Version 3

Extending Tlusty's method to the glycome: Tuning the repertoire of glycan determinants

Rodrick Wallace, PhD
Deborah Wallace, PhD
Division of Epidemiology
The New York State Psychiatric Institute [*]

May 1, 2011

Abstract

We apply Tlusty's information-theoretic analysis of the genetic code to the glycome, using a cognitive paradigm in which external information sources constrain and tune the glycan code error network, in the context of available metabolic energy. The resulting dynamic model suggests the possibility of observing spontaneous symmetry breaking of the glycan code as a function of metabolic energy intensity. These effects may be currently present, or embedded in evolutionary trajectory, recording large-scale ecosystem resilience shifts in energy availability such as the aerobic transition.

Key Words: carbohydrate, information theory, rate distortion manifold, spontaneous symmetry breaking.

1 Introduction

Glycomics is the study of the glycans and glycoconjugates – loosely, carbohydrates – produced by a cell or organism under specific conditions (e.g., Hart and Copeland, 2010). The glycome – the general body of such substances – is not well characterized. Mian and Rose (2011), for example, write that

> Superficially, the paucity of information- and coding-theoretic studies of carbohydrates can be explained by glycomics being a less mature field than genomics or proteonics... A deeper explanation is the more complex and dynamic nature of the glycome – the entire complement of carbohydrates... of an organism or a cell... Communication theoretic studies of the glycome and the [glycan code], the complex information conveyed by glycans and glycoconjugates, would increase understanding of the major events in macroevolution and extant molecular biology. For example, the biological communication mediated by glycans underlies diverse molecular, cellular, and tissue functions and plays critical roles in development, health and disease.

Tlusty (2007, 2008) examines the production of amino acids from codons using an information-theoretic formulation based on application of topological methods to a network error analysis, an approach that can be used to illuminate something of the difficulties currently facing glycomics.

Wallace (2010a, b) has, in fact, applied Tlusty's methods to models of protein folding, and we will, in some measure, extend that work to the glycome. The fundamental problem, however, is that both gene coding for amino acids, and many processes of protein folding, can be described as deterministic-but-for-errors. As Anfinsen (1973) has shown, an amino acid 'string' for a structured protein carries within it the information needed for correct folding, at least at near-zero aqueous concentrations of metabolites. The Hecht group (e.g., Hecht et al., 2004; Kim and Hecht, 2006) has shown that the famous protein α-helices and β-sheets are simply 'coded' by strings of alternating hydrophobic and hydrophilic amino acids having the digital signal forms 101100100110... and 101010101... respectively, where 1 indicates polar and 0 non-polar amino acid. The α-helix thus has a 3.6 residue/turn pattern, and the β-sheets alternate. Any polar/non-polar amino acids will suffice, although folding rates will vary greatly, and this permits application of Tlusty's error analysis. Glycans are different, as Hart and Copeland (2010) explain:

> Unlike nucleic acids and proteins, glycan structures are not hard-wired into the genome, depending upon a template for their synthesis. Rather, the glycan structures that end up on a polypeptide or lipid result from the concerted actions of highly specific glycosyltransferases... which are in turn dependent upon [a multiplicity of other processes]... Therefore, the glycoforms of a glycoprotein depend on many factors directly tied to both gene expression and cellular metabolism.

[*] Box 47, 1051 Riverside Dr., New York, NY, 10032 USA. wallace@pi.cpmc.columbia.edu

Elsewhere we have explored a cognitive paradigm for gene expression for signaling by environmental and developmental effects that turn genes on or off (Wallace and Wallace, 2008, 2009, 2010). Here we will extend that work to the glycome, recognizing that a broad class of cognitive processes can be represented in terms of 'dual' information sources, permitting generalization of Tlusty's methods via the tools of network information theory. We begin with a brief recapitulation of Tlusty's work, state the 'central problem' of glycomics from that perspective, and describe a strategy of theoretical attack loosely based on the rate distortion manifold work of Glazebrook and Wallace (2009).

2 Tlusty's stochastic topology

Tlusty (2007) describes the genetic code in terms of an error-network of equivalent and nearly-equivalent codons:

> The maximum [of a particular information-theory-based Morse Function] determines a single contiguous domain where a certain amino acid is encoded... Thus every mode [of the network] 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 points out (2010), is analogous, but not identical, 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):

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

(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.

