with each code species i are represented as information sources X_i. These information sources undergo a 'coevolutionary' interaction in the sense of Champagnat et al. (2006), producing a joint information source uncertainty (Cover and Thomas, 2006) for the full system as

$$H(X_1, ..., X_n, Z) \tag{3}$$

Feynman's (2000) insight that information is a form of free energy allows definition of an entropy-analog as

$$S \equiv H - Q_j \sum_j \partial H / \partial Q_j \tag{4}$$

The Q_i are taken as driving parameters that may include, but are not limited to, the Shannon uncertainties of the underlying information sources.

Again following Wallace (2014a), we can characterize the dynamics of the system in terms of Onsager-like nonequilibrium thermodynamics in the gradients of S as the set of stochastic differential equations (de Groot and Mazur, 1984),

$$dQ_t^i = L_i(\partial S/\partial Q^1 ... \partial S/\partial Q^m, t)dt + \sum_k \sigma_k^i(\partial S/\partial Q^1 ... \partial S/\partial Q^m, t)dB_k \tag{5}$$

where the B_k represent noise terms having particular forms of quadratic variation. See Protter (1990) or other standard references on stochastic differential equations for details.

This can be more simply written as

$$dQ_t^i = L_i(\mathbf{Q}, t)dt + \sum_k \sigma_k^i(\mathbf{Q}, t)dB_k \tag{6}$$

where $\mathbf{Q} \equiv (Q^1, ..., Q^m)$.

Following the arguments of Champagnat et al. (2006), this is a coevolutionary structure, where fundamental dynamics are determined by component interactions:

1. Setting the expectation of equations (5) equal to zero and solving for stationary points gives attractor states since the noise terms preclude unstable equilibria. These are analogous to the evolutionarily stable states of evolutionary game theory.
2. This system may, however, converge to limit cycle or pseudorandom 'strange attractor' behaviors similar to thrashing in which the system seems to chase its tail endlessly within a limited venue – the 'Red Queen'.
3. What is 'converged' to in any case is not a simple state or limit cycle of states. Rather it is an equivalence class, or set of them, of highly dynamic information sources coupled by through crosstalk and other mutual interactions. Thus 'stability' in this structure represents particular patterns of ongoing dynamics rather than some identifiable static configuration, that is, at best, a nonequilibrium steady state.
4. Applying Ito's chain rule for stochastic differential equations to the $(Q_t^j)^2$ and taking expectations allows calculation of variances. These may depend very powerfully on a system's defining structural constants, leading to significant instabilities (Khasminskii, 2012).

4 Large deviations: iterating the model

As Champagnat et al. (2006) note, shifts between the nonequilibrium steady states of a coevolutionary system can be addressed by the large deviations formalism. The dynamics of drift away from trajectories predicted by the canonical equation can be investigated by considering the asymptotic of the probability of 'rare events' for the sample paths of the diffusion.

'Rare events' are the diffusion paths drifting far away from the direct solutions of the canonical equation. The probability of such rare events is governed by a large deviation principle, driven by a 'rate function' $\mathcal{I}$ that can be expressed in terms of the parameters of the diffusion.

This result can be used to study long-time behavior of the diffusion process when there are multiple attractive singularities. Under proper conditions, the most likely path followed by the diffusion when exiting a basin of attraction is the one minimizing the rate function $\mathcal{I}$ over all the appropriate trajectories.

An essential fact of large deviations theory, however, is that the rate function $\mathcal{I}$ almost always has the canonical form

$$\mathcal{I} = -\sum_j P_j \log(P_j) \tag{7}$$

for some probability distribution, i.e., the uncertainty of an information source (Dembo and Zeitouni, 1998).

The argument directly complements equation (5), now seen as subject to large deviations that can themselves be described as the output of an information source L_D defining $\mathcal{I}$, driving or defining Q^j-parameters that can trigger punctuated shifts between quasi-stable nonequilibrium steady states.

Not all large deviations are possible: only those consistent with the high probability paths defined by the information source L_D will take place.

Recall from the Shannon-McMillan Theorem (Khinchin, 1957) that the output streams of an information source can be divided into two sets, one very large that represents nonsense statements of vanishingly small probability, and one very small of high probability representing those statements consistent with the inherent 'grammar' and 'syntax' of the information source. For example, whatever higher-order multicellular evolution takes place, some equivalent of backbone and blood remains.

Thus we could now rewrite equation (3) as

$$H_L(X_1, ..., X_n, Z, L_D) \tag{8}$$

where we have explicitly incorporated the 'large deviations' information source L_D that defines high probability evolutionary excursions for this system.

Again carrying out the argument leading to equation (5), we arrive at another set of quasi-stable modes, but possibly very much changed in number; either branched outward in time by a wave of speciation or quasi-speciation, or decreased through a wave of extinction. Iterating the models backwards in time constitutes a cladistic or coalescent analysis.

5 Evolution of codes under relaxed path-dependence

Following the arguments of Wallace (2014a), in general, for current organisms, the number of nonequilibrium steady states available to the system defined by equation (5), or to its generalization via equation (7), will be quite small – indeed, at most a handful – a consequence of code lock-in by path dependent evolutionary process. The same cannot be said, however, for earlier species or quasi-species, to which can be applied more general methods that may represent key processes acting three or four billion years in the past.

