THE POSSIBLE ROLE OF ASSIGNMENT CATALYSTS IN THE ORIGIN OF THE GENETIC CODE

VAHE BEDIAN*

Department of Biophysical Sciences, State University of New York at Buffalo, Buffalo, New York, 14214, U.S.A.

(Received 29 January, 1982; Accepted 5 April, 1982)

Abstract. A model is presented for the emergence of a primitive genetic code through the selection of a family of proteins capable of executing the code and catalyzing their own formation from polynucleotide templates. These proteins are assignment catalysts capable of modulating the rate of incorporation of different amino acids at the position of different codons. The starting point of the model is a polynucleotide based polypeptide construction process which maintains colinearity between template and product, but may not maintain a coded relationship between amino acids and codons. Among the primitive proteins made are assumed to be assignment catalysts characterized by structural and functional parameters which are used to formulate the production kinetics of these catalysts from available templates. Application of the model to the simple case of two letter codon and amino acid alphabets has been analyzed in detail. As the structural, functional, and kinetic parameters are varied, the dynamics undergoes many bifurcations, allowing an initially ambiguous system of catalysts to evolve to a coded, self-reproductive system. The proposed selective pressure of this evolution is the efficiency of utilization of monomers and energy. The model also simulates the qualitative features of suppression, in which a deleterious mutation is partly corrected by the introduction of translational error.

1. Introduction

Of the many complex problems associated with the origin of life, the origin of the genetic code is one of the most puzzling. The apparent universality of the code can be viewed as a consequence of a common ancestry of terrestrial life, while the remarkable accuracy of individual code assignments (i.e., codon-amino acid associations) can be ascribed to the specificity of aminoacyl-tRNA synthetases and other components of the translation system, resulting from the three-dimensional conformation of these macromolecules. In contrast, it is difficult to find a physical basis for the fact that the action of *all* aminoacyl-tRNA synthetases defines a *coded* relationship between the codon and amino acid alphabets. This relationship is the genetic code, which has the striling and important property of being unambiguous: only one amino acid is assigned to each codon, although several different codons may be related to the same amino acid (degeneracy).

The earliest attempts to understand the origin of this coded relationship were mechanistic theories, postulating stereochemical fits between amino acids and their codons or anticodons (Pelc and Welton, 1966). Although some specificity of interaction has been observed (Wagner and Arav, 1968; Saxinger and Ponnamperuma, 1971) the experimental evidence available to date does not substantiate a purely stereochemical theory of origin,

* Present Address: Department of Biochemistry and Biophysics, University of Pennsylvania School of Medicine, Philadelphia, Pennsylvania 19104, U.S.A.

Origins of Life 12 (1982) 181–204. 0302–1688/82/0122–181 $ 03.60.

according to which a unique and universal code would be prescribed by interaction energies of amino acids and nucleotides. A purely stochastic theory of origin of the present-day code, postulating equiprobability of all codon – amino acid assignments and therefore all possible codes, may be discounted because of very low probabilities for the nucleation of a code for 20 amino acids through a 'frozen accident' (Crick, 1968; Eigen, 1971; Hoffman, 1975). This problem may be circumvented by assuming that code nucleation is required only for a few amino acids (or types of amino acids), further additions (or distinctions) being introduced through evolution. While these two theories are useful as extreme limiting cases, it is more likely that both stereochemical and stochastic effects, as well as special initial and boundary conditions, played important roles during the complex stages of molecular evolution leading to the nucleation of a primitive code.

An alternative approach is to emphasize the contribution of primitive proteins capable of modulating rates of amino acid-codon associations (assignment catalysts), and assume that the selective advantages of prebiotic systems derived from their kinetic features. This paper is an attempt to model the enzymatic hypothesis of code origin, which was suggested through the error catastrophe models of Orgel (1963, 1970), Hoffman (1974), Kirkwood and Holliday (1975), and Goel and Ycas (1975, 1976). The hypothesis focuses on the role of assignment catalysts, which correspond best to present day aminoacyl-tRNA synthetases, elongation factors, and ribosomal proteins, in establishing and maintaining coded translation as an essential process of living organisms. These models of error propagation have already shown, that within the context of a single code, there is a threshold of accuracy above which the system can achieve stable and correct translation, but unfavorable initial conditions can drive the system to a stable but error laden state.

Mizutani and Ponnamperuma (1977) have already presented an enzymatic model for the evolution of a translation machinery in which polymerases are characterized by activities and selectivities (expressed as the grade of a polymerase) and a number of critical sites determining: these properties. Each polymerase is assumed to belong to a genetic code, and transition probabilities between different grade systems are defined. Their simulation results indicate that for reasonable values of selectivities and transition probabilities, the highest grade system will eventually dominate. However, the model treats transition probabilities between different grade components as arbitrary parameters, and does not reflect their dependence on the existing population of polymerases. Furthermore, it seems that transitions between systems that belong to the same or alternate codes need to be distinguished.

Our purpose is to construct a kinetic description of a system of assignment catalysts, such that assignment probabilities are expressed in terms of concentrations and activities of existing catalysts, and all possible codes (for a given choice of amino acid and codon alphabets) are explicitly accounted for. In accordance with the many analyses of the patterns of degeneracy in the code (Crick, 1968; Jukes, 1973, 1974; Walker, 1974; Wong, 1975, 1976), we will assume that the primitive code recognized (or distinguished) only a few (2–6) amino acids and codons.

2. The Context of the Origin of the Genetic Code

The sophistication of the translation machinery is such, that even the simplest possible system would require many complex processes that could only be established by a long history of molecular evolution. Figure 1 illustrates the general picture that has evolved during the last decade concerning processes that could lead to a protocell capable of reproduction and evolution. Below is a list of structural and functional components that could have evolved in a stepwise fashion to a protocell capable of *ambiguous* translation. The scheme combines features proposed by Fox (1974a) and Eigen and Schuster (1978b):

(a) Formation of amino acids and energy rich derivatives, nucleotides, and copolymers of each class. Although the exact details of prebiotic chemistry may remain unknown, a variety of plausible pathways and mechanisms for such processes have been demonstrated.

(b) Formation of protocells, possibly similar to coacervate droplets (Oparin, 1964) or proteinoid microspheres (Fox, 1974a,b), capable of propagation by uptake of polymers. Such a protocell would acquire obvious advantages by evolving the capacity to catalyze polymer formation (Fox *et al.*, 1974) and selective transport of monomers (Oparin, 1966).

(c) Polynucleotide replication by base pairing of complementary chains leading to stabilization of some families of polynucleotides. This process is analyzed in detail by Eigen (1971) and Eigen and Schuster (1977, 1978a).

