THEORETICAL MODELING

On the Spectrum of Prebiotic Chemical Systems

An Information Theory Treatment of Eigen's Paradox

Rodrick Wallace · Robert G. Wallace

Abstract We reexamine Eigen's paradox using the asymptotic limit theorems of information theory. Applying the homology between information source uncertainty and free energy density, under rate distortion constraints, the error catastrophe emerges as the lowest energy state for simple prebiotic systems without error correction. Invoking the usual compartmentalization – i.e., 'vesicles' – and using a Red Queen argument, suggests that information crosstalk between two or more properly interacting structures can initiate a coevolutionary dynamic having at least two quasi-stable states. The first is a low energy realm near the error threshold, and, depending on available energy, the second can approach zero error as a limit. A large deviations argument produces jet-like global transitions which, over sufficient time, may enable shifts between the many quasi-stable modes available to more complicated structures, 'locking in' to some subset of the various possible low error rate chemical systems, which become subject to development by selection and chance extinction. Energy availability, according to the model, is thus a powerful necessary condition for low error rate replication, suggesting that some fundamental prebiotic ecosystem transformation entrained reproductive fidelity. This work, then, supports speculation that our RNA/DNA world may indeed be only the chance result of a very broad prebiotic evolutionary phenomenon. Processes *in vitro*, or *ex planeta*, might have other outcomes.

Keywords Biochemical evolution · Eigen's paradox · Error catastrophe · Information theory · Rate distortion

R. Wallace (✉)
The New York State Psychiatric Institute, New York, NY, USA
e-mail: wallace@pi.cpmc.columbia.edu

R. G. Wallace
Department of Geography, University of Minnesota, Minneapolis, MN, USA

Introduction

Manfred Eigen's evolution model (Eigen 1971, 1996) stands as a landmark attempt at coherently relating evolution, molecular biology, and information theory. That work views selection as condensation in information space, and evolution as a succession of phase transformations. Indeed, Eigen's quasispecies model, with its error catastrophe, corresponds exactly to a phase transition in a two dimensional Ising system (Leuthausser 1986). The essence of Eigen's paradox is that the error catastrophe limits genome length in RNA precursor organisms to much less than observed in DNA organisms having error correcting enzymes, which, themselves, cannot be created in the absence of just such a long genome.

As Holmes (2005) put it,

> To create more genetic complexity, it is therefore necessary to encode more information in longer genes by using a replication system with greater fidelity. But there's the catch: to replicate with greater fidelity requires a more accurate and hence complex replication enzyme, but such an enzyme cannot be created because this will itself require a longer gene, and longer genes will breach the error threshold.

Here we will reconsider Eigen's paradox from a highly formal perspective which hews quite closely to the fundamental asymptotic limit theorems of information theory, in particular the Rate Distortion Theorem, and its zero-error limit, the Shannon–McMillan, or Source Coding Theorem. These, like the Central Limit Theorem in parametric statistics, permit derivation of 'regression-like' models which can be applied to real data. We use these theorems in a principled manner to derive the high rate of mutation inherent to RNA virus replication and suggest a plausible 'Red Queen' coevolutionary ratchet (Van Valen 1973) leading toward an evolutionary condensation resulting in effective error-correction mechanisms.

Our program has these components:

[1] An increasingly complicated network of simple interacting 'RNA-like' organisms creates a collective biochemical system – a 'vesicle' – which, as a parallel communication channel, can have a much higher channel capacity for low-error replication than do the individual components.

[2] Several such distinct, properly interacting, collectives – compartments in the sense of Eigen and Szathmary – become each others' most intimate environments, generating a coevolutionary ratchet resulting in a Red Queen structure which, given sufficient energy, can support quasi-stable states with very low reproductive error rates.

[3] High error rate, but low energy, systems-of-vesicles can become subject to systematic 'large deviations' excursions which, over sufficient time, can lead to the establishment of a distribution of low error rate, but higher energy, chemical systems: even prebiological quasi-organisms can, apparently, build pyramids, as it were. Thus many different chemistry-of-life solutions seem possible, and these, of course, would be subject to evolutionary processes of selection and chance extinction.

[4] The latter steps depend critically on the availability of adequate energy sources, which, we hold, will be largely driven by changes in the protoecosystem, i.e., ecosystem resilience domain shifts, in the sense of Holling (1973)

(Gunderson 2000; Wallace and Wallace 2008). One imagines availability of a new metabolic cycle, onset of predation, and/or crude photosynthesis as possible examples.

This is not, except perhaps for [4], a particularly new perspective, being essentially similar to the 'bags of genes' (Holmes 2005) model of Szathmary and colleagues (Szathmary and Demeter 1987; Szathmary 1989, 2006; Fontanari et al. 2006) and, of course, to Eigen's hypercycle compartment model (Eigen 1996), but with hypercycles generalized to broad coevolutionary interactions. Our innovations lie in a highly formal use of the asymptotic limit theorems of information theory and in the systematic extension of phase transition techniques from statistical physics to information theory via 'biological' renormalizations.

Although the basic line of reasoning is fairly straightforward, a number of mathematical tasks need to be confronted. The first is to reconfigure the error catastrophe of Eigen's model in terms of average distortion. That done, the famous homology between information source uncertainty and free energy density can be extended to the rate distortion function, and, since the rate distortion limit is always a non-increasing convex function of the distortion, the basic model emerges as a kind of energy minimization near the error catastrophe. Pettini's (2007) topological hypothesis regarding the relation between topological shifts in structure and phase transition can than be applied, using characteristic 'biological renormalizations' to specify different forms of phase transition. A theory of 'all possible' Eigen models is direct. For analogs to the Gaussian channel, zero distortion – no error at all – is unattainable, requiring infinite energy. Such systems would, then, always be subject to some mutational variation.

A standard coevolutionary argument applied to two or more vesicles interacting through a mutual information crosstalk produces the essential results, crudely similar to the set of stable strategies of elaborate stochastic coevolutionary game theory models (e.g., Champagnat et al. 2006). Again, the simplest possible system has two quasi-equilibrium points, one near the error limit, which is the low energy solution, and the other near zero error, the high energy solution.

The many different possible chemical strategies in this broad spread of possible solutions would themselves become higher order Darwinian individuals subject to the vagaries of evolutionary process in the context of large deviations, producing, on our planet, the familiar RNA/DNA chemistry basis of life. Other possibilities, however, seem particularly interesting, and may perhaps be observed *in vitro*, or *ex planeta*.

Reconsidering the Eigen Model

Following Campos and Fontanari (1999), the Eigen quasispecies model can be characterized in its binary version as follows:

A molecule is represented by a string of L digits, $(s_1, s_2, ..., s_L)$ with the variates s_α allowed to take only two different values, each of which represents a different type of monomer used to build the molecule. The concentrations x_i of molecules of type $i = 1, 2, ..., 2^L$ evolve in time according to the equations

$$dx_i/dt = \sum_j W_{i,j} x_j - \Phi(t) x_i \tag{1}$$

where $\Phi(t)$ is a dilution flux that keeps the total concentration constant, determined by the condition that $\sum_i dx_i/dt = 0$. Taking $\sum_i x_i = 1$ gives

$$\Phi = \sum_{i,j} W_{i,j} x_j. \tag{2}$$

The elements of $W_{i,j}$, the replication matrix, depend on the replication rate or fitness A_i of the strings of type i, as well as on the Hamming distance $d(i, j)$ between strings i and j. They are

$$W_{i,i} = A_i q^v \tag{3}$$

and

$$W_{i,j} = A_j q^{L-d(i,j)}[1-q]^{d(i,j)}, i \neq j. \tag{4}$$

Here $0 \leq q \leq 1$ is the single-digit replication accuracy, assumed the same for all digits. Again, Leuthausser (1986) shows this model corresponds exactly to that of a two dimensional Ising system.

The famous figure on p.83 of Eigen (1996), taken from Swetina and Schuster (1982), shows a numerical realization of the model for $L = 50$. The fittest sequence has Hamming distortion $d = 0$, and is characterized as the 'Master' sequence. As the error rate $(1 - q)$ increases from zero, the proportion of the population corresponding to the master sequence, x_0, declines, while those corresponding to other distortion values rise accordingly. The essential point, however, is that, for all values below the critical error threshold, the average across the quasispecies, all the x_d, what Eigen calls the 'consensus sequence', remains precisely the master sequence itself, even while the proportion constituting the master sequence itself falls. After the error threshold, however, that consensus average is lost, and the distributions become strictly random: the information of the master sequence has been dissipated.

Our interest is in reconsidering this effect from the perspective of the average distortion. As the error threshold is approached from below, the proportion corresponding the master sequence, having $d = 0$, declines, while the proportion of the population having $d > 0$ increases. Thus the average distortion, $D = \sum_j d_j x_j$, such that $\sum_i x_i = 1$, itself increases monotonically as the error rate approaches the error threshold, until a critical value of the average distortion D is reached.

The Rate Distortion Theorem, described in more detail below, states that there is a minimum necessary rate of transmission of information, $R(D)$, through an information channel for any given average distortion D. That is, rates of transmission – channel capacities – greater than $R(D)$ are guaranteed to have average distortion less than D for any chosen distortion measure (Cover and Thomas 1991; Dembo and Zeitouni 1998).

We have translated the Eigen model from a focus on a single error threshold into one involving a critical value of average distortion in the transmission of information, and this will prove to be important for our subsequent modeling exercise.

We will, like others, attempt to solve Eigen's paradox by invoking a population of related mutants as constituting a parallel transmission channel having a sufficiently high capacity to ensure collective reproductive fidelity in spite of individual molecular reproductive errors: The 'consensus average' becomes the essential reproductive message, and different such populations are separated into compartments or vesicles

in the sense of Szathmary and Demeter (1987). These interact to become each other's principal environments, engaging in a crosstalk that enables the coevolutionary ratchet. Many versions of this mechanism may have developed, becoming subject to evolutionary selection and chance extinction. Certainly both Eigen's compartment hypercycle Eigen (1996) and Szathmary's stochastic corrector (Szathmary and Demeter 1987) solve the essential problem. Likely, so too may a plethora of different structures, or, as Szathmary (1989) comments, 'while it is true that hypercycles need compartments, do compartments need hypercycles?'. Here we will attempt a kind of general compartment model, using a broad information-theoretic brush.

Genetic Inheritance as an Information Source

Eigen's and Szathmary's perspective is not the only information theoretic analysis of reproduction, although it certainly takes things to their earliest history. Adami et al. (2000) make a case for reinterpreting the transmission of genetic heritage in terms of a formal information process. They assert that genomic complexity can be identified with the amount of information a sequence stores about its environment: genetic complexity can be defined in a consistent information-theoretic manner. In their view, information cannot exist in a vacuum and must have an instantiation. For biological systems the instantiation of information is DNA. To some extent it is the blueprint of an organism and thus information about its own structure. More specifically, it is a blueprint of how to build an organism that can best survive in its native environment, and pass on that information to its progeny. They assert that an organism's DNA thus is not only a 'book' about the organism, but also a book about the environment it lives in, including the species it coevolves with. They identify the complexity of genomes by the amount of information they encode about the world in which they have evolved.

Ofria et al. (2003) continue in the same direction and argue that genomic complexity can be defined rigorously within standard information theory as the information the genome of an organism contains about its environment. From the point of view of information theory, it is convenient to view evolution on the molecular level as a collection of information transmission channels, subject to a number of constraints. In these channels, they state, the organism's genomes code for the information (a message) to be transmitted from progenitor to offspring, and are subject to noise due to an imperfect replication process. Information theory is concerned with analyzing the properties of such channels, how much information can be transmitted and how the rate of perfect information transmission of such a channel can be maximized.

