Pacific Symposium on Biocomputing 5:599-610 (2000)

INFORMATION DYNAMICS OF *IN VITRO* SELECTION-AMPLIFICATION SYSTEMS

T.G. DEWEY
Keck Graduate Institute of Applied Life Sciences, 535 Watson Drive, Claremont CA 91711, USA

Selection-amplifications systems provide a means of engineering biomacromolecules with new properties. The combination of stringent functional selection with the ability to amplify single molecules confers great specificity on the evolving population. Yet such systems like many complicated chemical kinetic mechanisms can show a range of unstable and metastable behavior. These instabilites can be investigated using the Shannon entropy of the evolving population. It is shown that the Shannon entropy provides a Lyapounov function for exploring dynamic stability. A simple model of in vitro evolution is presented and stability conditions are established. It is seen that fairly simple directed evolution models can exhibit a range of dynamical behavior.

1 Introduction

The technology currently exists to mutate, screen and amplify nucleic acids. These methods when used in combination can yield a number of different *in vitro* techniques for optimizing a complicated biological process. In most applications, a population of RNA or DNA sequences is screened for a specific interaction or function. This selected population can then be amplified using isothermal RNA amplification (3SR) and/or a polymerase chain reaction (PCR). The resulting population can then be subjected to further selective pressures and the attribute of interest can be continuously optimized. Repeated cycling of this procedure refines the population to be highly specific. A representative scheme for such a protocol is shown in Figure 1. This general approach has been used to select for optimal sequences for protein-DNA interactions[1 2 3 4 5] for protein-RNA binding [6] and for catalytic specificity of ribozymes[7 8 9] and deoxyribozymes[10 11].

The general procedure illustrated in Figure 1 describes two distinctly different selection-amplification schemes. The first approach, described as "pure selection", creates a large initial library consisting of a population of many different sequences that are equidistributed and uncorrelated. The selection procedure is used to screen a subpopulation that is subsequently amplified. This is done in the protein-DNA and protein-RNA binding optimization. Alternatively, one can start with a single sequence and introduce a mutational mechanism that allows the system to "evolve" and search out new sequences. Such mutations can be random point mutations introduced by using modified

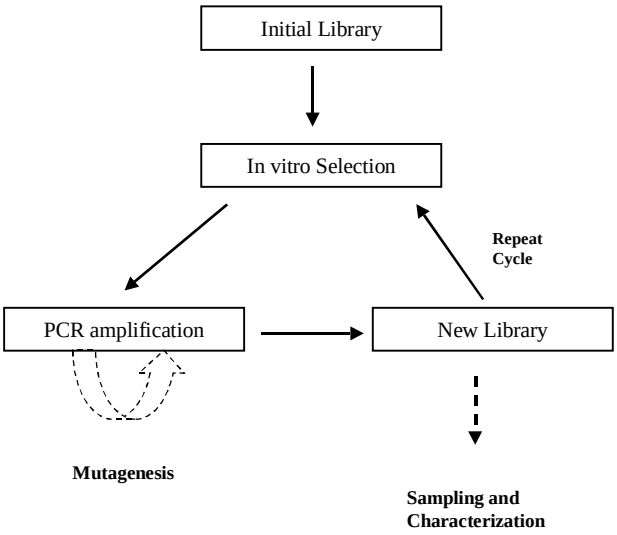

Figure 1: Schematic representation of the experimental design of selection-amplification of biological sequences (adapted from Sun et al., 1996).

nucleotides, block mutations incurred by artificial recombination[12 13] or can simply rely on intrinsic mutation rates in enzymatic steps such as the reverse transcriptase used in the RNA studies[7 10 14 15 16]. The "evolution" procedure differs from the "pure selection" procedure in that with evolution, the final selected sequence may not have existed in the initial library.

These selection-amplification experiments have been analyzed with specific mathematical models to explore the optimization of experimental design parameters.[17 18]. In these models, the screening process is treated in terms of the equilibrium thermodynamics of ligand-macromolecule binding. The amplification step is handled with a simple, probabilistic model of stepwise doubling of macromolecules. The combination of selection and amplification results in difference equations that describe the evolution of the population as the system is cycled. In the present work, the analysis of these selection-amplification systems is extended to explore the stability of the system. Of particular interest is the possibility of steady states away from the thermodynamic branch of the system To this end, the Shannon entropy of the cycling system is investigated. This proves to be a particularly useful quantity because it is a Lyapounov function of the difference equations governing the system's

dynamics. This means that the Shannon entropy can be used to establish stability criterion. A simple, heuristic treatment of the evolution of information in selection-amplification systems is described in Section 2. Section 3 establishes the connection between Shannon entropy and the Lyapounov function. The conditions required for stability are established in this section. Applications to *in vitro* evolution models is considered in Section 4. These examples build on a simple selection-amplification scheme without mutation that was described previously [18]. This model is extended to consider a two-species system that shows interchange as a result of mutations. Stability conditions for these models are discussed. Implications of these results for experimental protocols is discussed in the conclusion in Section 5.