(d) Coupling of polynucleotide and polypeptide formation processes. Even without the requirement of *coded* translation this is a complex process that requires many components and interactions. Since thermal proteinoids have been shown to exhibit a variety of catalytic activities (Fox and Dose, 1977), it is reasonable to assume that some may have acted as nucleotide polymerases. On the other hand, the first involvement of polynucleotides in catalysis of polypeptide formation may have been through primitive tRNAs. Crick *et al.* (1976) have proposed a scheme in which aminoacyl and peptidyl tRNAs bind the template with five base pairs, but in two different conformations. An experimental model system capable of polynucleotide based polypeptide production while maintaining colinearity is given by the work of Weber and Lacey (1974) and Lacey *et al.* (1975). They used aminoacylnucleoside anhydrides, and imidazole for catalyzing amino acid transfer to the bases of polynucleotides, to obtain some short peptides. Whatever its origin, the polynucleotide directed condensation of amino acids would have to ensure colinearity of the two polymer classes before a coded relationship between amino acids and codons could evolve.

Given a system with the above features, as well as the possibility that some polypeptides act as assignment catalysts, the thesis presented here is that an initially ambiguous population of assignment catalysts would evolve towards increased specificity and coded translation with concurrent evolution of their template polynucleotides and those of other functional components. This proposed evolution towards coded translation is based on increasing selective advantages by more efficient utilization of available monomers and energy.

3. A Primitive Protein Synthesizing Machinery

For purposes of focusing on the role of assigment catalysts, a stable population of self-replicating polynucleotides is assumed in the model without explicit representation of replication dynamics. As a further simplification, only one polynucleotide template for each functional protein is assumed. The polynucleotide based production of polypeptides maintains colinearity between template and product, but does not necessarily maintain a coded relationship between codons and amino acids; i.e., more than one type of amino acid may be incorporated at the position of a codon. This is a consequence of the assumption that catalysts for all possible assignments are present in the system, and a single catalyst may exhibit activity for all assignments. Thus, a variety of equal length polypeptides may be produced along the same nucleic acid.

In addition to the polynucleotide directed polymerization, two other processes are assumed to contribute to polypeptide dynamics: a linear decay term representing non-specific hydrolysis, and a constant production term representing contribution from random processes. Under these assumptions, the kinetic equation for a template derived protein (including assignment catalysts) is of the form:

$$\dot{x} = -Dx + C + KQ, \tag{1}$$

where

- x is the concentration of the protein
- $\dot{x}$ is its time derivative
- D is the decay rate constant
- C is the constant production rate due to random processes
- K is the overall production rate of polypeptides along the template polynucleotide in question
- Q is the fraction of polypeptides produced along the template that satisfy the structural requirements for the desired functional protein.

The factor Q is a function of assignment catalyst concentrations and specificities, as well as the structural relationship between template and desired product. Section 4 details the functional and structural characterization of assignment catalysts, and an explicit expression for Q is derived in Section 6. A more detailed and general derivation is given by Bedian (1979).

4. Functional and Structural Characterization of Catalysts

A striking feature of the present-day translation process is that essential protein components that effect code assignments (assignment catalysts) are themselves coded for in the genome, and are derived from their respective messenger RNAs by their own concerted and coherent action. The origin of such a self-reproductive set of assignment catalysts is the subject of this paper.

In a system where N_a amino acids and N_c codons are distinguished, the number of possible assignments (codon-amino acid associations) is given by $N_s = N_c \times N_a$. There will

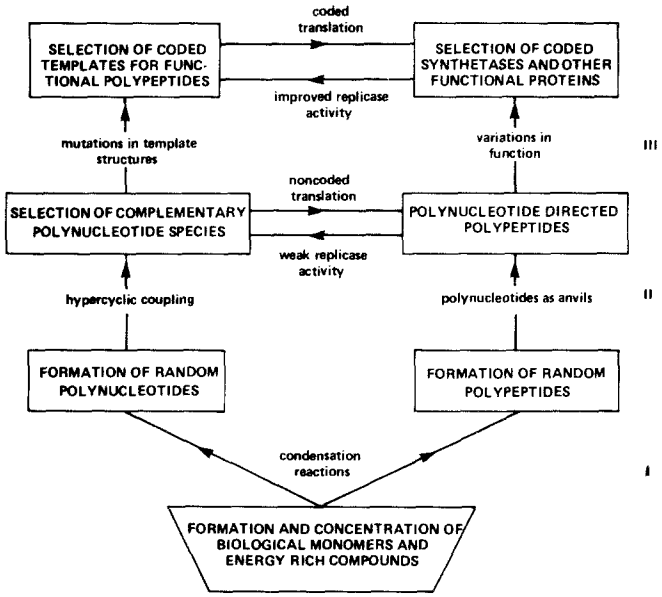

Fig. 1. General context of the model, indicating the hierarchy of events and processes assumed. The model pertains to stage III, but ignores replicase activities and their possible improvement.

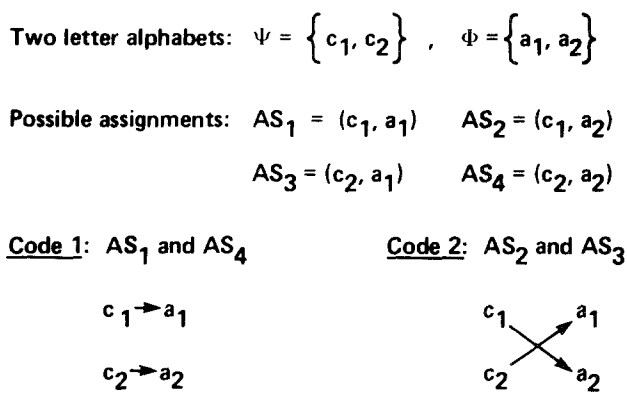

Fig. 2. Two letter codon and amino acid alphabets define four assignments (ordered pairs of codons and amino acids) and two alternative codes. Each code is a pair of assignments defining a single-valued mapping of codons into amino acids (i.e., no two arrows originate from the same codon).

be more than one subset of these N_s assignments that can define a code, as illustrated in Figure 2 for 2-letter codon and amino acid alphabets. Each assignment catalyst can be functionally characterized by N_s activity values for the N_s assignments, and a system of catalysts can be described by an activity matrix S. An element s_{ij} of S is defined as the contribution to the rate of assignment j due to the action of catalyst i, per unit concentration of the catalyst. The element s_{ij} is thus referred to as the activity of catalyst i for assignment j.

The *selectivity* of a catalyst is then represented by the relative values of activities for different assignments. A highly selective catalyst would have only one non-zero component in its activity vector. On the other hand, a non-specific catalyst will have equal activities for all assignments.