We note an important fact from Morse Theory (e.g., Matsumoto, 2002):

$$\gamma = 1 - \frac{1}{2}\chi,$$

(2)

where χ is the Euler characteristic of the underlying topological manifold. For a manifold having a Morse function f, χ can be expressed as the alternating sum of the function's Morse numbers: The Morse numbers $\mu_i (i = 0, 1, ..., m)$ of f on the manifold are the number of critical points $(df(x_c) = 0)$ of index i, the number of negative eigenvalues of the matrix $H_{i,j} = \partial f^2 / \partial x_i \partial x_j$. Then $\chi = \sum_{i=0}^{m} (-1)^i \mu_i$.

This holds true for any Morse function on the manifold M.

We reproduce part of Tlusty's Table 1, 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, to which Morse-theoretic 'free energy' functionals, like Tlusty's, 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.

Wallace (2010a, b) has applied Tlusty's approach to the classification of protein symmetries to derive the underlying topology of the 'protein folding code'. Following the seminal work of Levitt and Chothia (1976), there are four major globular protein structures: all-α helices, all-β sheets, α/β, $\alpha + \beta$, with obvious definitions. Chou and Maggioria (1998), using heroic methods on a much larger data set, identify a total of from 7 to 10 such classes, the majority of which seem fairly rare. From the second table, we infer that the normal globular 'protein folding code error network', in Tlusty's sense, is essentially a large connected 'sphere' – giving the four dominant structures – having one minor, and possibly as many as three more 'subminor' attachment handles in the Morse Theory sense (Matsumoto, 2002).

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

3 The glycomic conundrum

Richard Cummings (2009) reviews the major classes of glycan determinants recognized by glycan-binding proteins – the tips of carbohydrates on cellular surfaces that are recognized by other chemical species. These are made up of 2 to 6 linear monnosaccharides together with their potential side chains containing other sugars and modifications like sulfonation, phosphorylation, and acetylation. Glycosaminoglycans comprise repeating disaccharide motifs, where a linear sequence of 5 to 6 monosaccharides may be required for recognition. Cummings estimates that glycoproteins and glycolipids may contain a minimum of 3000 glycan determinants, with an additional minimum of 4000 theoretical polysaccharide sequences in glycosaminoglycans, say a total of 7-10 thousand glycan determinants. Figure 1 applies Heawood's formula to the case of 10000 'amino acids': the underlying 'glycome error code network' must be a grotesquely complicated topological object (GCTO), having between 4 and 8 million holes.

The solution to this extraordinary ambiguity is similar to that of the gene expression problem in general where external signals guide the timing of turning some tens of thousands of of genes on and off during development to produce a vast array of appropriate phenotypes: incoming information limits and channels developmental possibilities (Wallace and Wallace, 2008, 2009, 2010).

4 A cognitive paradigm

To reiterate, the glycoforms of a glycoprotein depend on many factors directly tied to both gene expression and cellular metabolism. This suggests the operation of a cellular-level process of chemical cognition, broadly analogous to the operation of the immune system as described by Atlan and Cohen (1998). That is, incoming information 'farms' glycoforms, in the context of metabolic energy intensity. That is, we take a two-step approach, first examining the effect of incoming signals, and then of the intensity of metabolic energy being made available. The argument becomes, not uncharacteristically, progressively more complicated as constraints increase.

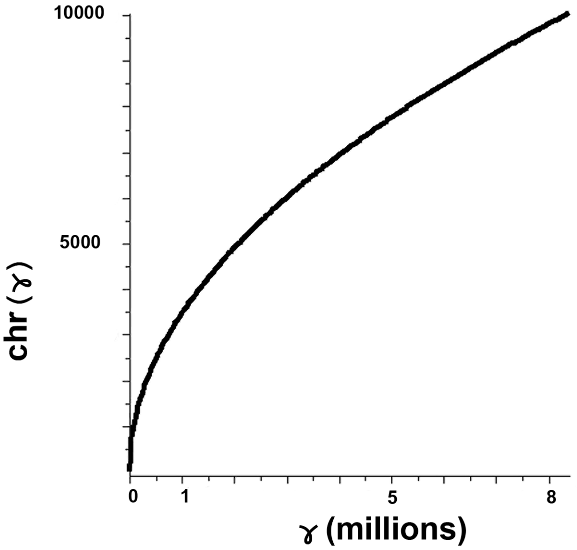

Figure 1: Heawood's topological coloring formula: Number of modes analogous to amino acids vs. number of 'holes' in the underlying error code network. The 10,000 modes representing glycan determinants require an underlying error code network with 8.3 million holes.