2 Information Evolution in Selection-Amplification

If the number of each species or sequence in a population is known, the Shannon entropy for the system can be defined. Since the composition of a population changes for every selection-amplification cycle, the entropy may also changes. The information content of the sequences in the selected population will evolve in a very specific way that is dependent on the procedure used (see Figure 2). The Shannon entropy, I, is defined as [19]:

$$I = -\sum_{i=1}^{N} p_i \log_2 p_i \tag{1}$$

where N is the number of sequences in the library and p_i is the probability of finding the ith sequence. In a randomly generated population, each sequence is generated with equal probability and $p_i = 1/N$. In this case, Eq. 1 becomes:

$$I_{\text{random}} = \log_2 N \tag{2}$$

This situation, equal probabilities for all sequences, maximizes Eq. 1 for a given N. Thus, any process, such as selection, that favors one set of sequences over another will reduce the Shannon entropy of the population. It should also be noted that the Shannon entropy is a property of the library. Because of this library-dependent property, an identical sequence occurring in two different libraries may make a different contribution to the Shannon entropy.

From general arguments, the qualitative behavior of the evolution of information can be seen for the pure selection method. Because this design starts with a randomly generated library, its information is at a maximum. Figure 2 shows a Venn diagram that represents the set of all possible sequences. As the system cycles, selective pressures restrict the sequences to a narrower, more

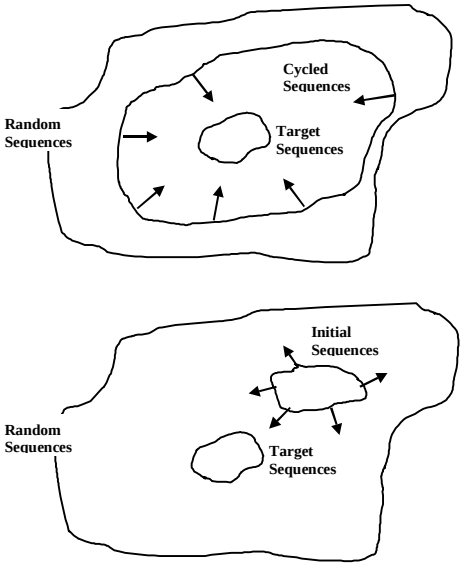

Figure 2: Evolution of information content as selection-amplification proceeds.

specific set (see Figure 2 top). When members in this set are not differentially amplified Eq. 2 holds. The Shannon entropy of the set drops with cycling merely because selection reduces the number of individual sequences, N. When differential amplification occurs, the information will be further reduced because of redundancy in the population, i.e., because not all probabilities are identical in Eq. 1. As the selection cycles proceed, the system will converge to the set of sequences that have the Shannon entropy of the "target" sequences. This is an example of "inward" evolution where the Shannon entropy of the library decreases as a result of the specific sequences required to optimize a given biological function. An analysis of the evolution of the information content of protein sequences show that biological molecular evolution is inward [20].

For the experimental design used in the evolution of Tetrahymena ribozyme, one starts with a "wild-type" sequence and introduces point mutations by adding modified nucleotides to the PCR. This produces a "Generation 0" in which all possible one-error mutants are present in high frequency along with increasingly smaller proportions of higher-error mutants. Because this library has a high frequency distribution around a single sequence, it will have a much lower total information than a randomly generated library. Selective pressures

then drive the sequences to evolve toward a specific target (see bottom of Figure 2). Unlike the pure selection process, when mutation rates are enhanced, this system can evolve and create new sequences. This evolutionary process will have more complicated information dynamics than the pure selection process. Initially, the mutagenesis will primarily result in a filling out of the wild type library. However, as cycling proceeds the library could expand to seek out the new set of target sequences. This process is also illustrated in Figure 2. Thus, there may be an "outward" evolution in the early stages of cycling. However, there is no reason that a successful library of target sequences should contain more information than Generation 0. Consequently, after the target region is discovered by the combination of evolution and selection, there may well be a decrease in Shannon entropy, just as in the pure selection process. An analysis of the information content of an evolving ribozyme system suggests that the evolution is "outward" [21].