Critical sites in primary structure

$...a_2\ a_1\ a_1\ a_2\ a_1 \cdots a_1\ a_2\ a_1 \cdots a_1\ a_2\ a_2\ a_1 \cdots$

$2\uparrow\ 1\uparrow\ 3\uparrow\ 2\uparrow\ 3\uparrow \quad 3\uparrow\ 4\uparrow\ 1\uparrow \quad 1\uparrow\ 2\uparrow\ 4\uparrow\ 3\uparrow$ **Required assignments**

$...c_1\ c_1\ c_2\ c_1\ c_2 \cdots c_2\ c_2\ c_1 \cdots c_1\ c_1\ c_2\ c_2 \cdots$

Corresponding codons in template

Assignment	No. of times required	m parameter
1	3	$m_{i1} = 3$
2	3	$m_{i2} = 3$
3	4	$m_{i3} = 4$
4	2	$m_{i4} = 2$
	Total number of critical sites:	$m_i = 12$

Fig. 3. Illustration of the template-product structural relationship. The sequence of amino acids required at the critical sites of the component are shown along with the corresponding codons in the template polynucleotide and the assignments required for obtaing the desired polypeptide from the given template. The table indicates the number of times each assignment is required for the production of component i. Note that all four assignments are required, although code 2 has a larger share ($m_{i2} + m_{i3} = 7$ compared to $m_{i1} + m_{i4} = 5$). In present day cells the production of a protein from its mRNA requires only assignments that belong to a single code.

Assignment catalysts (or any polypeptide with partially or totally specified primary structure) can also be structurally characterized by assuming that the occurrance of particular amino acids at a number of critical sites results in a protein with the desired function. The relationship of these critical regions of primary structure to the available 'messenger' structure can be summarized as follows: By comparing the amino acids at the critical sites with corresponding codons in the template polynucleotide, all the assignments required for the production of the protein from the template can be deduced as illustrated in Figure 3. Assuming independence between different polymerization steps, only the *total number* of occurrences of each assignment will be of significance in the model. The system of assignment catalysts can thus be structurally characterized by a second matrix, called the structural matrix M. An element m_{ij} of M is the total number of times assignment j is required for the production of catalyst i from its template. The total number of critical sites of catalyst i is then given by:

$$m_i = \sum_{j=1, N_s} m_{ij} \tag{2}$$

5. Assignments, Codes and Self-Coding Systems

It is important to clarify the implications of the definition of a code concerning the form of the S and M matrices. To achieve its unambiguous character, a code has to be a special subset of assignments that defines a single-valued function between codon and amino acid alphabets (Bedian and Herman, 1974). This is illustrated in Figure 2 for the simple case of two-letter alphabets in which two alternative codes are possible.

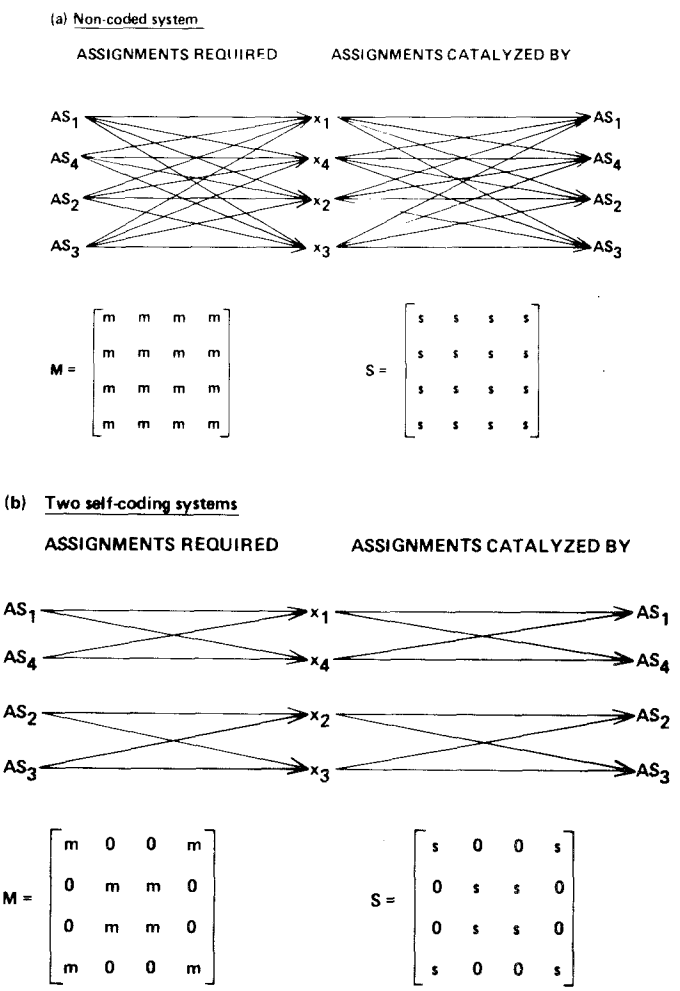

Fig. 4. Illustration of non-coding and self-coding systems for two letter alphabets and four catalysts: x_1, x_2, x_3, x_4. Arrows on the left represent assignments required for production of the four components (m parameters), while arrows on the right represent assignments catalyzed by the four components (s parameters). (a) represents the most stochastic, non-coded system, with all elements of M, as well as all elements of S, being equal. (b) represents the most coded system, with both pairs of catalysts (x_1, x_4) and (x_2, x_3) satisfying the conditions of self-coding. The restrictions on M can be viewed as selection of template structures, while the restrictions on S represent selection of functions. A formal definition of self-coding is given by Bedian and Herman (1974).

A *self-coding system* of assignment catalysts can be defined by the following conditions (see Figure 4):

(a) All assignments catalyzed by members of a self-coding system belong to a coded subset of all possible assignments. This implies that the S matrix has nonzero components only for the assignments that define a particular code;

(b) All assignments required for the production of any member of the self-coding system from its template also belong to the same code. Thus, the structural matrix M has nonzero values only for the same coding set of assignments as the S matrix.

It should be noted that all proteins involved in the translation process in present day cells satisfy the above properties. These restrictions on the two matrices were probably not present (or were present to much less extent) in pre-code molecular systems, such as the one modeled here. The emergence of coded translation must have involved a gradual evolution of functions (S matrix) and template-product structural relationships (M matrix) from a noncoded to a coded form. These matrices will therefore be used as parameters to investigate the behavior of simple catalyst systems, with the following general questions in mind:

(a) What are the qualitative and quantitative changes in the behavior of the system as the S and M parameters for a subset of components are varied from being noncoded to being coded?

(b) What is the behavior of a system in which all assignments necessary for *two* different codes can be catalyzed by components with coded S and M parameters? Will they coexist and produce ambiguity or will one code be selected? With all corresponding parameters equal, this would represent a system in which neither code is favored through specific interactions or boundary conditions;

(c) What are the general conditions for the dominance of a coded subset of assignment catalysts?