Under such a relaxation assumption, the large deviations information source L_D is far less constrained, and there will be very many possible quasi-stable nonequilibrium steady states available for transition, analogous to an ensemble in statistical mechanics. Again, for the current genetic code, involving 20 possible amino acids, following the arguments of Tlusty's (2007) Table 1, there are some 10^{84} possible alternative codings. Similar arguments surround protein folding and monosaccharide 'codes'.

The metabolic free energy index in KJ/mol, which we write as M, can, from the arguments of the Introduction, then be interpreted as a kind of temperature measure so that higher values permit higher equivalence class groupoid symmetries. This leads to a relatively simple statistical mechanics analog built on the H_L of equation (7).

Define a pseudoprobability for quasi-stable mode j as

$$P_j = \frac{\exp[-H_L^j/\kappa M]}{\sum_i \exp[-H_L^i/\kappa M]} \tag{9}$$

where κ is a scaling constant, j runs over the nonequilibirum quasi-steady states, and M is an index of available metabolic reaction energy intensity, typically measured as KJ/mol.

Next, define a Morse Function F as

$$\exp[-F/\kappa M] \equiv \sum_i \exp[-H_L^i/\kappa M] \tag{10}$$

Apply Pettini's (2007) topological hypothesis to F. Then M is seen as a very general temperature-like intensity measure whose changes drive punctuated topological alterations in the underlying ecological and coevolutionary structures associated with the Morse Function F.

Such topological changes, following Pettini's arguments, can be far more general than indexed by the simple Landau-type critical point phase transition in an order parameter.

Thus, the results of Wallace (2012a), regarding the complexity of the genetic code, can be directly reframed in terms of available metabolic free energy intensity leading to equations (9) and (10). Then M is a measure of metabolic free energy intensity, and the H_L^j represent the Shannon uncertainties in the transmission of information between codon machinery and amino acid machinery.

Increasing M then leads to the possibility of more complex codes, i.e., those having higher measures of symmetry, in the Landau sense, as calculated by Tlusty's methods (i.e., more holes), until competition, selection, and chance extinction leading to evolutionary lock-in took place at a relatively low level of coding efficiency. Canfield et al. (2006) speculate that the most active early ecosystems were probably driven by the cycling of H_2 and Fe^{2+}, providing relatively low free energy intensities for metabolic process.

6 Discussion and conclusions

Marshall (2006) characterizes the 'Cambrian explosion' in animal physical bauplan that took place 500 myr ago as follows:

> With the advent of ecological interactions between macroscopic adults... especially... predation..., the number of needs each organism had to meet must have increased markedly: Now there were myriad predators to contend with, and a myriad number of ways to avoid them, which in turn led to more specialized ways of predation as different species developed different avoidance strategies, etc... The combinatoric richness already present in the Ediacaran genome was extracted through the richness of biotic interaction as the Cambrian 'explosion' unfolded...

Here we argue that, in analogous fashion, the availability of myriad biochemical electron donor/acceptor cycles according to the schema of Canfield et al. (2006) created a rich chemical ecology. The explosive combinatoric richness present in the possible variety of genetic and other biological codes was, however, in this case extracted through the richness of biotic interaction downward in scale by the efficiency, effectiveness, or chance survival, of a single genetic and a single protein folding code (and perhaps a small number of basic monosaccharide codes) that emerged at a modest level of available chemical free energy to persist as the present dominant set of codes.

Wallace (2011) makes a similar argument for the evolution of homochirality – the most basic biochemical 'code' of all – formally invoking the standard groupoid approach to stereochemistry in a thermodynamic context that likewise generalizes Landau's spontaneous symmetry breaking arguments. On Earth, limited metabolic free energy density may have served as a low temperature-analog to 'freeze' the system in the lowest energy state, i.e., the set of simplest homochiral transitive groupoids representing reproductive chemistries. These engaged in Darwinian competition until a single configuration survived. Subsequent path-dependent evolutionary process locked-in this initial condition. Astrobiological outcomes, in the presence of higher initial metabolic free energy densities, could well be considerably richer, for example, of mixed chirality. One result would be a complicated distribution of biological chirality across a statistically large sample of extraterrestrial stereochemistry, in marked contrast with recent published analyses predicting a racemic average.

Indeed, there may well have been a 'protein singularity'. Following the observations of Chou and Maggiora (1998), as described above, proteins have four major, and perhaps as many as another six less frequent, structural classifications. This suggests a Tlusty code error network that is, essentially, a large 'sphere', having one minor, and possibly as many as three more subminor attachment handles, according to Heawood's formula. These basic structures build a highly complicated 'protein world' that cannot be simply characterized.

The prebiotic 'amyloid world' of Maury (2009), in contrast, is built on a single β-sheet structure, having the simplest possible underlying 'genetic code': 101010101010..., where 1 represents a polar, and 0 a non-polar amino acid (Hecht et al., 2004). In full extent, however, amyloid structures follow an eight-fold steric zipper (Sawaya et al., 2007), suggesting a one-step higher 'amyloid code' having, in Tlusty's sense from the Heawood formula, a double donut conformation, i.e., an error code of genus 2.