3 Shannon Entropy and System Stability

The change in the Shannon entropy provides a means of assessing the stability of the system. As will be seen, system cycling will result in a progression in the population that can be described using discrete difference equations. Ideally, one would hope that with repeated cycling, the system achieves a constant population that is optimal for a specific biological function. However, there is no assurance that repeated cycling will accomplish this goal. The solutions to the difference equations governing selection-amplification can be stable, marginally stable or unstable, depending on the initial conditions and the equation parameters. The *direct method* of Lyapounov provides a powerful means of assessing stability of non-linear differential [22] and difference equations [23]. To perform this analysis, a Lyapounov function must be identified and the weakness of the method is that there is no *a priori* method for identifying such a function. In the present case, we show that the Shannon entropy is a valid Lyapounov function for selection-amplification systems. It can, therefore, be used to investigate the stability of steady states of the evolving system.

To put these ideas into a concrete framework, the selection-amplification model of Sun et al. [18] is considered. They treated the case of a population of many DNA molecules of different sequences binding to a single protein species. The different sequences were separated into groups according to the binding affinity. The "initial library" contained N different groups. Each group contains n_i different molecules and each of these molecules has an association constant for binding to the protein of K_i. The subscript designating the group runs from $i = 1$ to N. Only the DNA molecules that are bound to protein

will be selected in the screening process and these are the molecules that get amplified. The fraction of selected molecules is given by:

$$f_i = \frac{K_i \left[protein\right]}{1 + K_i \left[protein\right]} \tag{3}$$

where $[protein]$ is the equilibrium concentration of free protein. The frequencies, $f_1, f_2, \ldots, f_N$ represents the probability distribution of molecules that are selected to be amplified. These frequencies will be used to estimate the Shannon entropy.

Amplification occurs as a result of PCR and the reaction cycle is considered to have an efficiency of λ. A screened molecule will be doubled with probability, λ, and will remain a single copy with probability, $1 - \lambda$. The amplification is assumed to be free of mutation and there will be l cycles of PCR amplification and m experimental cycles (combined selection-amplification cycles). It is assumed that the l PCR cycles are held constant through each experimental cycle. After m cycles, the number of molecules in the ith group is given by the following recursion:

$$n_{i,m} = f_i \left(1 + \lambda\right)^l n_{i,m-1} = f_i^m \left(1 + \lambda\right)^{lm} n_{i,0} \tag{4}$$

where $n_{i,0}$ is the number of molecules in the ith group of the initial, unscreened population. For the present purposes, the probability of the ith group is required. This is given by:

$$p_{i,m} = \frac{n_{i,m}}{\sum_{j=1}^{N} n_{j,m}} \tag{5}$$

The probability of a molecule belonging to the ith group gives a more complicated, non-linear recursion:

$$p_{i,m} = \frac{f_i p_{i,m-1}}{\sum_{j=1}^{N} f_j p_{j,m-1}} = \frac{f_i^m p_{i,0}}{\sum_{j=1}^{N} f_j^m p_{j,0}} \tag{6}$$

Interestingly, the PCR amplification term drops out of the expression for the probability. This is because each selected species is equally amplified.

The resulting recursion relationship, Eq. 6, is a difference equation involving the cycle number, m, as a variable. The experimentalist would hope to reach a situation were further cycling does not effect the population, i.e., a stable population exists. To mathematically determine if stability is possible, a Lyapounov function, $H\left(\mathbf{p}_m\right)$ is sought. This function must merely satisfy

the condition that $H \geq 0$ and that the partial derivatives with respect to $p_{i,m}$ are continuous. Lyapounov theorem states that stability conditions are:

$$\begin{array}{ll} -\sum_{j=1}^{N} \dot{p}_{j,m} \left(\frac{\partial H}{\partial p_{j,m}} \right) > 0 & \text{stable} \\ -\sum_{j=1}^{N} \dot{p}_{j,m} \left(\frac{\partial H}{\partial p_{j,m}} \right) = 0 & \text{marginally stable} \\ -\sum_{j=1}^{N} \dot{p}_{j,m} \left(\frac{\partial H}{\partial p_{j,m}} \right) < 0 & \text{unstable} \end{array} \tag{7}$$

where $\dot{p}_{j,m}$ is the discrete equivalent of the derivative with respect to m and is given by $p_{j,m} - p_{j,m-1}$.