(d) What is the dependence of the efficiency of production of desired components (Q in Equation (1)) on the structural, functional, and kinetic parameters of assignment catalysts?

These and other questions will be approached by formulating kinetic equations for the production of assignment catalysts with chosen function in terms of the required assignments and existing catalytic activities. Although the formulation is general, it has been applied only to the case of two letter codon and amino acid alphabets, and that in a reduced form. The purpose has been to address basic questions, such as the ones above. However, the formalism is applicable to other problems involving the appearance of functional proteins, since a similar structural and functional characterization could be applied.

6. Kinetic Equations for the Case of Two-Letter Alphabets

The detailed kinetics of the polymerization process in the general case has been given by Bedian (1979). Since all of the subsequent analysis refers to the case of two letter alphabets, we shall focus our attention on the dynamics of convergence to a code by construct-

ing a detailed model for a system that distinguishes two types of codons and two types of amino acids. Figure 2 describes the two-letter alphabet system and the four possible assignments for it. Note that assignments AS_1 and AS_4 define a code, and assignments AS_2 and AS_3 define an alternate code. With four assignments in the system, at least four catalysts are needed to describe it adequately. Thus both S and M matrices are of dimensions 4×4.

The reaction rate A_j for each of the four assignments depends on the concentrations and activities of all the catalysts:

$$A_j = \sum_{i=1,4} s_{ij} x_i \quad j = 1, 4, \tag{3}$$

where x_i is the concentration of catalyst i and s_{ij} is the activity of catalyst i for assignment j. Since AS_1 and AS_2 compete for c_1, and AS_3 and AS_4 compete for c_2, the probability of each assignment at the position of the two codons is given by:

$$\begin{aligned} P_j &= A_j/(A_1 + A_2) \quad \text{for } j = 1, 2 \\ P_j &= A_j/(A_3 + A_4) \quad \text{for } j = 3, 4\,. \end{aligned} \tag{4}$$

The fraction of polypeptides produced along a template and satisfying the structural requirements for catalyst n is given by the product of probabilities of all assignments required for its production:

$$Q_n = \prod_{j=1,4} P_j^{m_{nj}}. \tag{5}$$

The four catalyst system is described by the set of four equations:

$$\dot{x}_n = -D_n x_n + C_n + K_n Q_n \quad n = 1, 4\,. \tag{6}$$

7. Reduction of the Four Component System to Two Components

For the purpose of illustrating clearly the competition between two alternative codes, as well as obtaining two-dimensional state space representations of the model, we shall equate corresponding concentrations and parameters of components that belong to one code:

$$\begin{aligned} &x_1 = x_4 = x && x_2 = x_3 = y \\ &C_1 = C_4 = C_x && C_2 = C_3 = C_y \\ &D_1 = D_4 = D_x && D_2 = D_3 = D_y \\ &K_1 = K_4 = K_x && K_2 = K_3 = K_y \\ &s_{11} + s_{41} = s_{14} + s_{44} = s_{xx} && s_{12} + s_{42} = s_{13} + s_{43} = s_{xy} \\ &s_{21} + s_{31} = s_{24} + s_{34} = s_{yx} && s_{22} + s_{32} = s_{23} + s_{33} = s_{yy} \\ &m_{11} + m_{14} = m_{41} + m_{44} = m_{xx} && m_{12} + m_{13} = m_{42} + m_{43} = m_{xy} \\ &m_{21} + m_{24} = m_{31} + m_{34} = m_{yx} && m_{22} + m_{23} = m_{32} + m_{33} = m_{yy}. \end{aligned} \tag{7}$$

The indices 1 and 4, referring to code 1, have been replaces by x; similarly, indices 2 and 3, referring to code 2, have been replaced by y.

With the above substitutions we can define the probability of each code:

$$P_x = P_1 = P_4 = \frac{A_x}{A_x + A_y} = \frac{s_{xx}x + s_{yx}y}{(s_{xx} + s_{xy})x + (s_{yx} + s_{yy})y}$$
$$P_y = P_2 = P_3 = \frac{A_y}{A_x + A_y} = \frac{s_{xy}x + s_{yy}y}{(s_{xx} + s_{xy})x + (s_{yx} + s_{yy})y} . \tag{8}$$

The system of four equations (6) can now be reduced to two equations:

$$\dot{x} = -D_x x + C_x + K_x P_x^{m_{xx}} P_y^{m_{xy}}$$
$$\dot{y} = -D_y y + C_y + K_y P_x^{m_{yx}} P_y^{m_{yy}} . \tag{9}$$

The S and M matrices are of the form:

$$S = \begin{bmatrix} s_{xx} & s_{xy} \\ s_{yx} & s_{yy} \end{bmatrix} \qquad M = \begin{bmatrix} m_{xx} & m_{xy} \\ m_{yx} & m_{yy} \end{bmatrix} .$$

Note that conditions for self-coding for both codes are satisfied when the off-diagonal elements of S and M are zero.

The dynamical and steady state behavior of the system defined by the pair of Equations (9) was investigated by numerical integration of the equations and by linear analysis around steady state points. Prior to the detailed description of these results interpretation of the production efficiencies Q_x and Q_y as selection criteria is discussed.

8. Precode Selection Criteria

Selection criteria have meaning only in the context of a population of reproducing units under specified constraints and mutation rates. While our model pertains only to the internal dynamics of a localized unit, it does provide an explicit expression for the accuracy of self-reproduction of assignment catalysts. These are given by the efficiency functions for the two codes:

$$Q_x = P_x^{m_{xx}} P_y^{m_{xy}} \quad \text{and} \quad Q_y = P_x^{m_{yx}} P_y^{m_{yy}} . \tag{10}$$

$Q_x(Q_y)$ is the fraction of all polypeptides produced along the templates of $x(y)$ catalysts that satisfy the structural requirements for the catalysts. The Q functions are thus the efficiencies with which monomers and energy are utilized to produce code 1 (x) and code 2 (y) components, respectively. The steady state values of Q's are then the accuracies of self-reproduction of the two codes, for a given choice of parameters and initial conditions. Since this model of assignment catalysts can be viewed as a catalytic cycle (i.e., an autocatalytic unit), the results obtained by Eigen and Schuster (1977) would prescribe the selection of the unit with highest accuracy of self-reproduction. Thus, the steady state values of Q_x or Q_y can be used as a measure of the selective advantage of x or y components.