Goldschmidt et al. (2010) describe the fundamental situation in these terms:

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

Clearly, effective chaperoning requires considerable metabolic energy, and increasing availability of it would provide a sufficient condition for a explosive expansion of protein structure beyond the amyloid world.

Not everything works as simply, however. Using Tlusty's approach to examine Cummings' (2009) 5,000–7,500 glycan determinant amino acid analogs would involve a code manifold that Wallace (2012b) calls a 'Grossly Complex Topological Object'. That is, sophisticated cell surface glycan/lectin codings have astronomical topological genus – our chosen symmetry index. This probably required fairly late development of photosynthesis, predation, and/or the aerobic transition, to provide sufficient free energy intensities allowing complex multicellular organisms based on efficient information transmission between cells and tissues. Indeed, Wallace (2012b) finds that the group symmetry methods applicable to deterministic-but-for-error coding must be significantly extended, involving groupoid symmetry models appropriate to an intermediate layer of cognitive regulatory machinery (Wallace 2012c, 2014b).

Even given such complications, the essential insight is that 'Cambrian singularities' at different scales and levels of organization are inherently path dependent, with the evolutionarily or chance-selected outcomes of basic biochemical explosions likely to be locked in at a far earlier period than physical bauplan explosions. Possible examples beyond biochemical codes include explosive evolutionary variations in photosynthesis, mechanisms of tissue oxidation, and – of course –

complex ecological relations like predation, mutualism, and symbiosis. Different forms of all remain today, even after evolutionary pruning. The many variants in the 'bauplan' of oxidizer metabolism are of special interest (e.g., Flood et al. 2011; Campbell and Farrell, 2012).

This suggests – perhaps tautologically – that, under different astrobiological free energy ecologies, far richer spectra of reproductive and other biochemical or biological codes and mechanisms of information transmission and regulation may survive, likely across characteristic ecosystem niches.

7 Acknowledgments

The author thanks Dr. D.N. Wallace for useful discussions.

References

Campbell, M., S. Farrell, 2012, Biochemistry, Seventh Edition, Harcourt, New York.

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

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

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

Cover, T., Thomas, J., 2006, Elements of Information Theory, Wiley, New York.

Cummings, R., 2009, The repertoire of glycan determinants in the human glycome, Molecular BioSystems, 5:1087-1104.

De Groot, S., P. Mazur, 1984, Nonequilibrium Thermodynamics, Dover, New York.

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

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

Flood, P., J. Harbinson, M. Aarts, 2011, Natural genetic variation in plant photosynthesis, Trends in Plant Science, 16:327-225.

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

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

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

Hecht, M., A. Das, A. Go, L. Aradely, Y. Wei, 2004, De novo proteins from designeated combinatorial libraries, Protein Science, 13:1711-1723.

Khasminskii, R., 2012, Stochastic Stability of Differential Equations, Second Edition, Springer, New York.

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

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

Lee, J., 2000, Introduction to Topological Manifolds, Graduate Texts in Mathematics Series, Springer, New York.

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

Marshall, C., 2006, Explaining the Cambrian 'explosion' of animals, Annual Reviews of Earth and Planetary Science, 34:355-384.

Matsumoto, Y., 2002, An Introduction to Morse Theory, Translations of Mathematical Monographs, Vol. 208, American Mathematical Society.

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

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

Protter, P., 1990, Stochastic Integration and Differential Equations, Springer, New York.

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

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

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

Shmulevich, I., Dougherty, E., 2007, Genomic Signal Processing, Princeton University Press, Princeton, NJ.

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.

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; Casting polymer nets to optimize noisy molecular codes, Proceedings of the National Academy of Sciences, 105:8238-8243.

Tlusty, T., 2010, Personal communication.

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

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

Wallace, R., 2011, On the evolution of homochirality, Comptes Rendus Biologies, 334:263-268.

Wallace, R., 2012a, Metabolic constraints on the evolution of genetic codes: did multiple 'preaerobic' ecosystem transitions entrain richer dialects via serial endosymbiosis?

Transactions on Computational Systems Biology XIV, LNBI 7625:204-232.

Wallace, R., 2012b, Extending Tlusty's rate distortion index theorem method to the glycome: do even 'low level' biochemical phenomena require sophisticated cognitive paradigms?, BioSystems, 107:145-152.

Wallace, R., 2012c, Consciousness, crosstalk, and the mereological fallacy: an evolutionary perspective, Physics of Life Reviews, 9:426-453.

Wallace, R., 2012d, Spontaneous symmetry breaking in a non-rigid molecule approach to intrinsically disordered proteins, Molecular BioSystems, 8:374-377.

Wallace, R., 2014a, A new formal perspective on 'Cambrian explosions', Comptes Rendus Biologies, 337:1-5.

Wallace, R., 2014b, Cognition and biology: perspectives from information theory, Cognitive Processing, 15:1-12.