4.1 The dual information source

Cognition – here, the selection of a small part of the GCTO for actual implementation – involves choice that limits uncertainty, and thus a broad class of cognitive processes can be represented by 'dual' information sources. The underlying model, described by one observer as 'trivial but not unimportant', follows Atlan and Cohen (1998) and Wallace (2000).

Cognitive pattern recognition-and-selected response, from this perspective, proceeds by convoluting an incoming external 'sensory' signal with an internal 'ongoing activity' – which includes, but is not limited to, some learned or inherited picture of the world – and, at some point, triggering an appropriate action based on a decision that the pattern of sensory activity requires a response. It is not necessary to specify how the pattern recognition system is trained, and hence possible to adopt a 'weak' model, applicable regardless of learning paradigm. Fulfilling Atlan and Cohen's criterion of meaning-from-response, it is possible to define a language's contextual meaning entirely in terms of system output.

The model, a simplification of the standard neural network, is as follows.

A pattern of 'sensory' input – incorporating feedback from the external world – is expressed as an ordered sequence

$y_0, y_1, ...$ This is mixed in a systematic (but unspecified) algorithmic manner with internal 'ongoing' activity, a sequence $w_0, w_1, ...$, to create a path of composite signals $x = a_0, a_1, ..., a_n, ...$, where $a_j = f(y_j, w_j)$ for some function f. This path is then fed into a highly nonlinear, but otherwise similarly unspecified, decision oscillator generating an output $h(x)$ that is an element of one of two (presumably) disjoint sets B_0 and B_1. We take $B_0 \equiv \{b_0, ..., b_k\}, B_1 \equiv \{b_{k+1}, ..., b_m\}$.

Thus the model permits a graded response, supposing that if $h(x) \in B_0$ the pattern is not recognized, and if $h(x) \in B_1$ the pattern is recognized and some action $b_j, k+1 \leq j \leq m$ takes place.

This approach is broadly analogous to, but simpler than, the Hopfield/Hebb stochastic neuron in which series of inputs $y_i^j, i = 1...m$ from m nearby neurons at time j is convoluted with 'weights' $w_i^j, i = 1...m$, using an inner product $a_j = \mathbf{y}^j \cdot \mathbf{w}^j = \sum_{i=1}^m y_i^j w_i^j$ in the context of a 'transfer function' $f(\mathbf{y}^j \cdot \mathbf{w}^j)$ such that the probability of the neuron firing and having a discrete output $z^j = 1$ is $P(z^j = 1) = f(\mathbf{y}^j \cdot \mathbf{w}^j)$. Thus the probability that the neuron does not fire at time j is $1 - f(\mathbf{y}^j \cdot \mathbf{w}^j)$.

The m values y_i^j constitute 'sensory activity' and the m weights w_i^j the 'ongoing activity' at time j, with $a_j = \mathbf{y}^j \cdot \mathbf{w}^j$ and $x = a_0, a_1, ...a_n, ...$ A little more work leads to a fairly standard neural network model in which the network is trained by appropriately varying the $\mathbf{w}$ through least squares or other error minimization feedback.

The principal focus of the simpler model here is the composite paths x that trigger pattern recognition-and-response. That is, given a fixed initial state a_0, such that $h(a_0) \in B_0$, we examine all possible subsequent paths x beginning with a_0 and leading to the event $h(x) \in B_1$. Thus $h(a_0, ..., a_j) \in B_0$ for all $0 \leq j < m$, but $h(a_0, ..., a_m) \in B_1$. Recall that the y_j, the 'sensory' input convoluted with the internal w_j, contains feedback from the external world, i.e., how well h matches intent with need.

For each positive integer n let $N(n)$ be the number of grammatical and syntactic high probability paths of length n which begin with some particular a_0 having $h(a_0) \in B_0$ and lead to the condition $h(x) \in B_1$. Call such paths 'meaningful' and assume $N(n)$ to be considerably less than the number of all possible paths of length n – pattern recognition-and-response is comparatively rare.

The essential assumption is that the longitudinal finite limit

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

(3)

both exists and is independent of the path x. Call such a cognitive process *ergodic.*

Note that disjoint partition of state space may be possible according to sets of states which can be connected by meaningful paths from a particular base point, leading to a natural coset algebra of the system defining a groupoid (e.g., Glazebrook and Wallace, 2009).