9. General Topological Behavior of the System

Numerical integration of Equations (9) provided state space trajectories for different choices of parameters and initial conditions. The procedure used was a fourth-order Runge–Kutta method (see Ralston, 1965, Chapter 5). Such computations yielded three classes of results:

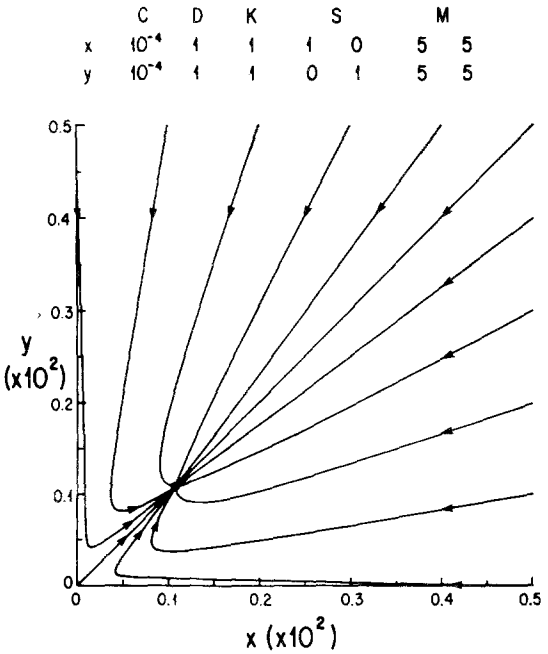

Fig. 5. State space trajectories from numerical integration, leading to a single, stable noncoded steady state on the diagonal. The S matrix is coded, but M is not. The system has the same topology when S is not coded, but M is.

(a) A single steady state which is a stable node can exist for the system. This node may correspond to a code ($P_x = 1$ or $P_x = 0$), or may have some intermediate value of P_x. Figure 5 illustrates the state space trajectories of a single, non-coded state.

(b) Three steady states, two of which are stable nodes separated by a saddle, are possible for a wide range of parameter values. When structural and functional parameters are coded for both x (code 1) and y (code 2) components, the two stable nodes each correspond to one of the two competing codes (Figure 6), the fate of the system being determined by initial conditions. Figure 7 illustrates an intermediate system in the transition from a single, ambiguous state (Figure 5) to two coded states (Figure 6).

(c) Under symmetric or nearly symmetric conditions, five steady states can be obtained: three stable nodes separated by two saddles (Figure 8). Thus, an ambiguous state can coexist with two coded states, and initial conditions will again determine the fate of the system.

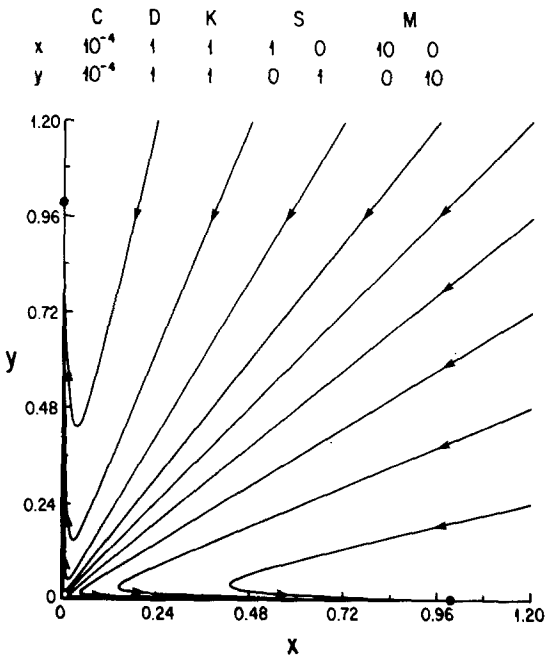

Fig. 6. Trajectories when S and M satisfy the conditions for self-coding for both codes. The steady state on the diagonal is now unstable (open circle) but two stable states (closed circles) are now available through a bifurcation. These two states correspond to the dominance of code 1 and code 2, respectively. Thus, ambiguity is not a necessary consequence of having coded catalysts for both codes.

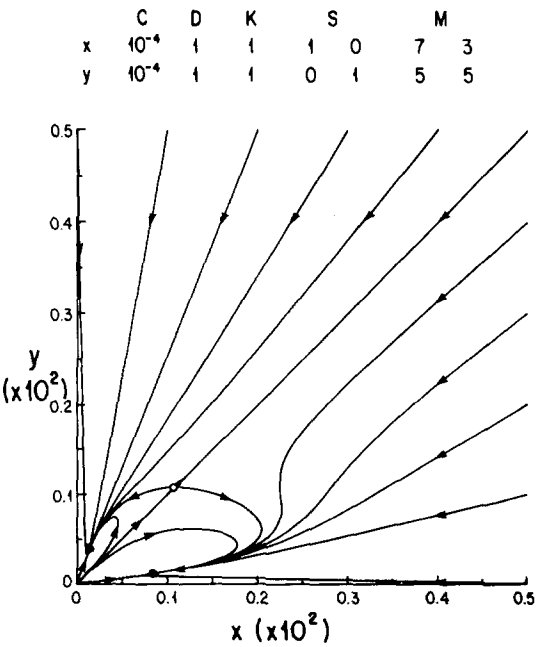

Fig. 7. Appearance of an asymmetrical bifurcation as the m-vector for x components starts favoring code 1.

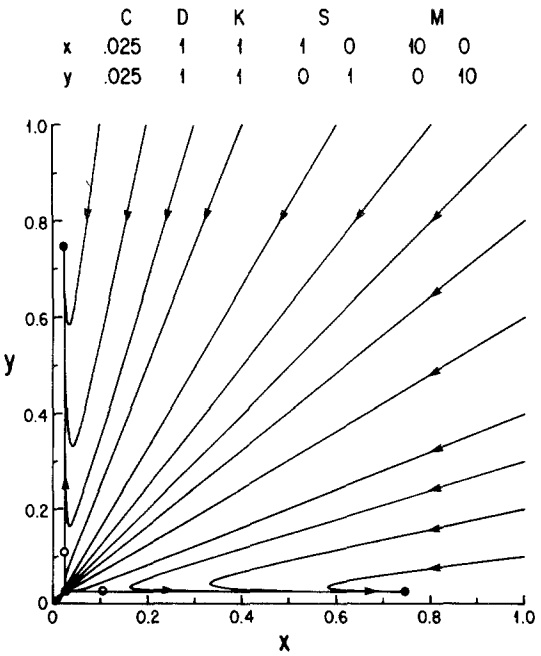

Fig. 8. Five steady states appear as C parameters are increased. The three stable states correspond to code 1 dominant, ambiguous and code 2 dominant, respectively, separated by two saddles. The fate of the system depends on initial conditons as well as the position of the two saddle points.