Adami and Cerf (2000) argue, using simple models of genetic structure, that the information content, or complexity, of a genomic string by itself (without referring to an environment) is a meaningless concept and a change in environment (catastrophic or otherwise) generally leads to a pathological reduction in complexity.

The transmission of genetic information is thus a contextual matter which involves operation of an information source which, according to this development, must interact with embedding (ecosystem) structures.

It is much in this spirit that we now address the Eigen paradox, beginning with a brief review of some facts and theorems from information theory.

Information Sources and Their Dynamics

Channel Capacity

Messages from a source, seen as symbols x_j from some alphabet, each having probabilities P_j associated with a random variable X, are 'encoded' into the language of a 'transmission channel', a random variable Y with symbols y_k, having probabilities P_k, possibly with error. Someone receiving the symbol y_k then retranslates it (without error) into some x_k, which may or may not be the same as the x_j that was sent.

More formally, the message sent along the channel is characterized by a random variable X having the distribution

$$P(X = x_j) = P_j, \, j = 1, ..., M.$$

The channel through which the message is sent is characterized by a second random variable Y having the distribution

$$P(Y = y_k) = P_k, k = 1, ..., L.$$

Let the joint probability distribution of X and Y be defined as

$$P(X = x_j, Y = y_k) = P(x_j, y_k) = P_{j,k}$$

and the conditional probability of Y given X as

$$P(Y = y_k \mid X = x_j) = P(y_k \mid x_j).$$

Then the Shannon uncertainty of X and Y independently and the joint uncertainty of X and Y together are defined respectively as

$$H(X) = -\sum_{j=1}^{M} P_j \log(P_j)$$

$$H(Y) = -\sum_{k=1}^{L} P_k \log(P_k)$$

$$H(X, Y) = -\sum_{j=1}^{M} \sum_{k=1}^{L} P_{j,k} \log(P_{j,k}). \tag{5}$$

The *conditional uncertainty* of Y given X is defined as

$$H(Y \mid X) = -\sum_{j=1}^{M} \sum_{k=1}^{L} P_{j,k} \log[P(y_k \mid x_j)] \tag{6}$$

For any two stochastic variates X and Y, $H(Y) \geq H(Y \mid X)$, as knowledge of X generally gives some knowledge of Y. Equality occurs only in the case of stochastic independence.

Since $P(x_j, y_k) = P(x_j) P(y_k \mid x_j)$, we have

$$H(X \mid Y) = H(X, Y) - H(Y)$$

The information transmitted by translating the variable X into the channel transmission variable Y – possibly with error – and then retranslating without error the transmitted Y back into X is defined as

$$I(X \mid Y) \equiv H(X) - H(X \mid Y) = H(X) + H(Y) - H(X, Y). \tag{7}$$

See, for example, Ash (1990), Khinchin (1957) or Cover and Thomas (1991) for a modern treatment. The original development can be found in Shannon and Weaver (1949) and Shannon (1957). The essential point is that if there is no uncertainty in X given the channel Y, then there is no loss of information through transmission.

In general this will not be true, and herein lies the essence of the theory.

Given a fixed vocabulary for the transmitted variable X, and a fixed vocabulary and probability distribution for the channel Y, we may vary the probability distribution of X in such a way as to maximize the information sent. The capacity of the channel is defined as

$$C \equiv \max_{P(X)} I(X \mid Y), \tag{8}$$

subject to the subsidiary condition that $\sum P(X) = 1$.

The central mechanism of the Shannon Coding Theorem which enables the sending of a message with arbitrarily small error along the channel Y at any rate $R < C$ is to encode it in longer and longer 'typical' sequences of the variable X; that is, those sequences whose distribution of symbols approximates the probability distribution $P(X)$ above which maximizes C.

If $S(n)$ is the number of such 'typical' sequences of length n, then

$$\log[S(n)] \approx nH(X)$$

where $H(X)$ is the uncertainty of the stochastic variable defined above. Some consideration shows that $S(n)$ is much less than the total number of possible messages of length n. Thus, as $n \rightarrow \infty$, only a vanishingly small fraction of all possible messages is meaningful in this sense. This observation, after some considerable development, is what allows the Coding Theorem to work so well. In sum, the prescription is to encode messages in typical sequences, which are sent at very nearly the capacity of the channel. As the encoded messages become longer and longer, their maximum possible rate of transmission without error approaches channel capacity as a limit. Again, Ash (1990), Khinchin (1957) and Cover and Thomas (1991) provide details.

A trivial but important generalization of the simple channel without memory is the Parallel Channel. Consider a set of K discrete memoryless channels having capacities $C_1, ..., C_K$. Assume these are connected in parallel in the sense that each unit of time an arbitrary symbol is transmitted and received over each channel. The input $\mathbf{x} = (x_1, ...x_K)$ to the parallel channel is a K-vector whose components are inputs to the individual channels, and the output $\mathbf{y} = (y_1, ..., y_K)$ is a K-vector whose components are the individual channel outputs. The capacity of the parallel channel is

$$C_{Total} = \sum_{i=1}^{K} C_i. \tag{9}$$

Thus the consensus sequence in a compartment or vesicle, which is clearly a parallel channel, can be transmitted at a much higher overall rate than is possible using individual replicators. This is no small matter.

The Shannon–McMillan Theorem

Not all statements – sequences of the random variable X – are equivalent. According to the structure of the underlying language of which the message is a particular expression, some messages are more 'meaningful' than others, that is, in accord with

the grammar and syntax of the language. The other principal result from information theory, the Shannon–McMillan or Asymptotic Equipartition Theorem, describes how messages themselves are to be classified.

Suppose a long sequence of symbols is chosen, using the output of the random variable X above, so that an output sequence of length n, with the form

$$x_n = (\alpha_0, \alpha_1, ..., \alpha_{n-1})$$

has joint and conditional probabilities

$$P(X_0 = \alpha_0, X_1 = \alpha_1, ..., X_{n-1} = \alpha_{n-1})$$
$$P(X_n = \alpha_n \mid X_0 = \alpha_0, ..., X_{n-1} = \alpha_{n-1}). \tag{10}$$

Using these probabilities we may calculate the conditional uncertainty

$$H(X_n \mid X_0, X_1, ..., X_{n-1}).$$

The uncertainty of the *information source*, $H[\mathbf{X}]$, is defined as

$$H[\mathbf{X}] \equiv \lim_{n \to \infty} H(X_n \mid X_0, X_1, ..., X_{n-1}). \tag{11}$$

In general

$$H(X_n \mid X_0, X_1, ..., X_{n-1}) \leq H(X_n).$$

Only if the random variables X_j are all stochastically independent does equality hold. If there is a maximum n such that, for all $m > 0$

$$H(X_{n+m} \mid X_0, ..., X_{n+m-1}) = H(X_n \mid X_0, ..., X_{n-1}),$$

then the source is said to be of *order* n. It is easy to show that

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

In general the outputs of the X_j, $j = 0, 1, ..., n$ are *dependent*. That is, the output of the communication process at step n depends on previous steps. Such serial correlation, in fact, is the very structure which enables most of what follows.

Here, however, the processes are all assumed stationary in time, that is, the serial correlations do not change in time, and the system is *stationary*.

A very broad class of such self-correlated, memoryless, information sources, the so-called *ergodic* sources for which the long-run relative frequency of a sequence converges stochastically to the probability assigned to it, have a particularly interesting property:

It is possible, in the limit of large n, to divide all sequences of outputs of an ergodic information source into two distinct sets, S_1 and S_2, having, respectively, very high and very low probabilities of occurrence, with the source uncertainty providing the splitting criterion. In particular the Shannon–McMillan Theorem states that, for a (long) sequence having n (serially correlated) elements, the number of 'meaningful' sequences, $N(n)$ – those belonging to set S_1 – will satisfy the relation

$$\log[N(n)]/n \approx H[\mathbf{X}]. \tag{12}$$

More formally,

$$\lim_{n \to \infty} \log[N(n)]/n = H[\mathbf{X}]$$
$$= \lim_{n \to \infty} H(X_n \mid X_0, ..., X_{n-1})$$
$$= \lim_{n \to \infty} H(X_0, ..., X_n)/(n+1). \tag{13}$$

The Shannon Coding theorem, by means of an analogous splitting argument, shows that for any rate $R < C$, where C is the channel capacity, a message may be sent without error, using the probability distribution for X which maximizes $I(X \mid Y)$ as the coding scheme. Using the internal structures of the information source permits *limiting attention only to meaningful sequences of symbols*. This restriction can greatly raise the maximum possible rate at which information can be transmitted with arbitrarily small error: if there are M possible symbols and the uncertainty of the source is $H[\mathbf{X}]$, then the effective capacity of the channel C, using this 'source coding,' becomes (Ash 1990)

$$C_E = C \log(M)/H[\mathbf{X}]. \tag{14}$$

As $H[\mathbf{X}] \leq \log(M)$, with equality only for stochastically independent, uniformly distributed random variables,

$$C_E \geq C. \tag{15}$$

Note that, for a given channel capacity, the condition

$$H[\mathbf{X}] \leq C \tag{16}$$

always holds.

Source uncertainty has a very important heuristic interpretation. As Ash (1990) puts it,

...[W]e may regard a portion of text in a particular language as being produced by an information source. The probabilities $P[X_n = \alpha_n \mid X_0 = \alpha_0, ..., X_{n-1} = \alpha_{n-1})$ 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.

The Rate Distortion Theorem

The Shannon–McMillan Theorem is the zero error limit of the Rate Distortion Theorem (Dembo and Zeitouni 1998; Cover and Thomas 1991), which can be expressed in terms of a splitting criterion that identifies high probability pairs of sequences. We follow closely the treatment of Cover and Thomas (1991).

The origin of the problem is the question of representing one information source by a simpler one in such a way that the least information is lost. For example we

might have a continuous variate between 0 and 100, and wish to represent it in terms of a small set of integers in a way that minimizes the inevitable distortion that process creates. Typically, for example, an analog audio signal will be replaced by a 'digital' one. The problem is to do this in a way which least distorts the *reconstructed* audio waveform.

Suppose the original stationary, ergodic information source Y with output from a particular alphabet generates sequences of the form

$$y^n = y_1, ..., y_n.$$

These are 'digitized,' in some sense, producing a chain of 'digitized values'

$$b^n = b_1, ..., b_n,$$

where the b-alphabet is much more restricted than the y-alphabet.

b^n is, in turn, *deterministically retranslated* into a reproduction of the original signal y^n. That is, each b^m is mapped on to a unique n-length y-sequence in the alphabet of the information source Y:

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

Note, however, that many y^n sequences may be mapped onto the *same* retranslation sequence $\hat{y}^n$, so that information will, in general, be lost.

The central problem is to minimize that loss.

The retranslation process defines a new stationary, ergodic information source, $\hat{Y}$.

The next step is to define a *distortion measure*, $d(y, \hat{y})$, which compares the original to the retranslated path. For example the *Hamming distortion* is

$$d(y, \hat{y}) = 1, \, y \neq \hat{y}$$
$$d(y, \hat{y}) = 0, \, y = \hat{y}. \tag{17}$$

For continuous variates the *Squared error distortion* is

$$d(y, \hat{y}) = (y - \hat{y})^2. \tag{18}$$

The Rate Distortion Theorem applies to a broad class of possible distortion measures.