The Shannon entropy for the population of sequences obeying Eq. 6 is a Lyapounov function. It is always positive and its partial differential with respect to the probabilities is continuous. Thus, the Shannon entropy can provide a useful tool in examining the dynamics of the selection-amplification system. The Lyapounov function is then defined as:

$$H(\mathbf{p}_m) = I_m = -\sum_{j=1}^{N} p_{j,m} \log_2 p_{j,m} \tag{8}$$

The Shannon entropy also has the attractive quality of being an "extensive" function. That is, the information content of two independent systems is additive. The stability condition using the Shannon entropy as a Lyapounov function takes a relatively simple form:

$$\begin{array}{ll} \sum_{j=1}^{N} \dot{p}_{j,m} \log_2 p_{j,m} > 0 & \text{stable} \\ \sum_{j=1}^{N} \dot{p}_{j,m} \log_2 p_{j,m} = 0 & \text{marginally stable} \\ \sum_{j=1}^{N} \dot{p}_{j,m} \log_2 p_{j,m} < 0 & \text{unstable} \end{array} \tag{9}$$

These stability conditions are analogous in form to those of non-equilibrium thermodynamics [24] and are a consequence of defining the Lyapounov function as an "entropy-like" function. Using the analogy with non-equilibrium thermodynamics, $\dot{p}_{j,m}$ is identified with a component of the thermodynamic flux and $\log_2 p_{j,m}$ is the respective affinity. As in the thermodynamic case, it is possible to explore stability of steady states far from equilibrium using a fluctuation analysis [24]. In such an analysis, the variation in the Lyapounov function is considered using an expansion:

$$I_m = I_{m,ss} + \delta I_m + (1/2)\, \delta^2 I_m \tag{10}$$

where $I_{m,ss}$ is the Shannon entropy at steady state values of $p_{i,m}$ and δI_m and $\delta^2 I_m$ are first and second order variations about the steady state. When the

steady state is an extremum, $\delta I_m = 0$ and $\delta^2 I_m$ is the Lyapounov function used to establish the stability of the steady state. In the present case, a new Lyapounov function is defined by:

$$\delta^2 I_m = \sum_{j=1}^{N} \left(\frac{\partial^2 I_m}{\partial p_{j,m}^2} \right) \delta p_{j,m}^2 \tag{11}$$

To establish the extremum, recursion relationships such as Eq. 6 are used. These give equations of the form: $p_{i,m} = F\left(p_{1,m-1}, p_{2,m-1}, \ldots p_{N,m-1}\right)$. Steady state is achieved when $p_{i,m} = p_{i,m-1}$, so the steady state condition is established by solving the N equations:

$$p_{i,ss} = F\left(p_{1,ss}, p_{2,ss}, \ldots p_{N,ss}\right) \tag{12}$$

For stability about the steady state, the following two conditions must be fulfilled[24]:

$$\delta^2 I_m = -\sum_{j=1}^{N} \frac{\delta p_{j,m}^2}{p_{j,m}} < 0 \tag{13}$$

$$\frac{d}{dm}\left(\delta^2 I_m\right) = -2\sum_{j=1}^{N} \frac{\delta p_{j,m} \delta \dot{p}_{j,m}}{p_{j,m}} \geq 0 \tag{14}$$

where the continuum limit is used in Eq. 14. The first condition, Eq. 13 will always be satisfied because $p_{j,m} \geq 0$. Consequently, the second condition, Eq. 14, represents the main computational tool for assessing stability. This will be used in the next section to investigated a selection-amplification system that can evolve as a result of mutations.

4 Stability Analysis of Directed Evolution Models

The selection-amplification scheme described above for optimizing DNA binding to a specific sequence can be analyzed using the Lyapounov approach. This is a particularly simple system that basically converges to a single steady state. As the number of cycles increases, the dominant species will be the one with the greatest affinity (largest f_i) for the protein. As $m \to \infty$, one of the probabilities approaches unity and all other approach zero. Thus, this simple selection-amplification scheme achieves the goal of finding the optimal population with respect to DNA binding.