Two or all three of the above topological behaviors are generally observed as one of the parameters is varied. Transitions between 1, 3 and 5 steady state situations occur through production or annihilation of pairs of steady states. The five steady state behavior may appear in a region of parameter space connecting two regions both with 3 steady states, or a region with 1 and another with 3 steady states. When more than one stable state is available, the positions and separatrices of saddle points determine the regions of state space that lead to each of the stable points.

For purposes of systematically investigating the effect of different parameters, the steady states of the system were found and linear analysis was performed around them to determine their asymptotic stability (Minorsky, 1962, Chapter 1). Steady states of the system were found by imposing the condition $\dot{x} = \dot{y} = 0$ on Equations (9), which could be expressed as a polynomial in P_x. Each real root of the polynomial between zero and one corresponded to a steady state of the system.

In the following section, graphs of steady state values of P_x are presented as a function of different parameters. P_x is a convenient function to describe changes in steady state behavior with respect to a given parameter, as well as an indicator of the type and degree of 'codedness' of each steady state. $P_x = 1 (P_y = 0)$ corresponds to dominance of code 1, while $P_x = 0 (P_y = 1)$ corresponds to dominance of code 2. $P_x = P_y = 0.5$ is a non-coded, ambiguous state. These graphs are presented in pairs: one where the M matrix is symmetrical in x and y (i.e., $m_{xy} = m_{yx}$); and another where M is asymmetrical with the constraint $m_{yx} = m_{yy}$ (i.e., y genetics do not satisfy the self-coding conditions for code 2).

Even this simplified system has a large parameter space (13 independent parameters), and a complete search would be very time consuming. Therefore, only situations and parameter variations relevant to the code origin and stability problems are presented.

10. Conditions for Coding

Variation of the kinetic parameters C, D, K, as well as the structural and functional matrices M and S, reveals the conditions necessary for the appearance of coded states (i.e., stable steady states where $P_x = 1$ or $P_x = 0$). One general and obvious result is that one of the codes can dominate if its catalysts are favored through a relatively low decay constant or relatively high production rate. The more interesting question, however, is to find the necessary conditions without relying on such preferential kinetics. These conditions are described below:

(a) Strong stereochemical effects can determine the dominant code. Figure 9 demonstrates the behavior of the system when s_{xx} is varied while s_{yy} is held constant, corresponding to varying the relative rates of assignments of the two codes. Dominance of a code can be ensured if stereochemical factors lead to relatively high activities for that code. When both codes are possible, the position of the saddle varies such that more of the state space (i.e., initial conditions) leads to the favored code. Two points are worth noting: First, when the structural matrix is coded for x (code 1) but not for y (code 2), only a *strong* stereochemical bias can ensure the dominance of code 2 ($P_x = 0$ only when $s < 0.01$). Second, with the structural matrix M coded for both x and y, stable states for *both* codes are possible for a whole range of stereochemical biases ($0.01 < s < 100$), suggesting that special initial conditions could have led to a 'frozen accident' even in the presence of stereochemical biases. However, this range is smaller, and the effect of biases stronger, when M is less coded. Thus, stereochemical factors could

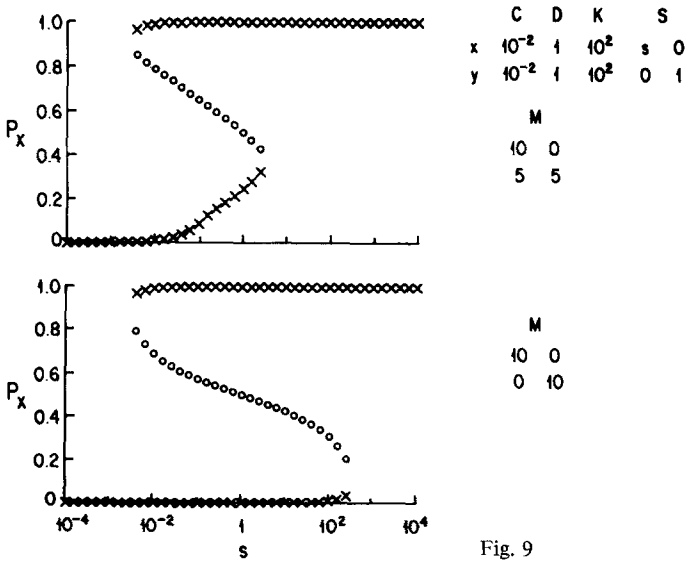

Fig. 9

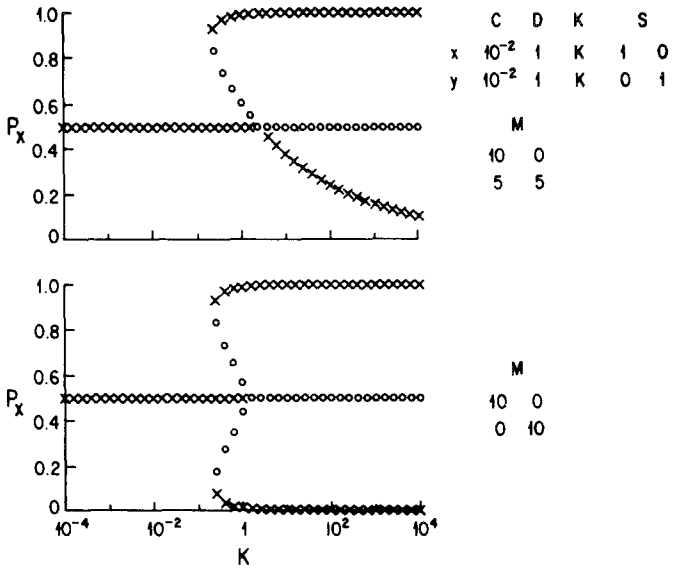
C D K S
x 10^{-2} 1 K 1 0
y 10^{-2} 1 K 0 1
M
10 0
5 5
M
10 0
0 10
P_x
K
10^{-4}
10^{-2}
1
10^{2}
10^{4}

Fig. 10

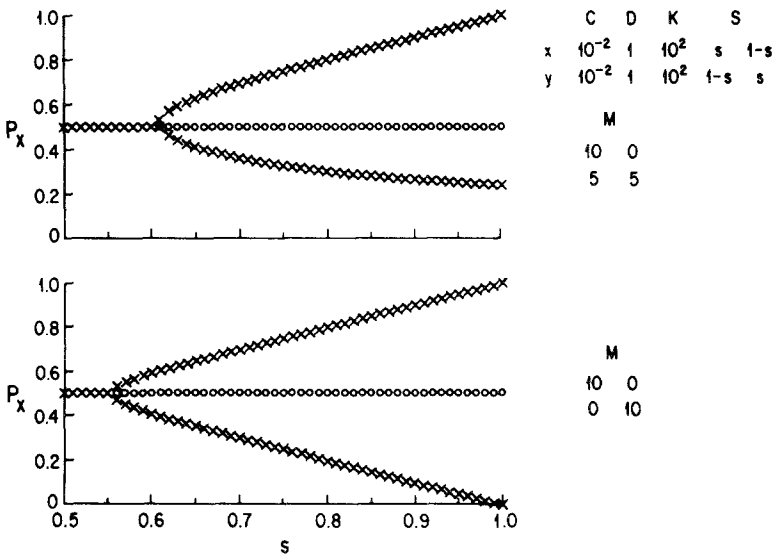
C D K S
x 10^{-2} 1 10^{2} s 1-s
y 10^{-2} 1 10^{2} 1-s s
M
10 0
5 5
M
10 0
0 10
P_x
s
0.5
0.6
0.7
0.8
0.9
1.0