The distortion between paths y^n and $\hat{y}^n$ is defined as

$$d(y^n, \hat{y}^n) = (1/n) \sum_{j=1}^{n} d(y_j, \hat{y}_j). \tag{19}$$

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

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

The *average distortion* is defined as

$$D \equiv \sum_{y^n} p(y^n) d(y^n, \hat{y}^n). \tag{20}$$

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

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

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

In general, this will not be true.

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

$$R(D) = \min_{p(y, \hat{y}); \sum_{(y, \hat{y})} p(y) p(y|\hat{y}) d(y, \hat{y}) \leq D} I(Y, \hat{Y}). \tag{21}$$

The minimization is over all conditional distributions $p(y \mid \hat{y})$ for which the joint distribution $p(y, \hat{y}) = p(y) p(y \mid \hat{y})$ 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 (effectively, the channel capacity) so that the average distortion does not exceed the distortion D. Cover and Thomas (1991) or Dembo and Zeitouni (1998) provide details.

Pairs of sequences $(y^n, \hat{y}^n)$ can be defined as *distortion typical*; that is, for a given average distortion D, defined in terms of a particular measure, pairs of sequences can be divided into two sets, a high probability one containing a relatively small number of (matched) pairs with $d(y^n, \hat{y}^n) \leq D$, and a low probability one containing most pairs. As $n \to \infty$, the smaller set approaches unit probability, and, for those pairs,

$$p(y^n) \geq p(\hat{y}^n \mid y^n) \exp[-nI(Y, \hat{Y})]. \tag{22}$$

Thus, roughly speaking, $I(Y, \hat{Y})$ embodies the splitting criterion between high and low probability pairs of paths.

For the theory of interacting information sources, then, $I(Y, \hat{Y})$ can play the role of H in the dynamic treatment that follows.

The rate distortion function of Eq. 21 can actually be calculated in many cases by using a Lagrange multiplier method – see Section 13.7 of Cover and Thomas (1991). For a simple Gaussian channel having noise with zero mean and variance σ^2 then

$$R(D) = 1/2 \log[\sigma^2/D], 0 \leq D \leq \sigma^2$$
$$R(D) = 0, D > \sigma^2. \tag{23}$$

For this particular channel, zero distortion, no mutations at all, requires an infinite channel capacity, which, we will show, requires infinite energy.

A second important observation is that *any* rate distortion function $R(D)$, following the arguments of Cover and Thomas (1991, Lemma 13.4.1) is necessarily a *decreasing convex function* of D, that is, a reverse-J-shaped curve. This requirement, like the singularity of Gaussian-like channels at zero distortion, has profound consequences for replication dynamics.

Information System Phase Transitions

As many have noted, the first part of Eq. 13,

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

is homologous to the thermodynamic limit in the definition of the free energy density of a physical system. This has the form

$$F(K) = \lim_{V \to \infty} \log[Z(K)]/V, \tag{24}$$

where F is the free energy density, K the inverse temperature, V the system volume, and $Z(K)$ is the partition function defined by the system Hamiltonian. Any good statistical mechanics text will provide details (e.g., Landau and Lifshitz 2007).

The Appendix shows how this homology permits a highly natural generalization of standard renormalization methods from statistical mechanics to information theory, producing phase transitions and analogs to evolutionary punctuation in systems characterized by piecewise, adiabatically stationary, ergodic information sources. Thus biological phase changes appear to be ubiquitous in natural systems.

The approach uses a mean field approximation in which average strength (or probability) of nondisjunctive linkages between interacting nodes (vesicles) – crosstalk – serves as a kind of inverse temperature parameter. Phase transitions can then be described using various 'biological' renormalization strategies, in which universality class tuning becomes a principal second order mechanism. These are continuous phase transitions in that there is no latent heat required, as in boiling water until it turns to steam, although a treatment of ecosystem resilience in dynamic manifolds (e.g., Wallace and Wallace 2008) can perhaps be viewed as an analog. Under a resilience domain shift the underlying system topology changes, either within or between dynamic manifolds, and this can be viewed as similar to the fundamental geometric structural change water undergoes when it transforms from crystal to liquid, or from liquid to gas. Indeed, recent work (Franzosi and Pettini 2004; Pettini 2007) uses Morse theory from differential topology to identify necessary conditions for topological shifts on dynamic manifolds of systems undergoing first and second order phase transition. The converse, sufficiency, is an open question. See the Appendix for a brief introduction to Morse Theory.

The important point is that the homology between Eqs. 12 and 24 ensures that some form of emergent behavior, akin to a physical phase transition, is inevitable for a broad class of systems which transmit information.

Phenomenological Onsager Theory

The homology between the information source uncertainty and the free energy density of a physical system arises, in part, from the formal similarity between their definitions in the asymptotic limit. Information source uncertainty can be defined as in Eq. 13. This is, as noted, quite analogous to the free energy density of a physical system, Eq. 24.

Feynman (1996) provides a series of physical examples, based on Bennett's work, where this homology is, in fact, an identity, albeit for very simple systems. Bennett argues, in terms of idealized irreducibly elementary computing machines, that the

information contained in a message can be viewed as the work saved by not needing to recompute what has been transmitted.

Feynman explores in some detail Bennett's ideal microscopic machine designed to extract useful work from a transmitted message. The essential argument is that computing, in any form, takes work. Thus the more complicated a process, measured by its information source uncertainty, the greater its energy consumption, and our ability to provide energy is limited. In a psychological context, the striking phenomenon of inattentional blindness, Wallace (2007) argues, emerges as a thermodynamic limit on processing capacity in a focused cognitive system, that is, one which has been strongly configured about a particular task. Other biological generalizations seem obvious.

Understanding the time dynamics of information systems away from the kind of phase transition critical points described above and in the Appendix requires a phenomenology similar to the Onsager relations of nonequilibrium thermodynamics. This will lead to a more general theory involving large-scale topological changes in the sense of Morse theory.

If the source uncertainty associated with a biological process is parametized by some vector of quantities $\mathbf{K} \equiv (K_1, ..., K_m)$, then, in analogy with nonequilibrium thermodynamics, gradients in the K_j of the *disorder*, defined as

$$S \equiv H(\mathbf{K}) - \sum_{j=1}^{m} K_j \partial H / \partial K_j \qquad (25)$$

become of central interest.

Equation 25 is similar to the definition of entropy in terms of the free energy density of a physical system, as suggested by the homology between free energy density and information source uncertainty described above.

Pursuing the homology further, the generalized Onsager relations defining temporal dynamics become

$$dK_j/dt = \sum_i L_{j,i} \partial S / \partial K_i, \qquad (26)$$

where the $L_{j,i}$ are, in first order, constants reflecting the nature of the underlying phenomena. The L-matrix is to be viewed empirically, in the same spirit as the slope and intercept of a regression model, and may have structure far different than familiar from simpler chemical or physical processes. The $\partial S / \partial K$ are analogous to thermodynamic forces in a chemical system, and may be subject to override by external physiological driving mechanisms: biological systems, unlike physical systems, can make choices as to energy allocation.

That is, an essential contrast with simple physical systems driven by (say) entropy maximization is that complex biological structures can make decisions about resource allocation, to the extent resources are available. Thus resource availability is a context, not a determinant.

Equations 25 and 26 can be derived in a simple parameter-free covariant manner which relies on the underlying topology of the information source space implicit to the development (e.g., Wallace and Wallace 2008). We will not pursue that development here.

The dynamics, as we have presented them so far, have been noiseless, while biological systems are always very noisy. Equation 26 might be rewritten as

$$dK_j/dt = \sum_i L_{j,i}\partial S/\partial K_i + \sigma W(t)$$

where σ is a constant and $W(t)$ represents white noise.

This leads directly to a family of classic stochastic differential equations having the form

$$dK_t^j = \sum_i L^{j,i}\big(t,\,\partial S/\partial K^i\big)dt + \sigma^j(t)dB_t^j, \tag{27}$$

where the $L^{i,j}$ and σ are appropriately regular functions of t and K^i, dB_t^j represents the noise structure, and we have readjusted the indices.

Further progress in this direction requires introduction of methods from stochastic differential geometry and related topics in the sense of Emery (1989), which will not be required in our approach. The obvious inference is that noise, which need not be 'white', can serve as a tool to shift the system between various modes, as a kind of crosstalk and the source of a generalized stochastic resonance.

Effectively, topological shifts between and within dynamic manifolds constitute another theory of phase transitions. Indeed, similar considerations have become central in the study of phase changes for physical systems (Franzosi and Pettini 2004).

Our phenomenological Onsager treatment would likely be much enriched by adoption of a Morse theory perspective.

Rate Distortion Dynamics

It is possible to restate Eqs. 26 and 27 in a manner which relates them closely to our central concerns. First recall the relation between information source uncertainty and channel capacity from Eq. 16:

$$H[\mathbf{X}] \leq C.$$

Next, the definition of channel capacity, Eq. 8:

$$C \equiv \max_{P(X)} I(X \mid Y).$$

Finally, the definition of the rate distortion function, from Eq. 21:

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

$R(D)$ defines the minimum channel capacity necessary for average distortion D, placing limits on information source uncertainty. Thus, we suggest that, as in Eigen's original model, distortion measures can drive information system dynamics.

We are led to propose, as a heuristic, that the dynamics of Eqs. 25, 26, and 27 will, through Eq. 16, be constrained by the system as described in terms of a parametized rate distortion function. To do this, take R as parametized, not only by the distortion D, but by some vector of variates $\mathbf{D} = (D_1, ..., D_k)$, for which the first component

is the average distortion. The assumed dynamics are then driven by gradients in the rate distortion disorder defined as

$$S_R \equiv R(\mathbf{D}) - \sum_{i=1}^{k} D_i \partial R / \partial D_i, \qquad (28)$$

leading to the deterministic and stochastic systems of equations

$$dD_j/dt = \sum_i L_{j,i} \partial S_R / \partial D_i \qquad (29)$$

and

$$dD_t^j = \sum_i L^{j,i}\big(t, \partial S_R / \partial D^i\big)dt + \sigma^J(t)dB_t^j. \qquad (30)$$

A simple Gaussian channel, taking $\sigma^2 = 1$, has a Rate Distortion function

$$R(D) = 1/2 \log[1/D],$$

so that,

$$S_R(D) = R(D) - D dS_R/dD = 1/2 \log(1/D) + 1/2. \qquad (31)$$

The simplest possible Onsager relation becomes

$$dD/dt \propto -dS_R/dD = 1/(2D), \qquad (32)$$

where $-dS_R/dD$ represents the force of the 'biochemical wind'.

This has the solution

$$D \propto \sqrt{t}. \qquad (33)$$

Similar results will accrue to any of the reverse-J-shaped relations which must inevitably characterize simple RNA-like replicators. The implication is that all simple RNA(-like) organisms – including RNA viruses – will inevitably be subject to a relentless biochemical force, a constant torrent, which can drive them very close to their critical mutation rates. That is, in general, absent a contravening biological or other constraint,

$$D(t) = f(t), \qquad (34)$$

with $f(t)$ monotonic increasing in t.

It is not surprising, then, that so many RNA viruses are found close to their error thresholds, which do indeed constitute a powerful biological constraint (Holmes 2003).

In the next section we will show how two or more interacting RNA-like vesicles can, given enough energy, by virtue of Van Valen's (1973) Red Queen dynamic, oppose this fierce chemical hurricane.

Prebiotic Coevolution

Natural systems subject to coevolutionary interaction may become enmeshed in the Red Queen dilemma of Alice in Wonderland, in that they must undergo constant evolutionary change in order to avoid extinction – they must constantly run just

to stay in the same place. An example would be a competitive arms race between predator and prey: Each evolutionary advance in predation must be met with a coevolutionary adaptation which allows the prey to avoid the more efficient predator. Otherwise the system will collapse, since a highly specialized predator can literally eat itself to extinction. Similarly, each prey defense must be matched by a predator adaptation for the system to persist.