Slightly more complicated dynamics will occur when the system is allowed to evolve as a result of mutations. A DNA-protein selection model is again

considered for our evolving system. For simplicity, a two-component system is considered. One component has a low-affinity for the protein and the other has a high-affinity. This is not such an unrealistic model, if the binding is extremely sensitive to sequence and is close to being "all or nothing". During the amplification steps, the system is allowed to mutate. The probability that no mutation occurs during amplification is γ. If a mutation occurs, it converts one affinity state into the other. This occurs during amplification with probability, $1-\gamma$. In this model, the mutation rates between species are both identical. The probability of a sequence doubling is: $\lambda\gamma$. The probability of the sequence being amplified into the other affinity state is: $\lambda\,(1-\gamma)$ In this case, the number of sequences in the initial state remains the same. The probability of a sequence not being amplified is, as before, $(1-\lambda)$.

The recursion relationships for this case are now much more complicated. After significant algebra, the following recursion relationship is obtained;

$$p_{1,m} = \frac{[A+B]\, f_1 p_{1,m-1} + [A-B]\, f_2 p_{2,m-1}}{2A\,(f_1 p_{1,m-1} + f_2 p_{2,m-1})} \tag{15}$$

with

$$A = (1/2)\,(1+\lambda)^l \tag{16}$$

and

$$B = (1/2)\,(1-\lambda+2\lambda\gamma)^l \tag{17}$$

A similar recursion holds for $p_{2,m}$. However, it is not needed because one can use: $p_{1,m} = 1 - p_{2,m}$.

For a two state system, the stability condition, Eq. 14 takes a particularly simple form. It amounts to establishing that $\delta\dot{p}_{1,m} \leq 0$. This is determined by first taking the variation of Eq. 15 and treating the derivative with respect to m as:

$$\delta\dot{p}_{1,m} = \delta p_{1,m} - \delta p_{1,m-1} \tag{18}$$

This yields expressions for $\delta^2 I_m$ that equal $(\delta p_{1,m})^2$ times a function in $p_{1,m}$. The steady state values of $p_{1,m}$ are than substituted into this function to evaluate the inequality.

Using Eqs. 12 and 14, the stability conditions for this evolving system can be established. For this case, Eq, 12 gives a quadratic equation. For physically plausible situations, only one of these roots gives an acceptable steady state probability, i.e., it must lie between 0 and 1. This root is asymptotically stable. As a sample calculation, we consider the case were $B = (1/2)\,A$ and $f_2/f_1 = 0.1$. At steady state, the high affinity species has $p_{1,m} = 0.73$. Interestingly, introducing mutation into the system means that the system will

not converge on a single species, i.e. $p_{1,m} \neq 1$. There can be a significant difference in binding affinities of the two populations and yet the lower affinity component never dies out. This shows that it is not possible for the system to maintain an optimal population with respect to binding affinity. Another interesting condition is $B = (1/2)\,A$ and $f_2/f_1 = 1$. This case has no selective advantage of one species over the other and a single root giving a 50:50 mix is obtained. However, this solution is only marginally stable and can support oscillations.

Introducing more complicated features into the model admits more complicated dynamics. For instance, if the mutation rates differ with the high affinity (favored) species having a high mutation rate while the lower affinity species does not favor mutation, two physically acceptable roots can be obtained and the possibility of oscillatory behavior exists. Similarly if the selection step involves multiple binding of DNA to protein then additional non-linearity is introduced into the system. This creates situations that have the potential for chaotic and self-organizing behavior. These more complicated systems remain to be analyzed in more detail.

5 Conclusion

This work demonstrates the utility of using the Shannon entropy to investigate the dynamics of *in vitro* selection-amplification systems. It is shown that for a range of simple selection-amplification systems, the Shannon entropy is a Lyapounov function for the difference equations governing the temporal evolution. As such, it can be used to establish conditions for dynamic stability. Using this approach, a variety of model systems can be investigated and the requirements for stability can be established. This provides the experimentalist with the necessary tools to model and design such system. Perhaps the most important consequences of this analysis is that even simple models show that the selective pressure alone does not drive the system to a single, optimal chemical species. There will always be design constraints in mutational systems that can prevent optimization of the system. It is also possible to create simple experimental situations where oscillatory and self-organized temporal behavior exist. Such situations can allow for the creation of robust experimental systems that evolve to an attractor. This analysis can aid the experimentalist in finding the conditions for such behavior.

Acknowledgments

This is work was supported in part by NIH grant # 1R15GM55910.