It is thus possible to define an ergodic information source $\mathbf{X}$ associated with stochastic variates X_j having joint and conditional probabilities $P(a_0, ..., a_n)$ and $P(a_n | a_0, ..., a_{n-1})$ such that appropriate joint and conditional Shannon uncertainties may be defined which satisfy the standard relations of the Shannon-McMillan Theorem (Cover and Thomas, 2006):

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

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

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

(4)

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

Recall that the Shannon-McMillan Theorem and its variants provide 'laws of large numbers' that permit definition of the Shannon uncertainties in terms of cross-sectional sums of the form $H = -\sum P_k \log[P_k]$, where the P_k constitute a probability distribution (Ash, 1990; Cover and Thomas, 2006).

Different quasi-languages will be defined by different divisions of the total universe of possible responses into various pairs of sets B_0 and B_1. Like the use of different distortion measures in the Rate Distortion Theorem, however, it seems obvious that the underlying dynamics will all be qualitatively similar.

Nonetheless, dividing the full set of possible responses into the sets B_0 and B_1 may itself require higher order cognitive decisions by another module or modules, suggesting the necessity of choice within a more or less broad set of possible quasi-languages. This would directly reflect the need to shift gears according to the different challenges faced by the organism or organic subsystem. A critical problem then becomes the choice of a normal zero-mode language among a very large set of possible languages representing accessible excited states. This is a fundamental matter that mirrors, for isolated cognitive systems, the resilience arguments applicable to more conventional ecosystems, that is, the possibility of more than one zero state to a cognitive system. Identification of an excited state as the zero mode becomes, then, a kind of generalized autoimmune disorder that can be triggered by linkage with external ecological information sources representing various kinds of structured stress.

In sum, meaningful paths – creating an inherent grammar and syntax of cognitive process – have been defined entirely in terms of system response, as Atlan and Cohen propose.

4.2 Network information theory

We reiterate that the central function of the cognitive chemical selection process defined in the section above is to restrict the play of the information-theoretic Morse Function, in Tlusty's sense, whose modes define the amino acid analogs of the 10000 glycan determinants.

The essential point is that the dual information source representing the cognitive selection of some small zone of the GCTO is itself acted on by gene expression and metabolic contexts. Cognitive gene expression, in the sense of Wallace and Wallace (2008, 2009), can be represented by impinging dual information sources.

The tool for the action of such external context on the GCTO selector is network information theory (e.g., Cover and Thomas, 2006; El Gamal and Kim, 2010).

Given three interacting information sources, Y_1, Y_2, Z, the splitting criterion for tripartite jointly typical sequences, taking Z as an external context, is (Cover and Thomas, 2006, p. 524)

$$I(Y_1; Y_2|Z) = H(Z) + H(Y_1|Z) + H(Y_2|Z) - H(Y_1, Y_2, Z), \tag{5}$$

where $H(...|...)$ and $H(..., ..., ...)$ represent conditional and joint uncertainties (Ash, 1990; Khinchin, 1957; Cover and Thomas, 2006).

More complicated multivariate typical sequences receive much the same treatment (El Gamel and Kim, 2010, p.2-26). Given a basic set of information sources $(X_1, ..., X_k)$ that one partitions into two ordered sets $X(J)$ and $X(J')$, then the splitting criterion becomes $H(X(J)|X(J'))$. Generalization to a greater number of ordered sets is straightforward.

Then the joint splitting criterion – I, H above – however expressed as a composite of the underlying information sources and their interactions, satisfies a relation analogous to equation (3), where $N(n)$ is the number of high probability jointly typical paths of length n. The joint splitting criterion is given as a functional composition of the underlying information sources and their interactions as affected by the embedding contextual cognitive information sources indexed by J'.

4.3 Metabolic effects

The final iteration of the argument involves incorporating the effects of metabolic energy on the cognitive GCTO Tlusty Morse Function selector, now convoluted through network information theory with cognitive gene expression dual information sources, in the sense of Wallace and Wallace (2008, 2009). We do this in a painfully standard manner. Assume there is a tunable metabolic energy bottleneck that delivers a limited intensity of metabolic energy per unit reactor, an intensive measure M. The usual assumption, following a simple Gibbs model, is that the probability of $H(X(J)|X(J'))$ is given by an expression

$$Pr[H] \propto \frac{\exp[-H(J|J')/M]}{\sum_{J'} \exp[-H(J|J')/M]}. \tag{6}$$

The $X(J')$ represent the effects of embedding cognitive gene expression mechanisms, via their dual information sources, and M that of available cell energy metabolism.

The denominator of this expression is much like the partition function in statistical physics.