Here we present a fairly elaborate model of coevolution, in terms of interacting information sources. Interaction events, we will argue, can be highly punctuated. These are assumed to be between prebiotic vesicles which engage in crosstalk.

We begin by examining ergodic information sources and their dynamics under the self-similarity of a renormalization transformation near a punctuated phase transition, using a simple version of the more elaborate arguments in the Appendix. We then study the linked interaction of two information sources in which the richness of the quasi-language of each affects the other, that is, when two information sources have become one another's primary environments. This leads directly and naturally to a coevolutionary Red Queen. We will generalize the development to a 'block diagonal' structure of several interacting sources, and extend the model to second order, using a large deviations argument, suggesting a means for the necessary series of punctuated biochemical evolutionary events.

Fragmentation and Coalescence

The structures of interest to us here can be most weakly, and hence universally, described in terms of an adiabatically, piecewise stationary, ergodic information source involving a stochastic variate X which, in some general sense, sends symbols α in correlated sequences $\alpha_0, \alpha_1...\alpha_{n-1}$ of length n (which may vary), according to a joint probability distribution, and its associated conditional probability distribution,

$$P\big[X_0 = \alpha_0, X_1 = \alpha_1, ... X_{n-1} = \alpha_{n-1}\big],$$

$$P\big[X_{n-1} = \alpha_{n-1} \mid X_0 = \alpha_0, ... X_{n-2} = \alpha_{n-2}\big].$$

If the conditional probability distribution depends only on m previous values of X, then the information source is said to be of order m (Ash 1990).

By 'ergodic' we mean that, in the long term, correlated sequences of symbols are generated at an average rate equal to their (joint) probabilities. 'Adiabatic' means that changes are slow enough to allow the necessary limit theorems to function, 'stationary' means that, between pieces, probabilities don't change (much), and 'piecewise' means that these properties hold between phase transitions, which are described using renormalization methods.

As the length of the (correlated) sequences increases without limit, the Shannon–McMillan Theorem permits division of all possible streams of symbols into two groups, a relatively small number characterized as meaningful, whose long-time behavior matches the underlying probability distribution, and an increasingly large set of gibberish with vanishingly small probability. Let $N(n)$ be the number of possible meaningful sequences of length n emitted by the source $\mathbf{X}$. Again, uncertainty of the source, $H[\mathbf{X}]$, can be defined by the relation

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

The Shannon–McMillan Theorem shows how to characterize $H[\mathbf{X}]$ directly in terms of the joint probability distribution of the source $\mathbf{X}$.

Let $P[x_i \mid y_j]$ be the conditional probability that stochastic variate $X = x_i$ given that stochastic variate $Y = y_j$ and let $P[x_i, y_j]$ be the joint probability that $X = x_i$ and $Y = y_j$. Then the joint and conditional uncertainties of X and Y, $H(X, Y)$, and $H(X \mid Y)$ are given by expressions like those of Eqs. 5 and 6.

And again, the Shannon–McMillan Theorem of states that the function for $H[\mathbf{X}]$ is determined according to the limits of Eq. 13.

Estimating the probabilities of the sequences $\alpha_0, \ldots \alpha_{n-1}$ from observation, the ergodic property allows us to use them to estimate the uncertainty of the information source. Thus $H[\mathbf{X}]$ is directly measurable.

Equation 16, $H[\mathbf{X}] \leq C$, establishes C as the maximum rate at which the external world can transmit information originating with the information source, or that internal vesicles can communicate. Much of the subsequent development could, in fact, be expressed using this relation, and the analog in terms of the rate distortion Eqs. 28–30.

Again recall the relation between the expression for source uncertainty and the free energy density of a physical system which undergoes a phase transition depending on an inverse temperature parameter $K = 1/T$ at a critical temperature T_C.

Imposition of a renormalization symmetry on $F(K)$ in Eq. 24 describes, in the infinite volume limit, the behavior of the system at the phase transition in terms of scaling laws (Wilson 1971). After some development, taking the limit $n \to \infty$ as an analog to the infinite volume limit of a physical system, we will apply this approach to a parametized source uncertainty. We will examine changes in structure as a fundamental 'inverse temperature' changes across the underlying system.

We use three parameters to describe the relations between an information source and its environment or between different interacting sources.

The first, $J \geq 0$, measures the degree to which acquired characteristics are transmitted. $J \approx 0$ thus represents a high degree of genetic as opposed to epigenetic inheritance.

J will always remain distinguished, a kind of inherent direction or external field strength in the sense of Wilson (1971).

The second parameter, $Q = 1/\mathscr{C} \geq 0$, represents the inverse availability of resources. $Q \approx 0$ thus represents a high ability to renew and maintain a particular system.

The third parameter, $K = 1/T$, is an inverse index of a generalized temperature T, which we will more directly specify below in terms of the richness of interacting information sources.

We suppose further that the structure of interest is implicitly embedded in, and operates within the context of, a larger manifold stratified by metric distances.

Take these as multidimensional vector quantities $\mathbf{A}$, $\mathbf{B}$, $\mathbf{C}$.... $\mathbf{A}$ may represent location in space, time delay, or the like, and $\mathbf{B}$ may be determined through multivariate analysis of a spectrum of observed behavioral or other factors, in the largest sense, etc.

It may be possible to reduce the effects of these vectors to a function of their magnitudes $a = \mid \mathbf{A} \mid$, $b = \mid \mathbf{B} \mid$ and $c = \mid \mathbf{C} \mid$, etc. Define a generalized distance r as

$$r^2 = a^2 + b^2 + c^2 + \ldots \; .$$ (35)

To be more explicit, we assume an ergodic information source $\mathbf{X}$ is associated with the reproduction and/or persistence of a biological population or other structure. The source $\mathbf{X}$, its uncertainty $H[J, K, Q, \mathbf{X}]$ and its parameters J, K, Q all are assumed to depend *implicitly* on the embedding manifold, in particular on the metric r of Eq. 35.

A particularly elegant and natural formalism for generating such punctuation in our context involves application of Wilson's (1971) program of renormalization symmetry – invariance under the renormalization transform – to source uncertainty defined on the r-manifold. The results predict that language in the most general sense, which includes the transfer of information within a biological or prebiotic context, will undergo sudden changes in structure analogous to phase transitions in physical systems.

We emphasize that this is an argument by abduction from physical theory: Much current development surrounding physical phase change is based on the assumption that at transition a system looks the same under renormalization. That is, phase transition represents a stationary point for a renormalization transform in the sense that the transformed quantities are related by simple scaling laws to the original values.

Renormalization is a clustering semigroup transformation in which individual components of a system are combined according to a particular set of rules into a 'clumped' system whose behavior is a simplified average of those components. Since such clumping is a many-to-one condensation, there can be no unique inverse renormalization, and, as the Appendix shows, many possible forms of condensation.

Assume it possible to redefine characteristics of the information source $\mathbf{X}$ and J, K, Q as functions of averages across the manifold having metric r, which we write as R. That is, 'renormalize' by clustering the entire system in terms of blocks of different sized R.

Let $N(K, J, Q, n)$ be the number of high probability meaningful correlated sequences of length n *across the entire community* in the r-manifold of Eq. 35, given parameter values K, J, Q. We study changes in

$$H[K, J, Q, \mathbf{X}] \equiv \lim_{n \to \infty} \frac{\log[N(K, J, Q, n)]}{n}$$

as $K \to K_C$ and/or $Q \to Q_C$ for critical values K_C, Q_C at which the system begins to undergo a marked transformation from one kind of structure to another.

Given the metric of Eq. 35, a *correlation length*, $\chi(K, J, Q)$, can be defined as the average length in r-space over which structures involving a particular phase dominate.

Now clump the 'community' into blocks of average size R in the multivariate r-manifold, the 'space' in which the system of interest is implicitly embedded.

Following the classic argument of Wilson (1971), reproduced and expanded in the Appendix, it is possible to impose renormalization symmetry on the source uncertainty on H and χ by assuming at transition the relations

$$H[K_R, J_R, Q_R, \mathbf{X}] = R^\delta H[K, J, Q, \mathbf{X}] \tag{36}$$

and

$$\chi(K_R, J_R, Q_R) = \chi(K, J, Q)/R \tag{37}$$

hold, where K_R, J_R and Q_R are the transformed values of K, J and Q after the clumping of renormalization. We take $K_1, J_1, Q_1 \equiv K, J, Q$ and permit the characteristic exponent $\delta > 0$ to be nonintegral. The Mathematical Appendix provides examples of other possible relations.

Equations 36 and 37 are assumed to hold in a neighborhood of the transition values K_C and Q_C.

Differentiating these with respect to R gives expressions for dK_R/dR, dJ_R/dR and dQ_R/dR depending simply on R which we write as

$$dK_R/dR = u(K_R, J_R, Q_R)/R$$
$$dQ_R/dR = w(K_R, J_R, Q_R)/R$$
$$dJ_R/dR = [v(K_R, J_R, Q_R)/R]J_R. \tag{38}$$

Solving these differential equations gives K_R, J_R and Q_R as functions of J, K, Q and R.

Substituting back into Eqs. 36 and 37 and expanding in a first order Taylor series near the critical values K_C and Q_C gives power laws much like the Widom-Kadanoff relations for physical systems (Wilson 1971). For example, letting $J = Q = 0$ and taking $\kappa \equiv (K_C - K)/K_C$ gives, in first order near K_C,

$$H = \kappa^{\delta/y} H_0$$
$$\chi = \kappa^{-1/y} \chi_0 \tag{39}$$

where y is a constant arising from the series expansion.

Note that there are only two fundamental equations – Eqs. 36 and 37 – in $n > 2$ unknowns: The critical 'point' is, in this formulation, most likely to be a complicated implicitly defined critical surface in J, K, Q, ...-space. The 'external field strength' J remains distinguished in this treatment, i.e., the inverse of the degree to which acquired characteristics are inherited, but *neither K, Q nor other parameters are, by themselves, fundamental*, rather their joint interaction defines critical behavior along this surface.

That surface is a fundamental object, not the particular set of parameters (except for J) used to define it, which may be subject to any set of transformations which leave the surface invariant. Thus inverse generalized temperature, resource availability, or whatever other parameters may be identified as affecting H, are inextricably intertwined and mutually interacting, according to the form of this critical surface. That surface, in turn, is unlikely to remain fixed, and should vary with time or other extrinsic parameters, including, but not likely limited to, J.

At the critical surface a Taylor expansion of the renormalization Eqs. 36 and 37 gives a first order matrix of derivatives whose eigenstructure defines fundamental system behavior. For physical systems the surface is a saddle point (Wilson 1971), but more complicated behavior seems likely in what we study. See Binney et al. (1986) for some details of this differential geometry.

Taking, for the moment, the simplest formulation, $(J = Q = 0)$, that is, a well-provisioned system with memory, as K increases toward a threshold value K_C, the source uncertainty declines and, at K_C, the average regime dominated by the 'other phase' grows. That is, the system begins to freeze into one having a large correlation

length for the second phase. The two phenomena are linked at criticality in physical systems by the scaling exponent y.