Fig. 11

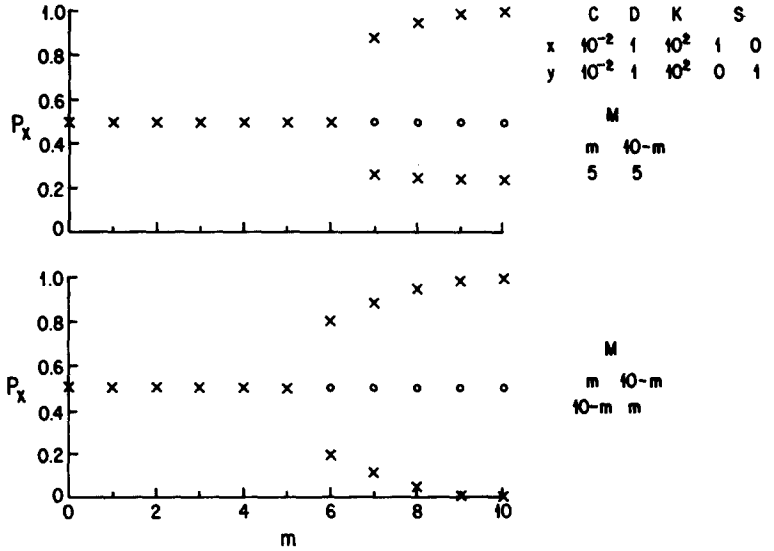

Fig. 12

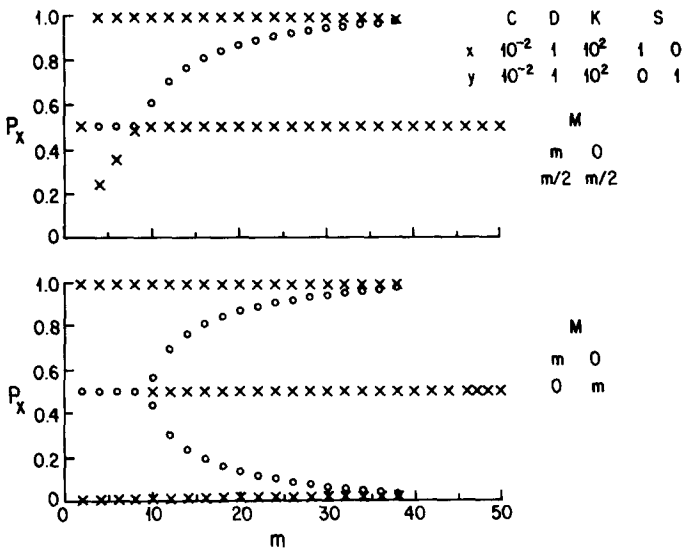

Fig. 13

Figs. 9–13. Bifurcation plots represented by steady state values of P_x as one of the parameters is varied. Stable states are represented by crosses, unstable states by circles. A stable state at $P_x = 1$ is code 1 dominant, at $P_x = 0$ is code 2 dominant, at $P_x = 0.5$ is ambiguous. Each figure is a pair of diagrams, one where M is coded for x components only, another where M is coded for both x and y.

This interpretation of the particular parameter varied in each figure is given in the text.

have been important in early events of code nucleation, while kinetic bifurcations could have played a significant role in the process of evolution towards a code. While it is reassuring that the model can account for stereochemical biases, no such bias is assumed when investigating the effects of other parameters.

(b) The polynucleotide directed production of polypeptides should be the dominant process. This condition is demonstrated in Figure 10, where K parameters are varied while holding C constant. For our model system, coded states are possible only when $K/C > 40$, where K and C represent the contributions of polynucleotide directed and random processes, respectively, to polypeptide production. Note that as K increases, the stable, ambiguous state at $P_x = 0.5$ becomes unstable through a sequence of two bifurcations, and one or two stable, coded states appear.

(c) The functional and structural matrices S and M should satisfy the conditions for self-coding. These two conditions are demonstrated in Figures 11 and 12. The S matrix is varied from the non-coded to the coded form in Figure 11, while M is coded. In Figure 12, S is kept coded, while M is varied from non-coded to coded forms. The bifurcations resulting in these two situations are of the same general form: the stable state at $P_x = 0.5$ is transformed to a saddle, while one (or both) of the two new stable states approach coding ($P_x = 1$ or $P_x = 0$). It is worthwhile to note again that when self-coding conditions are satisfied for catalysts of both codes, ambiguity need not result. Depending on where it starts, the system will approach one of two stable states, and one of the two codes will dominate.

(d) The total number of critical sites should not be large. Figure 13 demonstrates the bifurcations resulting from a progressive increase in the total number of critical sites. Referring to the case with symmetrical M, it can be seen that the two coded states are the only possible stable states until the number of critical sites reaches 10. At this point, bifurcation into five steady states allows coexistence of two coded and one ambiguous stable state. As the number of critical sites increases further, the saddle points approach the coded states (less of state space leads to coding) and eventually annihilate them, leaving behind a single, ambiguous stable state. These results imply that the more stringent structural requirements are for a desired function, the less likely is the system to be at a stable, coded state.

11. Evolution to a Code

Of particular interest is the dependence of the selection criterion Q_x on the S and M parameters. Keeping the kinetic parameters at values favorable to a code, Q_x (Equation (10)) was evaluated at the steady state approaching code 1 (when more than one stable state was possible), using S and M matrices of the form:

$$S = \begin{bmatrix} s & 1-s \\ 1-s & s \end{bmatrix} \quad \text{and} \quad M = \begin{bmatrix} m & 10-m \\ 5 & 5 \end{bmatrix}.$$

The S matrix was kept symmetrical because similar selectivities could be expected from x and y components when the total number of critical sites is the same. The s parameter

was varied between 0.5 (non-coding) and 1.0 (coding); and m was varied between 5 (non-coding) and 10 (coding for x). Figure 14 shows that Q_x increases sharply as S and M approach the coding forms (i.e., $s = 1$, $m = 10$). Thus, in a population of localized units where m and s could vary, some units would be expected to evolve to a coded state by progressive increases in s and m.