Letting

$$\exp[-F/M] = \sum_{J'} \exp[-H(J|J')/M] \equiv Z \tag{7}$$

we can define a new Morse Function that is analogous to the free energy of a simple physical system as

$$F = -M \log[Z]. \tag{8}$$

Here cognitive gene expression mechanisms and cell energy metabolism have been combined to dynamically tune the cognitive process that chooses the venue of the Tlusty Morse Function defining the repertoire of glycan determinants. This chain of processes is needed to collapse the GCTO onto something both physiologically manageable and appropriate, another retina-like rate distortion manifold in the sense of Glazebrook and Wallace (2009). In effect, the selector has to both choose and power the subrepertoire, and these are significant restrictions.

5 Discussion and conclusions

We have, after a long chain of argument, invoked the cognitive rate distortion manifold of Glazebrook and Wallace (2009) to dynamically tune and power a glycan code subset according to the demands of gene expression and metabolic energy, a glycan fovea, as it were.

The introduction of a metabolically-driven free energy in equation (8) carries certain standard implications regarding punctuated phase transitions.

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

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

(9)

with $A_2 \approx \alpha(T - T_c)$ sufficiently close to the critical temperature T_c. This mean field approach can be used to describe a great variety of effects.

Minimization of F_0 with respect to the order parameter yields the average value of ϕ, $< \phi >$, which is zero above the critical temperature and non-zero below it. In the absence of external fields, the second-order transition occurs at $T = T_c$.

The Landau formalism quickly enters deep topological waters (e.g., Pettini, 2007, pp. 42-43; Landau and Lifshitz, 2007, pp. 459-466). The essence of Landau's insight was that phase transitions without latent heat – second order transitions – were usually in the context of a significant symmetry change in the physical states of a system, with one phase, at higher temperature, being far more symmetric than the other. A symmetry is lost in the transition, a phenomenon called spontaneous symmetry breaking. The greatest possible set of symmetries in a physical system is that of the Hamiltonian describing its energy states. Usually states accessible at lower temperatures will lack symmetries available at higher temperatures, so that the lower temperature phase is the less symmetric: The randomization of higher temperatures ensures that higher symmetry/energy states will then be accessible to the system.

Equation (8) suggests that increasing the intensity of available metabolic energy is a necessary condition for using larger sections of the GCTO defining the repertoire of glycan determinants, and that the accession to higher 'symmetries' within that repertoire is almost assuredly highly punctuated with increasing energy in a series of discernible phase transitions. An alternative, less explicit, formulation would simply focus on the delivery of metabolic energy as determined by an external information source, restricting the Morse Function to something like $H(J|J')$ above, for which phase transition behavior would be more difficult to characterize.

The punctuated glycan fovea we postulate should be observable, either in present systems, or frozen within evolutionary trajectory, as ecosystem shifts like the aerobic transition made available greater intensity of metabolic energy. At the very least, a search for it provides a principled approach to functional glycomics likely to yield fundamental advances, if only disproof of our simple model.

6 References

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

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

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

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

Cover, T., H. Thomas, 2006, Elements of Information Theory, Second Edition, Wiley, New York.

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

El Gamal, A., Y. Kim, 2010, Lecture Notes on Network Information Theory, ArXiv:1001.3404v4.

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

Hart, G., R. Copeland, 2010, Glycomics hits the big time, Cell, 143:672-676.

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

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

Kim, W., M. Hecht, 2006, Generic hydrophobic residues are sufficient to promote aggregation of the Alzheimer's Aβ42 peptide, Proceedings of the National Academy of Sciences, 103:15824-15829.

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

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

Matsumoto, Y., 2002, An Introduction to Morse Theory, Transactions of the American Mathematical Society 208, Providence, RI.

Mian, I., C. Rose, 2011, Communication theory and multicellular biology, Integrative Biology, 3:350-367.

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

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

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.

Tlusty, T., 2010, Personal communication.

Wallace, R., 2000, Language and coherent neural amplification in hierarchical systems: renormalization and the dual information source of a generalized stochastic resonance, International Journal of Bifurcation and Chaos, 10:493-502.

Wallace, R., 2010a, Structure and dynamics of the 'protein folding code' inferred using Tlusty's topological rate distortion approach, BioSystems, 103:18-26.

Wallace, R., 2010b, Protein folding disorders: Toward a basic biological paradigm, Journal of Theoretical Biology, 267:582-594.

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

Wallace, R., D. Wallace, 2010, Gene Expression and Its Discontents, Springer, New York.