Assume the rate of change of $\kappa = (K_C - K)/K_C$ remains constant, $| d\kappa/dt |= 1/\tau_K$. Analogs with physical theory suggest there is a characteristic time constant for the phase transition, $\tau \equiv \tau_0/\kappa$, such that if changes in κ take place on a timescale longer than τ for any given κ, we may expect the correlation length $\chi = \chi_0\kappa^{-s}$, $s = 1/y$, will be in equilibrium with internal changes and result in a very large fragment in r-space. Following Zurek (1985, 1996), the 'critical' freezout time, $\hat{t}$, will occur at a 'system time' $\hat{t} = \chi / | d\chi/dt |$ such that $\hat{t} = \tau$. Taking the derivative $d\chi/dt$, remembering that by definition $d\kappa/dt = 1/\tau_K$, gives

$$\frac{\chi}{| d\chi/dt |} = \frac{\kappa \tau_K}{s} = \frac{\tau_0}{\kappa}$$

so that

$$\kappa = \sqrt{s\tau_0/\tau_K}.$$

Substituting this value of κ into the equation for correlation length, the expected size of fragments in r-space, $d(\hat{t})$, becomes

$$d \approx \chi_0 \left(\frac{\tau_K}{s\tau_0} \right)^{s/2}$$

with $s = 1/y > 0$. The more rapidly K approaches K_C the smaller is τ_K and the smaller and more numerous are the resulting r-space fragments. Thus rapid change produces small fragments more likely to risk extinction in a system dominated by economies of scale.

An absolutely central observation is that this argument suggests the possibility of a reverse transformation: from a set of weakly linked, quasi-independent vesicles in simple loose association, into a large, coherent, strongly-linked structure in which the vesicles are now intimate, mutually-supporting, subcomponents. From a Morse Theory perspective, this must involve a fundamental topological transformation of the underlying biochemical networks.

The next section explores the dynamics of coevolutionary coherence in more detail.

Recursive Interaction

The next iteration of the argument involves envisioning reciprocally interacting biological information sources – vesicles – as subject to a coevolutionary Red Queen by treating their respective reproductive source uncertainties as recursively parametized by each other. That is, we *assume the information sources are each other's primary environments*. These are, respectively, characterized by information sources $\mathbf{X}$ and $\mathbf{Y}$, whose uncertainties are parametized

[1]   by measures of both inheritance and inverse resources – J, Q as above – and, most critically,

[2]   by each others inverse uncertainties, $\mathscr{H}_X \equiv 1/H[\mathbf{X}]$ and $\mathscr{H}_Y \equiv 1/H[\mathbf{Y}]$, i.e.,

$$H[\mathbf{X}] = H[Q, J, \mathscr{H}_Y, \mathbf{X}]$$

$$H[\mathbf{Y}] = H[Q, J, \mathscr{H}_X, \mathbf{Y}]. \qquad (40)$$
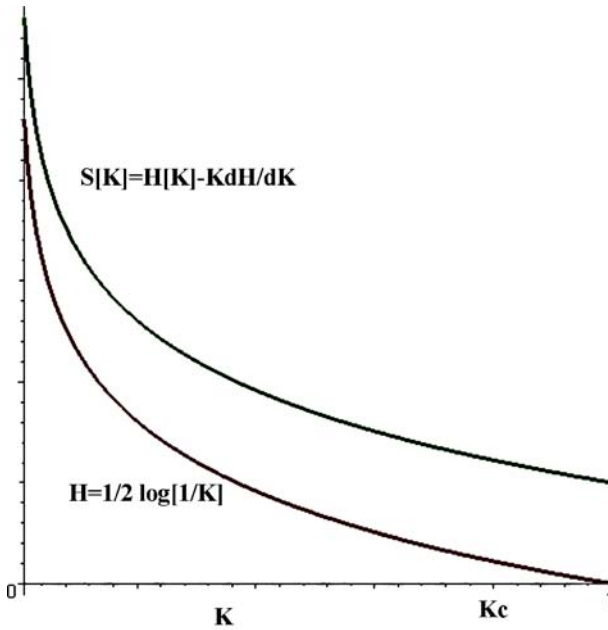
Fig. 1 A reverse-*J*-shaped curve for source uncertainty $H[\mathbf{X}]$ – measuring language richness – as a function of an inverse temperature parameter $K = 1/H[\mathbf{Y}]$. To the right of the critical point K_C the system breaks into fragments in *r*-space whose size is determined by the rate at which K approaches K_C. A collection of fragments already to the right of K_C, however, would be seen as condensing into a single unit as K declined below the critical point. If K is an inverse source uncertainty itself, i.e., $K = 1/H[\mathbf{Y}]$ for some information source $\mathbf{Y}$, then under such conditions a Red Queen dynamic can become enabled, driving the system strongly to the left. No intermediate points are asymptotically stable in this development. To the right of the critical point K_C the system is locked into disjoint fragments. 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

This is a recursive system having complicated behaviors.

Assume a strongly heritable genetic system, $J = 0$, with fixed inverse resource base, Q, for which $H[\mathbf{X}]$ follows something like the lower graph in Fig. 1, a reverse J-shaped curve with $K \equiv \mathcal{H}_Y = 1/H[\mathbf{Y}]$, and similarly $H[\mathbf{Y}]$ depends on $\mathcal{H}_X$. That is, increase or decline in the source uncertainty of one system leads to increase or decline in the source uncertainty of the other, and vice versa. The richness of the two information sources is closely linked.

Start at the right of the lower graph for $H[\mathbf{X}]$ in Fig. 1, the source uncertainty of the first system, but to the left of the critical point K_C. Assume $H[\mathbf{Y}]$ increases so $\mathcal{H}_Y$ decreases, and thus $H[\mathbf{X}]$ increases, walking up the lower curve of Fig. 1 from the right: the richness of the first system's internal language increases.

The increase of $H[\mathbf{X}]$ leads, in turn, to a decline in $\mathcal{H}_X$ and triggers an increase of $H[\mathbf{Y}]$, whose increase leads to a further increase of $H[\mathbf{X}]$ and vice versa: The Red Queen, taking the system from the right of Fig. 1 to the left, up the lower curve as the two systems mutually interact.

Now enlarge the scale of the argument, and consider the possibility of other interactions.

The upper graph of Fig. 1 represents the disorder

$$S = H[K, \mathbf{X}] - K dH[K, \mathbf{X}]/dK, \ K \equiv 1/H[\mathbf{Y}],$$

whose gradient, in the absence of the Red Queen interaction, would simply drive the system toward the minimum energy critical point for this system, as in the section on rate distortion dynamics.

Clearly this ratchet can also be reversed. Thus the system has two quasi-stable limit points, a low energy solution near the error limit phase transition point, and a high energy state near to, but never at, the zero error limit, depending on the availability of sufficient energy to the system. Absent a relatively high energy source, low error rate reproduction would be impossible, according to the model. This suggests that some major, large scale, ecosystem transformation in energy availability was a necessary condition for low error rate reproduction.

This simple model has a surprising number of implications that are not confined to a prebiotic context. It is possible to formulate the contest between a highly mutating pathogen, operating at a low energy configuration, and a complicated immune system, in terms of phenotype, rather than simply genotype, coevolution, in the sense of West-Eberhard (2003). Then the phenotype-phenotype 'two-vesicle' coevolutionary ratchet has, again, two quasi-fixed points: fragmentation near a high variability phenotype, and immune/viral phenotype coalescence near a low variability phenotype. The latter will, ultimately, write an image of itself, in the sense of Adami et al. (2000), on the pathogen gene in spite of its high mutation rate, producing protected zones defining that phenotype.

Extending the Model

The model directly generalizes to multiple interacting information sources.

First consider a matrix of crosstalk measures between a set of information sources. Assume the matrix of mutual information crosstalk can be block diagonalized into two major components, characterized by mutual information source measures like

$$I_m(X_1...X_i), m = 1, 2.$$

Then apply the two-component theory above.

Extending the development to multiple, recursively interacting information sources resulting from a more general block diagonalization seems direct. First use inverse measures $\mathscr{I}_j \equiv 1/I_j$, $j \neq m$ as parameters for each of the other blocks, writing

$$I_m = I_m(K_1...K_s, ...\mathscr{I}_j...), \ j \neq m$$

where the K_s represent other relevant parameters.

Next segregate the $\mathscr{I}_j$ according to their relative rates of change.

The dynamics of such a system becomes a recursive network of stochastic differential equations, similar to those used to study many other highly parallel dynamic structures (e.g., Wymer 1997).

Letting the K_j and $\mathscr{I}_m$ all be represented as parameters Q_j, (with the caveat that I_m not depend on $\mathscr{I}_m$), one can define

$$S_I^m \equiv I_m - \sum_i Q_i \partial I_m/\partial Q_i$$

to obtain a complicated recursive system of phenomenological 'Onsager relations' stochastic differential equations like Eq. 27 or, in terms of rate distortion functions, like Eq. 30,

$$dQ_t^j = \sum_i \left[L_{j,i}\left(t, ...\partial S_l^m/\partial Q^i...\right)dt + \sigma_{j,i}\left(t, ...\partial S_l^m/\partial Q^i...\right)dB_t^i \right], \qquad (41)$$

where, again, for notational simplicity only, we have expressed both the reciprocal $\mathscr{I}$'s and the external K's in terms of the same Q_j.

The index m ranges over the crosstalk and we could allow different kinds of 'noise' dB_t^i, having particular forms of quadratic variation which may, in fact, represent a projection of environmental factors under something like a rate distortion manifold (Wallace and Wallace 2008).

Indeed, the I_m and/or the derived S^m might, in some circumstances, be combined into a Morse function, permitting application of Pettini's Topological Hypothesis.

The model rapidly becomes unwieldy, probably requiring some clever combinatorial or groupoid convolution algebra and related diagrams for concise expression, much as in the usual field theoretic perturbation expansions (Hopf algebras, for example). The virtual reaction method of Zhu et al. (2007) is another possible approach.

As in the simple model above, there will be, first, multiple quasi-stable points within a given system's I_m, representing a class of generalized resilience modes accessible via punctuation, enmeshing the crosstalking vesicles in a manner analogous to the simple model of Fig. 1, and depending critically on available energy.

Second, however, will be analogs to the fragmentation of Fig. 1 when the system exceeds the critical value K_c. That is, the K-parameter structure will represent full-scale fragmentation of the entire structure, and not just punctuation within it.

There are other possible patterns:

[1] Setting Eq. 41 equal to zero and solving for stationary points – 'coevolutionary stable strategies' – again gives attractor states since the noise terms preclude unstable equilibria.

[2] However, this system may also converge to limit cycle or 'strange attractor' behaviors in which it seems to chase its tail endlessly.

[3] What is converged to in both cases is not a simple state or limit cycle of states. Rather it is an equivalence class, or set of them, of highly dynamic information sources coupled by mutual interaction through crosstalk. Thus 'stability' in this structure represents particular patterns of ongoing dynamics rather than some identifiable 'state'. This leads directly to elaborate groupoid descriptions of dynamic manifolds, and to consideration of directed homotopy equivalence classes which we will not explore further here (Wallace and Wallace 2008).

We are, then, deeply enmeshed in a highly recursive phenomenological stochastic differential equations, but in a dynamic rather than static manner: the objects of this dynamical system are equivalence classes of information sources and their crosstalk, rather than simple 'states' of a dynamical or reactive chemical system.

Other formulations may well be possible, but our work here serves to illustrate the method.

It is, however, interesting to compare our results to those of Dieckmann and Law (1996), who invoke evolutionary game dynamics to obtain a first order coevolutionary canonical equation having the form

$$ds_i/dt = K_i(s)\partial W_i(s_i', s) \mid_{s_i'=s_i} . \tag{42}$$

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

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

Dieckmann and Law contend that their formal derivation of this equation satisfies four critical requirements:

[1]   The evolutionary process needs to be considered in a coevolutionary context.
[2]   A proper mathematical theory of evolution should be dynamical.
[3]   The coevolutionary dynamics ought to be underpinned by a microscopic theory.
[4]   The evolutionary process has important stochastic elements.