References

1. K.W. Kinzler and B. Vogelstein, Whole genome PCR: application to the identification of sequences bound by gene regulatory proteins. *Nucl. Acids Res.* **17**, 3645-3653 (1989).
2. T.K. Blackwell and H. Weintraub, Differences and Similarities in DNA-Binding Preferences of MyoD and E2A Protein Complexes Revealed by Binding Site Selection. *Science* **250**, 1104-1110 (1990).
3. H.J. Thiesen and C. Bach, Target detection assay (TDA): a versatile procedure to determine DNA binding sites as demonstrated on SP1 protein. *Nucl. Acids Res.* **18**, 3202-3209 (1990).
4. R. Pollock and R. Treisman, A sensitive method for the determination of protein DNA binding specificities. *Nucl. Acids Res.* **18**, 6197-6204 (1990).
5. E.J. Rebar and C.O. Pabo, Zinc finger phage: affinity selection of fingers with new DNA-binding specificities. *Science* **263**, 671-673 (1994).
6. C. Tuerk and L. Gold, Systematic Evolution of Ligands by Exponential Enrichment: RNA Ligands to Bacteriophage T4 DNA Polymerase. *Science* **249**, 505-510 (1990).
7. D.P. Beaudry and G.F. Joyce, Directed evolution of an RNA enzyme. *Science* **257**, 635-641 (1992).
8. D.P. Bartel and J.W. Szostak, Isolation of new ribozymes from a large pool of random sequences.*Science* **261**, 1411-1418 (1993).
9. J.M. Burke and A. Berzal-Herranz, In vitro selection and evolution of RNA: Applications for catalytic RNA, molecular recognition and drug discovery. *FASEB J.* **7**, 106-112 (1993).
10. R.R. Breaker and G.F. Joyce, A DNA enzyme that cleaves RNA. *Chem. Biol.* **1**, 223-239 (1994).
11. C.R. Geyer and D. Sen, Evidence for the metal-cofactor independence of an RNA phosphodiesterase-cleaving DNA enzyme. *Chem. Biol.* **1**, 579-593.
12. W.P.C. Stemmer, DNA shuffling by random fragmentation and reassembly: In vitro recombination for molecular evolution. *Proc. Natl. Acad. Sci. USA* **91**, 10747-10751.
13. D.H. Burke and J.H. Willis, Recombination, RNA evolution and bifunctional RNA molecules isolated through chimeric SELEX. *RNA* **4**, 1165-1175 (1998).
14. N. Lehman and G.F. Joyce, Evolution in vitro of an RNA enzyme with altered metal dependence. *Nature* **361**, 182-185 (1993).
15. N. Lehman and G.F. Joyce. Evolution in vitro: analysis of a lineage of

ribozymes. *Curr. Biol.* **3**, 723-734 (1993).
16. Tsang J, Joyce GF (1996) In vitro evolution of randomized ribozymes. Meth. Enzymol. 267:410-426 (1996).
17. D. Irvine,C. Tuerk and L. Gold, SELEXION Systematic Evolution of Ligands by Exponential Enrichment with Integrated Optimization by Non-Linear Analysis. *J. Mol. Biol.* **222**, 739-761 (1991).
18. F. Sun, D. Galas and M.S. Waterman, A Mathematical Analysis of in Vitro Molecular Selection-Amplification. *J. Mol. Biol.* **258**, 650-660 (1996).
19. T.M. Cover and J.A. Thomas in *Elements of Information Theory* (John Wiley & Sons, New York, 1991).
20. T.G. Dewey and M. Delle Donne, Non-equilibrium Thermodynamics of Molecular Evolution. *J. Theor. Biol.* **193**, 593-599(1998).
21. N. Lehman, M. Delle Donne, M. West and T.G. Dewey, The Genotypic Landscape during *In Vitro* Evolution of a Catalytic RNA: Implications for Phenotypic Buffering (submitted, 1999).
22. P.G. Drazin in *Nonlinear Systems* (Cambridge University Press, Cambridge 1992) pp. 178-180.
23. N.G. van Kampen in *Stochastic Processes in Physics and Chemistry* (Elvsevier Publishing Co. Amsterdam, 1981) pp. 110-121.
24. I. Prigogine in *From Being to Becoming* (W.H. Freeman and Co., San Francisco, 1980) pp. 77-101.
25. B.J. Strait and T.G. Dewey, The Shannon Information Entropy of Protein Sequences. *Biophys. J.* **71**, 148-155 (1996).