Our Eq. 41 seems clearly within this same ballpark, although we have taken a much different route, one which indeed produces elaborate patterns of phase transition punctuation in a highly natural manner. Champagnat et al. (2006), in fact, derive a higher order canonical approximation extending Eq. 42 which is very much closer equation to Eq. 41, that is, a stochastic differential equation describing evolutionary dynamics.

Champagnat et al. (2006) go even further, using a large deviations argument to analyze dynamical coevolutionary paths, not merely evolutionary singularities. They contend that in general, the issue of evolutionary dynamics drifting away from trajectories predicted by the canonical equation can be investigated by considering the asymptotic of the probability of 'rare events' for the sample paths of the diffusion. By 'rare events' they mean diffusion paths drifting far away from the canonical equation.

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

An essential fact of large deviations theory is that the rate function I which Champagnat et al. (2006) invoke can almost always be expressed as a kind of entropy, that is, in the form $I = -\sum_j P_j \log(P_j)$ for some probability distribution. This result

goes under a number of names; Sanov's Theorem, Cramer's Theorem, the Gartner-Ellis Theorem, the Shannon–McMillan Theorem, and so forth (Dembo and Zeitouni 1998). Here we will use it to suggest the possibility of second order effects in prebiotic coevolutionary process. The fluctuational paths defined by the system of equations in Eq. 41 may, over sufficient time, lead to relatively sudden, and very marked, evolutionary excursions.

Large Deviations in Prebiotic Evolution

We begin with a recapitulation of large deviations and fluctuation formalism.

Information source uncertainty, according to the Shannon–McMillan Theorem, serves as a splitting criterion between high and low probability sequences (or pairs of them) and displays the fundamental characteristic of a growing body of work in applied probability often termed the Large Deviations Program, (LDP) which seeks to unite information theory, statistical mechanics and the theory of fluctuations under a single umbrella.

Following Dembo and Zeitouni (1998, p.2),

Let $X_1, X_2, ... X_n$ be a sequence of independent, standard Normal, real-valued random variables and let

$$S_n = (1/n) \sum_{j=1}^{n} X_j. \tag{43}$$

Since S_n is again a Normal random variable with zero mean and variance $1/n$, for all $\delta > 0$

$$\lim_{n \to \infty} P(| S_n | \geq \delta) = 0, \tag{44}$$

where P is the probability that the absolute value of S_n is greater or equal to δ. Some manipulation, however, gives

$$P(| S_n | \geq \delta) = 1 - \frac{1}{\sqrt{2\pi}} \int_{-\delta\sqrt{n}}^{\delta\sqrt{n}} \exp\left(-x^2/2\right) dx, \tag{45}$$

so that

$$\lim_{n \to \infty} \frac{\log P(| S_n | \geq \delta)}{n} = -\delta^2/2. \tag{46}$$

This can be rewritten for large n as

$$P(| S_n | \geq \delta) \approx \exp\left(-n\delta^2/2\right). \tag{47}$$

That is, for large n, the probability of a large deviation in S_n follows something much like the asymptotic equipartition relation of the Shannon–McMillan Theorem, i.e., that meaningful paths of length n all have approximately the same probability $P(n) \propto \exp(-nH[\mathbf{X}])$.

Questions about meaningful paths appear suddenly as formally isomorphic to the central argument of the LDP which encompasses statistical mechanics, fluctuation theory, and information theory into a single structure (Dembo and Zeitouni 1998).

Again, the cardinal tenet of large deviation theory is that the rate function $-\delta^2/2$ can, under proper circumstances, be expressed as a mathematical entropy having the standard form

$$-\sum_k p_k \log p_k, \tag{48}$$

for some distribution p_k.

Next we briefly recapitulate part of the standard treatment of large fluctuations (Onsager and Machlup 1953; Fredlin and Wentzell 1998).

The macroscopic behavior of a complicated physical system in time is assumed to be described by the phenomenological Onsager relations giving large-scale fluxes as

$$\sum_i R_{i,j} dK_j/dt = \partial S/\partial K_i, \tag{49}$$

where the $R_{i,j}$ are appropriate constants, S is the system entropy and the K_i are the generalized coordinates which parametize the system's free energy.

Entropy is defined from free energy F by a Legendre transform – more of which follows below:

$$S \equiv F - \sum_j K_j \partial F/\partial K_j,$$

where the K_j are appropriate system parameters.

Neglecting volume problems for the moment, free energy can be defined from the system's partition function Z as

$$F(K) = \log[Z(K)].$$

The partition function Z, in turn, is defined from the system Hamiltonian – defining the energy states – as

$$Z(K) = \sum_j \exp[-KE_j],$$

where K is an inverse temperature or other parameter and the E_j are the energy states.

Inverting the Onsager relations gives

$$dK_i/dt = \sum_j L_{i,j} \partial S/\partial K_j = L_i(K_1, ..., K_m, t) \equiv L_i(K, t). \tag{50}$$

The terms $\partial S/\partial K_i$ are macroscopic driving forces dependent on the entropy gradient.

Let a white Brownian noise $\varepsilon(t)$ perturb the system, so that

$$dK_i/dt = \sum_j L_{i,j} \partial S/\partial K_j + \varepsilon(t) = L_i(K, t) + \varepsilon(t), \tag{51}$$

where the time averages of ε are $<\varepsilon(t)> = 0$ and $<\varepsilon(t)\varepsilon(0)> = D\delta(t)$. $\delta(t)$ is the Dirac delta function, and we take K as a vector in the K_i.

Following Luchinsky (1997), if the probability that the system starts at some initial macroscopic parameter state K_0 at time $t = 0$ and gets to the state $K(t)$ at time t is

$P(K, t)$, then a somewhat subtle development (e.g., Feller 1971) gives the forward Fokker–Planck equation for P:

$$\partial P(K, t)/\partial t = -\nabla \cdot (L(K, t)P(K, t)) + (D/2)\nabla^2 P(K, t). \tag{52}$$

In the limit of weak noise intensity this can be solved using the WKB, i.e., the eikonal, approximation, as follows: take

$$P(K, t) = z(K, t) \exp(-s(K, t)/D). \tag{53}$$

$z(K, t)$ is a prefactor and $s(K, t)$ is a classical action satisfying the Hamilton–Jacobi equation, which can be solved by integrating the Hamiltonian equations of motion. The equation reexpresses $P(K, t)$ in the usual parametized negative exponential format.

Let $p \equiv \nabla s$. Substituting and collecting terms of similar order in D gives

$$dK/dt = p + L, \quad dp/dt = -\partial L/\partial Kp \tag{54}$$

and

$$-\partial s/\partial t \equiv h(K, p, t) = pL(K, t) + \frac{p^2}{2}, \tag{55}$$

with $h(K, t)$ the Hamiltonian for appropriate boundary conditions.

Again following Luchinsky (1997), these Hamiltonian equations have two different types of solution, depending on p. For $p = 0$, $dK/dt = L(K, t)$ which describes the system in the absence of noise. We expect that with finite noise intensity the system will give rise to a distribution about this deterministic path. Solutions for which $p \neq 0$ correspond to optimal paths along which the system will move with overwhelming probability.

In sum, to again paraphrase Luchinsky (1997), large fluctuations, although infrequent, are fundamental in a broad range of processes, and it was recognized by Onsager and Machlup (1953) that insight into the problem could be gained from studying the distribution of fluctuational paths along which the system moves to a given state. This distribution is a fundamental characteristic of the fluctuational dynamics, and its understanding leads toward control of fluctuations. Fluctuational motion from the vicinity of a stable state may occur along different paths. For large fluctuations, the distribution of these paths peaks sharply along an optimal, most probable, path. In the theory of large fluctuations, the pattern of optimal paths plays a role similar to that of the phase portrait in nonlinear dynamics.

The essential generalization is that, by invoking the results of the section on rate distortion dynamics, we can substitute the rate distortion function for free energy in this argument. For the two-vesicle system of Fig. 1 this suggests the possibility of a transition from the low energy, high error state to the high energy, low error condition, assuming enough time and energy are available.

For more complicated cooperative prebiotic assemblies of RNA-like vesicles, these sudden, relatively large, jet-like transitions might – again given enough time and available energy – lead to far more complicated structures having error correction capabilities within the overall constraints of the particular chemical system. A

central question then revolves around what classes of systems would be more likely to make such a transition. These might be the ones best sought via observation *ex planeta*.

Discussion and Conclusions

Like many others, we have resolved Eigen's paradox by proposing appropriate inter-action between compartments consisting of closely related populations of prebiotic mutants carrying a 'consensus sequence'. These are isolated in vesicles, which are themselves linked. Our innovation has been to impose a formalism driven by the asymptotic limit theorems of information theory. Rather than abduct spin glass, two dimensional Ising, or similar models from physics, we have used the homology between information source uncertainty and free energy density to invoke highly modified biological phase change and generalized Onsager relation schemes in order to describe the dynamics of information processing systems at, and away from, phase transition analogs. A restatement of Eigen's model in rate distortion terms, compounded with a coevolutionary perspective, suggests RNA virus-like structures are likely to be found as either a fragmented community, with each mutational population driven toward the mutation error limit, or else in a cooperative state of linked vesicles which, collectively, can have a relatively low reproductive error rate, depending on available energy.

More complicated coevolutionary dynamics seem possible, not only 'quasi-evolutionarily stable strategies', but even analogs to limit cycles or pseudorandom strange attractors, all having various rates of reproductive error.

A large deviations argument suggests that certain of these prebiotic chemical systems may, given time, undergo highly structured excursions among their various possible equilibrium or quasi-equilibrium modes, depending on available energy. The more viable of these collective modes having low error rates would then become subject to further evolutionary process, resulting in a condensation from RNA-like to DNA-like error correcting structures, where the various vesicles become ever more closely intertwined. Energy availability may indeed be the critical matter, with the probability of a 'large deviation' dependent on development of new energy sources. The inverse perspective is that the availability of a new energy source – a new metabolic cycle, or elementary photofixation, for example – can drive large deviations, and hence evolutionary process. One can argue that large scale energy availability is enmeshed with large scale ecosystem resilience shifts, sensu Holling (1973), another example in which resilience drives evolution, albeit in a prebiotic circumstance (Wallace and Wallace 2008).

Our relatively systematic invocation of core concepts from information theory apparently marks a departure from current theorizing on these subjects, although we have, in fact, purged the groupoid and topological foundations of a 'resilience' analy-sis of similar topics (Wallace and Wallace 2008). These permit further classification of quasi-stable sets within the dynamic manifolds implicit to the development here. That is, a far richer theoretical structure is available, and will probably be required: One is, in spite of all the formal heavy lifting of this paper, still somehow reminded of the famous cartoon in which a white-coated scientist, standing in front of a blackboard filled with mathematical scribbles, turns to his audience, chalk in hand, saying "...and here a miracle occurs...".

Nonetheless, we have added significantly to the store of needed theoretical building blocks. The keys to further progress, however, seem, as usual, to lie in more extensive and penetrating observational and laboratory studies.

Mathematical Appendix

Morse Theory

Morse theory examines relations between analytic behavior of a function – the location and character of its critical points – and the underlying topology of the manifold on which the function is defined. We are interested in a number of such functions, for example information source uncertainty on a parameter space and 'second order' iterations involving parameter manifolds determining critical behavior, for example sudden onset of a giant component in the mean number model (Wallace and Wallace 2008), and universality class tuning in the mean field model of the next section. These can be reformulated from a Morse theory perspective. Here we follow closely the elegant treatments of Pettini (2007) and Kastner (2006).

The essential idea of Morse theory is to examine an n-dimensional manifold M as decomposed into level sets of some function $f : M \to \mathbf{R}$ where $\mathbf{R}$ is the set of real numbers. The a-level set of f is defined as

$$f^{-1}(a) = \{x \in M : f(x) = a\},$$

the set of all points in M with $f(x) = a$. If M is compact, then the whole manifold can be decomposed into such slices in a canonical fashion between two limits, defined by the minimum and maximum of f on M. Let the part of M below a be defined as

$$M_a = f^{-1}(-\infty, a] = \{x \in M : f(x) \le a\}.$$

These sets describe the whole manifold as a varies between the minimum and maximum of f.

Morse functions are defined as a particular set of smooth functions $f : M \to \mathbf{R}$ as follows. Suppose a function f has a critical point x_c, so that the derivative $df(x_c) = 0$, with critical value $f(x_c)$. Then f is a Morse function if its critical points are nondegenerate in the sense that the Hessian matrix of second derivatives at x_c, whose elements, in terms of local coordinates are

$$H_{i,j} = \partial^2 f / \partial x^i \partial x^j,$$

has rank n, which means that it has only nonzero eigenvalues, so that there are no lines or surfaces of critical points and, ultimately, critical points are isolated.

The index of the critical point is the number of negative eigenvalues of H at x_c.

A level set $f^{-1}(a)$ of f is called a critical level if a is a critical value of f, that is, if there is at least one critical point $x_c \in f^{-1}(a)$.

Again following Pettini (2007), the essential results of Morse theory are:

[1] If an interval $[a, b]$ contains no critical values of f, then the topology of $f^{-1}[a, v]$ does not change for any $v \in (a, b)$. Importantly, the result is valid even if f is not a Morse function, but only a smooth function.

[2] If the interval $[a, b]$ contains critical values, the topology of $f^{-1}[a, v]$ changes in a manner determined by the properties of the matrix H at the critical points.

[3] If $f : M \to \mathbf{R}$ is a Morse function, the set of all the critical points of f is a discrete subset of M, i.e., critical points are isolated. This is Sard's Theorem.

[4] If $f : M \to \mathbf{R}$ is a Morse function, with M compact, then on a finite interval $[a, b] \subset \mathbf{R}$, there is only a finite number of critical points p of f such that $f(p) \in [a, b]$. The set of critical values of f is a discrete set of $\mathbf{R}$.

[5] For any differentiable manifold M, the set of Morse functions on M is an open dense set in the set of real functions of M of differentiability class r for $0 \le r \le \infty$.

[6] Some topological invariants of M, that is, quantities that are the same for all the manifolds that have the same topology as M, can be estimated and sometimes computed exactly once all the critical points of f are known: Let the Morse numbers $\mu_i (i = 1, ..., m)$ of a function f on M be the number of critical points of f of index i, (the number of negative eigenvalues of H). The Euler characteristic of the complicated manifold M can be expressed as the alternating sum of the Morse numbers of any Morse function on M,

$$\chi = \sum_{i=0}^{m} (-1)^i \mu_i.$$

The Euler characteristic reduces, in the case of a simple polyhedron, to

$$\chi = V - E + F$$

where V, E, and F are the numbers of vertices, edges, and faces in the polyhedron.

[7] Another important theorem states that, if the interval $[a, b]$ contains a critical value of f with a single critical point x_c, then the topology of the set M_b defined above differs from that of M_a in a way which is determined by the index, i, of the critical point. Then M_b is homeomorphic to the manifold obtained from attaching to M_a an i-handle, i.e. the direct product of an i-disk and an $(m - i)$-disk.

Again, Pettini (2007) contains both mathematical details and further references. See, for example, Matsumoto (2002) or the classic by Milnor (1963).

The Mean Field Model

Wallace and Wallace (1998, 1999) have addressed how a language, in a large sense, 'spoken' on a network structure, responds as properties of the network change. The properties of interest were the magnitude of 'strong' or 'weak' ties which, respectively, either disjointly partitioned the network or linked it across such partitioning. These would be analogous to local and mean-field couplings in physical systems.

Fix the magnitude of strong ties – again, those which disjointly partition the underlying network into submodules – but vary the index of nondisjunctive weak ties, P, between these components, taking $K = 1/P$.

Assume the piecewise, adiabatically stationary ergodic information source (or sources) of interest depends on three parameters, two explicit and one implicit. The explicit are K as above and, as a calculational device, an 'external field strength' analog J, which gives a 'direction' to the system. We will, in the limit, set $J = 0$. Note that many other approaches may well be possible, since renormalization techniques allow a variety of approaches.

The implicit parameter, r, is an inherent generalized 'length' characteristic of the phenomenon, on which J and K are defined. That is, J and K are written as functions of averages of the parameter r, which may be quite complex, having nothing at all to do with conventional ideas of space. For example r may be defined by the degree of niche partitioning in ecosystems or separation in social structures.

For a given generalized language of interest having a well defined (adiabatically, piecewise stationary) ergodic source uncertainty, $H = H[K, J, \mathbf{X}]$.

To summarize a long train of standard argument (Binney et al. 1986; Wilson 1971), imposition of invariance of H under a renormalization transform in the implicit parameter r leads to expectation of both a critical point in K, written K_C, reflecting a phase transition to or from collective behavior across the entire array, and of power laws for system behavior near K_C. Addition of other parameters to the system results in a 'critical line' or surface.

Let $\kappa \equiv (K_C - K)/K_C$ and take χ as the 'correlation length' defining the average domain in r-space for which the information source is primarily dominated by 'strong' ties. The first step is to average across r-space in terms of 'clumps' of length $R = <r>$. Then $H[J, K, \mathbf{X}] \rightarrow H[J_R, K_R, \mathbf{X}]$.

Taking Wilson's (1971) analysis as a starting point, the 'renormalization relations' used here are:

$$H[K_R, J_R, \mathbf{X}] = f(R)H[K, J, \mathbf{X}]$$

$$\chi(K_R, J_R) = \chi(K, J)/R, \qquad (56)$$

with $f(1) = 1$ and $J_1 = J$, $K_1 = K$. The first equation significantly extends Wilson's treatment. It states that 'processing capacity,' as indexed by the source uncertainty of the system, representing the 'richness' of the generalized language, grows monotonically as $f(R)$, which must itself be a dimensionless function in R, since both $H[K_R, J_R]$ and $H[K, J]$ are themselves dimensionless. Most simply, this requires replacing R by R/R_0, where R_0 is the 'characteristic length' for the system over which renormalization procedures are reasonable, then setting $R_0 \equiv 1$, hence measuring length in units of R_0.

Wilson's original analysis focused on free energy density. Under 'clumping,' densities must remain the same, so that if $F[K_R, J_R]$ is the free energy of the clumped system, and $F[K, J]$ is the free energy density before clumping, then Wilson's Eq. 4 is $F[K, J] = R^{-3}F[K_R, J_R]$,

$$F[K_R, J_R] = R^3 F[K, J].$$

Remarkably, the renormalization equations are solvable for a broad class of functions $f(R)$, or more precisely, $f(R/R_0)$, $R_0 \equiv 1$.

The second equation just states that the correlation length simply scales as R.

Again, the central feature of renormalization in this context is the assumption that, at criticality, the system looks the same at all scales, that is, it is *invariant under renormalization* at the critical point. All else flows from this.

There is no unique renormalization procedure for information sources: other, very subtle, symmetry relations – not necessarily based on the elementary physical analog we use here – may well be possible. For example, McCauly (1993, p.168) describes the highly counterintuitive renormalizations needed to understand phase transition in simple 'chaotic' systems. This is important, since biological systems may well alter their renormalization properties – equivalent to tuning their phase transition dynamics – in response to external signals.

To begin, following Wilson, take $f(R) = R^d$, d some real number $d > 0$, and restrict K to near the 'critical value' K_C. If $J \to 0$, a simple series expansion and some clever algebra (Wilson 1971; Binney et al. 1986) gives

$$H = H_0 \kappa^\alpha$$

$$\chi = \chi_0/\kappa^s, \tag{57}$$

where α, s are positive constants. More biologically relevant examples appear below.

Further from the critical point, matters are more complicated, appearing to involve Generalized Onsager Relations, 'dynamical groupoids', and a kind of thermodynamics associated with a Legendre transform of H: $S[K] \equiv H[K] - KdH/dK$.

Regardless of the particular renormalization properties, sudden critical point transition is possible in the opposite direction for this model. That is, going from a number of independent, isolated and fragmented systems operating individually and more or less at random, into a single large, interlocked, coherent structure, once the parameter K, the inverse strength of weak ties, falls below threshold, or, conversely, once the strength of weak ties parameter $P = 1/K$ becomes large enough.

This heuristic insight can be made more exact using a rate distortion argument (or, more generally, using the Joint Asymptotic Equipartition Theorem) as follows:

Suppose that two ergodic information sources **Y** and **B** begin to interact, to 'talk' to each other, to influence each other in some way so that it is possible, for example, to look at the output of **B** – strings b – and infer something about the behavior of **Y** from it – strings y. We suppose it possible to define a retranslation from the B-language into the Y-language through a deterministic code book, and call $\hat{\mathbf{Y}}$ the translated information source, as mirrored by **B**.

Define some distortion measure comparing paths y to paths $\hat{y}$, $d(y, \hat{y})$. Invoke the Rate Distortion Theorem's mutual information $I(Y, \hat{Y})$, which is the splitting criterion between high and low probability pairs of paths. Impose, now, a parametization by an inverse coupling strength K, and a renormalization representing the global structure of the system coupling. This may be much different from the renormalization behavior of the individual components. If $K < K_C$, where K_C is a critical point (or surface), the two information sources will be closely coupled enough to be characterized as condensed.

In the absence of a distortion measure, the Joint Asymptotic Equipartition Theorem gives a similar result.

Detailed coupling mechanisms will be sharply constrained through regularities of grammar and syntax imposed by limit theorems associated with phase transition.

Biological Renormalization

Equation 56 states that the information source and the correlation length, the degree of coherence on the underlying network, scale under renormalization clustering in chunks of size R as

$$H[K_R, J_R]/f(R) = H[J, K]$$
$$\chi[K_R, J_R]R = \chi(K, J),$$

with $f(1) = 1$, $K_1 = K$, $J_1 = J$, where we have slightly rearranged terms.

Differentiating these two equations with respect to R, so that the right hand sides are zero, and solving for dK_R/dR and dJ_R/dR gives, after some consolidation, expressions of the form

$$dK_R/dR = u_1 d\log(f)/dR + u_2/R$$
$$dJ_R/dR = v_1 J_R d\log(f)/dR + \frac{v_2}{R} J_R. \tag{58}$$

The u_i, v_i, $i = 1, 2$ are functions of K_R, J_R, but not explicitly of R itself.

We expand these equations about the critical value $K_R = K_C$ and about $J_R = 0$, obtaining

$$dK_R/dR = (K_R - K_C)y d\log(f)/dR + (K_R - K_C)z/R$$
$$dJ_R/dR = w J_R d\log(f)/dR + x J_R/R. \tag{59}$$

The terms $y = du_1/dK_R \mid_{K_R=K_C}$, $z = du_2/dK_R \mid_{K_R=K_C}$, $w = v_1(K_C, 0)$, $x = v_2(K_C, 0)$ are constants.

Solving the first of these equations gives

$$K_R = K_C + (K - K_C)R^z f(R)^y, \tag{60}$$

again remembering that $K_1 = K$, $J_1 = J$, $f(1) = 1$.

Wilson's method is to iterate on this relation, which is supposed to converge rapidly near the critical point (Binney et al. 1986), assuming that for K_R near K_C, we have

$$K_C/2 \approx K_C + (K - K_C)R^z f(R)^y. \tag{61}$$

We iterate in two steps, first solving this for $f(R)$ in terms of known values, and then solving for R, finding a value R_C that we then substitute into the first of Eq. 56 to obtain an expression for $H[K, 0]$ in terms of known functions and parameter values.

The first step gives the general result

$$f(R_C) \approx \frac{[K_C/(K_C - K)]^{1/y}}{2^{1/y} R_C^{z/y}}. \tag{62}$$

Solving this for R_C and substituting into the first expression of Eq. 56 gives, as a first iteration of a far more general procedure (Shirkov and Kovalev 2001), the result

$$H[K, 0] \approx \frac{H[K_C/2, 0]}{f(R_C)} = \frac{H_0}{f(R_C)}$$
$$\chi(K, 0) \approx \chi(K_C/2, 0)R_C = \chi_0 R_C, \tag{63}$$

which are the essential relationships.

Note that a power law of the form $f(R) = R^m, m = 3$, which is the direct physical analog, may not be biologically reasonable, since it says that 'language richness' can grow very rapidly as a function of increased network size. Such rapid growth is simply not observed.

Taking the biologically realistic example of non-integral 'fractal' exponential growth,

$$f(R) = R^\delta, \tag{64}$$

where $\delta > 0$ is a real number which may be quite small, Eq. 62 can be solved for R_C, obtaining

$$R_C = \frac{\left[K_C/(K_C - K)\right]^{[1/(\delta y + z)]}}{2^{1/(\delta y + z)}} \tag{65}$$

for K near K_C. Note that, for a given value of y, one might characterize the relation $\alpha \equiv \delta y + z = \text{constant}$ as a 'tunable universality class relation' in the sense of Albert and Barabasi (2002).

Substituting this value for R_C back into Eq. 62 provides an expression for H, having three parameters: δ, y, z.

A more biologically interesting choice for $f(R)$ is a logarithmic curve that 'tops out', for example

$$f(R) = m \log(R) + 1. \tag{66}$$

Again $f(1) = 1$.

Using Mathematica 4.2 or above to solve Eq. 62 for R_C gives

$$R_C = \left[\frac{Q}{W[Q \exp(z/my)]}\right]^{y/z}, \tag{67}$$

where

$$Q \equiv (z/my)2^{-1/y} \left[K_C/(K_C - K)\right]^{1/y}.$$

Lambert's transcendental function $W(x)$ is defined by the relation

$$W(x) \exp(W(x)) = x.$$

It arises in the theory of random networks and in renormalization strategies for quantum field theories.

An asymptotic relation for $f(R)$ would be of particular biological interest, implying that 'language richness' increases to a limiting value with population growth. Such a pattern is broadly consistent with calculations of the degree of allelic heterozygosity as a function of population size under a balance between genetic drift and neutral mutation (Hartl and Clark 1997; Ridley 1996). Taking

$$f(R) = \exp[m(R - 1)/R] \tag{68}$$

gives a system which begins at 1 when $R = 1$, and approaches the asymptotic limit $\exp(m)$ as $R \to \infty$. Mathematica 4.2 finds

$$R_C = \frac{my/z}{W[A]}, \tag{69}$$

where

$$A \equiv (my/z) \exp(my/z) \big[2^{1/y} [K_C/(K_C - K)]^{-1/y} \big]^{y/z}.$$

These developments indicate the possibility of taking the theory significantly beyond arguments by abduction from simple physical models, although the notorious difficulty of implementing information theory existence arguments will undoubtedly persist.

Universality Class Distribution

Physical systems undergoing phase transition usually have relatively pure renormalization properties, with quite different systems clumped into the same 'universality class,' having fixed exponents at transition (Binney et al. 1986). Biological and social phenomena may be far more complicated:

If the system of interest is a mix of subgroups with different values of some significant renormalization parameter m in the expression for $f(R, m)$, according to a distribution $\rho(m)$, then the first expression in Eq. 56 should generalize, at least to first order, as

$$H[K_R, J_R] = E[f(R, m)] H[K, J]$$
$$\equiv H[K, J] \int f(R, m) \rho(m) dm, \tag{70}$$

so that $E[g]$ represents the mean of g over some distribution.

If $f(R) = 1 + m \log(R)$ then, given any distribution for m,

$$E[f(R)] = 1 + E[m] \log(R). \tag{71}$$

Other forms of $f(R)$ having more complicated dependencies on the distributed parameter or parameters, like the power law R^δ, do not produce such a simple result. Taking $\rho(\delta)$ as a normal distribution, for example, gives

$$E[R^\delta] = R^{E[\delta]} \exp \big[(1/2)(\log(R^\sigma))^2 \big], \tag{72}$$

where σ^2 is the distribution variance. The renormalization properties of this function can be determined from Eq. 62.

Thus the information dynamic phase transition properties of mixed systems will not in general be simply related to those of a single subcomponent, a matter of possible empirical importance: If sets of relevant parameters defining renormalization universality classes are indeed distributed, experiments observing pure phase changes may be very difficult. Tuning among different possible renormalization strategies in response to external signals would result in even greater ambiguity in recognizing and classifying information dynamic phase transitions.

Important aspects of mechanism may be reflected in the combination of renormalization properties and the details of their distribution across subsystems.

In summary, real biological, or prebiotic, systems are likely to have very rich patterns of phase transition which may not display the simplistic, indeed, literally elemental, purity familiar to physicists. Overall mechanisms will, however, still remain significantly constrained by the theory, in the general sense of probability limit theorems.

References

Adami C, Cerf N (2000) Physical complexity of symbolic sequences. Physica D 137:62–69
Adami C, Ofria C, Collier T (2000) Evolution of biological complexity. Proc Natl Acad Sci U S A 97:4463–4468
Albert R, Barabasi A (2002) Statistical mechanics of complex networks. Rev Mod Phys 74:47–97
Ash R (1990) Information theory. Dover, New York
Binney J, Dowrick N, Fisher A, Newman M (1986) The theory of critical phenomena. Clarendon, Oxford
Campos P, Fontanari J (1999) Finite-size scaling of the error threshold transition in finite populations. J Phys A 32:L1–L7
Champagnat N, Ferriere R, Meleard S (2006) Unifying evolutionary dynamics: From individual stochastic processes to macroscopic models. Theor Popul Biol 69:297–321
Cover T, Thomas J (1991) Elements of information theory. Wiley, New York
Dembo A, Zeitouni O (1998) Large deviations: techniques and applications, 2nd edn. Springer, New York
Dieckmann U, Law R (1996) The dynamical theory of coevolution: a derivation from stochastic ecological processes. J Math Biol 34:579–612
Eigen M (1971) Self-organization of matter and the evolution of biological macromolecules. Naturwissenschaft 58:465–523
Eigen M (1996) Steps towards life: a perspective on evolution. Oxford University Press, New York
Emery M (1989) Stochastic calculus in manifolds. Universitext series. Springer, New York
Feller W (1971) An introduction to probability theory and its applications. Wiley, New York
Feynman R (1996) Feynman lectures on computation. Addison-Wesley, Reading
Fontanari J, Santos M, Szathmary E (2006) Coexistence and error propagation in pre-biotic vesicle models: A group selection approach. J Theor Biol 239:247–256
Franzosi R, Pettini M (2004) Theorem on the origin of phase transitions. Phys Rev Lett 92:060601
Fredlin M, Wentzell A (1998) Random perturbations of dynamical systems. Springer, New York
Gunderson L (2000) Ecological resilience—in theory and application. Ann Rev Ecolog Syst 31:425–439
Hartl D, Clark A (1997) Principles of population genetics. Sinaur, Sunderland
Holling C (1973) Resilience and stability of ecological systems. Ann Rev Ecolog Syst 4:1–23
Holmes E (2003) Error thresholds and the constraints to RNA virus evolution. Trends Microbiol 11:543–546
Holmes E (2005) On being the right size. Nat Genet 37:923–924
Kastner M (2006) Phase transitions and configuration space topology. arXiv preprint cond-mat/0703401
Khinchin A (1957) Mathematical foundations of information theory. Dover, New York
Landau L, Lifshitz E (2007) Statistical physics, 3rd edn, part 1. Elsevier, New York
Leuthausser I (1986) An exact correspondence between Eigen's evolution model and a two-dimensional Ising system. J Chem Phys 84:1884–1885
Luchinsky D (1997) On the nature of large fluctuations in equilibrium systems: observations of an optimal force. J Phys A 30:L577–L583
Matsumoto Y (2002) An introduction to morse theory. In: Translations of Mathematical Monographs, vol 208. American Mathematical Society, Providence
McCauly L (1993) Chaos, dynamics and fractals: an algorithmic approach to deterministic chaos. Cambridge University Press, Cambridge
Milnor J (1963) Morse theory. In: Annals of Mathematical Studies, vol 51. Princeton University Press, Princeton
Ofria C, Adami C, Collier T (2003) Selective pressures on genomes in molecular evolution. J Theor Biol 222:477–483

Onsager L, Machlup S (1953) Fluctuations and irreversible processes. Phys Rev 91:1505–1512

Pettini M (2007) Geometry and topology in hamiltonian dynamics and statistical mechanics. Springer, New York

Ridley M (1996) Evolution, 2nd edn. Blackwell Science, Oxford University Press, Oxford

Shannon C, Weaver W (1949) The mathematical theory of communication. University of Illinois Press, Urbana

Shannon C (1957) Certain results in coding theory for noisy channels. Inf Control 1:6–30

Shirkov D, Kovalev V (2001) The Bogoliubov renormalization group and solution symmetry in mathematical physics. Phys Rep 352:219–249

Swetina J, Schuster P (1982) Self-replication with error – a model for polynucleotide replication. Biophys Chem 16:329–340

Szathmary E, Demeter L (1987) Group selection of early replicators and the origin of life. J Theor Biol 128:463–486

Szathmary E (1989) The integration of the earliest genetic information. Tree 4:200–204

Szathmary E (2006) The origin of replicators and reproducers. Trans R Soc B 361:1761–1776

Van Valen L (1973) A new evolutionary law. Evol Theory 1:1–30

Wallace R (2007) Culture and inattentional blindness: a global workspace perspective. J Theor Biol 245:378–390

Wallace R, Wallace RG (1999) Organisms, organizations, and interactions: an information theory approach to biocultural evolution. BioSystems 51:101–119

Wallace R, Wallace RG (1998) Information theory, scaling laws and the thermodynamics of evolution. J Theor Biol 192:545–559

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. Trans Comput Syst Biol IX, LNBI 5121:23–85

West-Eberhard M (2003) Developmental plasticity and evolution. Oxford University Press, New York

Wilson K (1971) Renormalization group and critical phenomena. I Renormalization group and the Kadanoff scaling picture. Phys Rev B 4:3174–3183

Wymer C (1997) Structural nonlinear continuous-time models in econometrics. Macroecon Dyn 1:518–548

Zhu R, Riberio A, Salahub D, Kauffman S (2007) Studying genetic regulatory networks at the molecular level: delayed reaction stochastic models. J Theor Biol 246:725–745

Zurek W (1985) Cosmological experiments in superfluid helium? Nature 317:505–508

Zurek W (1996) Shards of broken symmetry. Nature 382:296–298