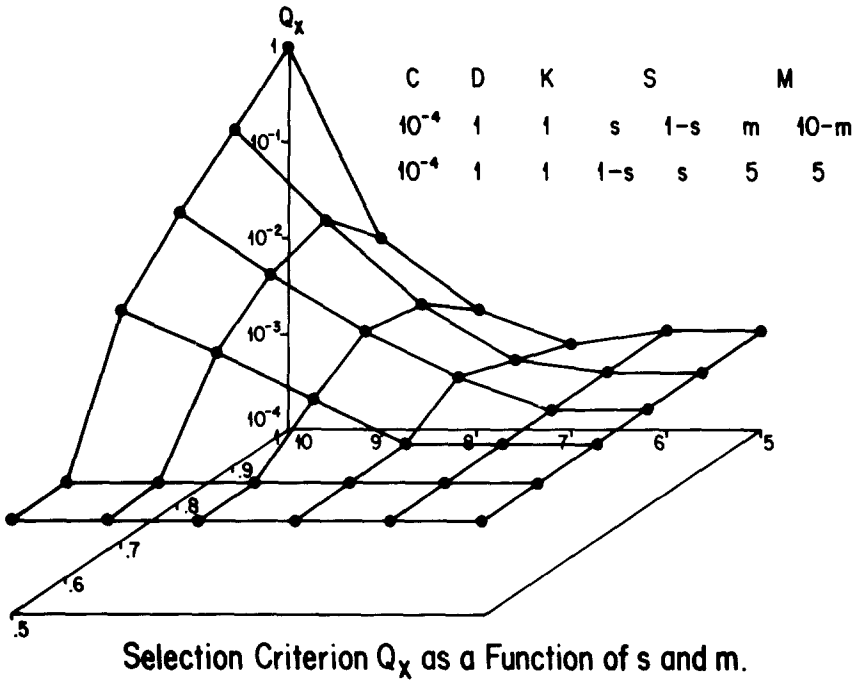

Fig. 14. Steady state values of the selection criterion Q_x as a function of translational accuracy (s) and template-product structural relationship (m). Q_x has its maximum when both s and m are coded ($s = 1$, $m = 10$), and an optimal pathway (variation of s and m) can lead to optimization of the selection criterion and establishment of coded translation with concurrent selection of coded templates.

Several important conclusions can be drawn from these results:

(a) The self-reproductive efficiency Q_x has the appropriate shape for acting as a selection criterion, and may have been the driving force for the selection of a coded set of assignment catalysts and their respective templates;

(b) The surface described by Q_x versus s and m has a 'ridge' such that for each choice of template-product structural relationship (m) there exists a value of translational accuracy (s) that optimizes the selection criterion (Q_x). These values of m and s correspond to an optimal pathway of evolution to a code. The gradual changes in a population of protocells towards coded translation would allow for coevolution of templates of other essential components to coded forms. Even if the selective advantage depended on an essential component which was not an assignment catalyst, a similar Q surface would be obtained by using its structural parameters and the probabilities of the two code assignments at a steady state;

(c) When structural parameter (m) is not coded (i.e., $m < 10$) the optimal value of translational accuracy (s) is less than 1. While present day cells generally do not tolerate a significant level of errors in translation, suppression through a mutation or external agents are well-known examples of *optimized* translational ambiguity.

12. Simulating Suppression of Deleterious Mutations

A deleterious point mutation, such as a nonsense mutation resulting in a defective essential enzyme, can be suppressed by introducing some ambiguity in the assignment of the codon at the mutation site. By misreading the mutated codon, enough active enzyme is produced to make the organism viable. This can be achieved either through a second mutation affecting translational accuracy, or through external agents such as antibiotics (Gorini and Kataja, 1965; Gorini, 1974) and organic solvents (Gado and Horvath, 1963). The phenotypic rescue induced by streptomycin in an arginine auxotroph of *E. coli* is exhibited by the ability of the defective strain to grow in the presence of streptomycin, while no growth is observed in minimal medium (Gorini and Kataja, 1964a). The mutation is in the structural gene for the metabolic enzyme ornithine transcarbamylase (OTC), which is not detected in the mutant strain but was measured in increasing amounts with increasing concentrations of streptomycin in the medium.

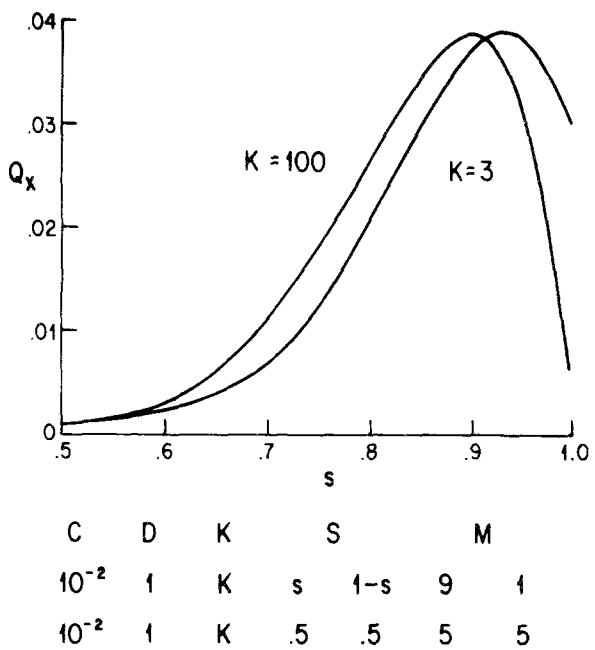

Fig. 15. Selection criterion Q_x as a function of translation accuracy (s) when the m-vector is a single step from being coded for x components ($m_{xx} = 9$, $m_{xy} = 1$), simulating the conditions for suppression of a deleterious mutation. The selection criterion peaks when there is some ambiguity in translation.

The beneficial effects of translational ambiguity when an essential protein does not have a perfectly coded gene is represented in the model by the shape of the selection criterion Q_x as a function of gene structure (m) and translational accuracy (s). Figure 15 is a cross-section of the Q_x surface at $m = 9$, representing a point mutation in the structural gene of an essential component. Under these conditions Q_x is not a monotonically increasing function of accuracy, but achieves a maximum at a finite rate of translational errors ($s < 1$). The loss of selective value at higher translational accuracies (higher s) is very dramatic, particularly when the polynucleotide directed production of polypeptides is the dominant process (large K). This corresponds to the mutant ($m = 9$) not being able to grow (low Q) without the translational errors induced by streptomycin ($s = 1$). The highest value achieved by Q_x is significantly lower than the maximum value achievable when the gene is perfectly coded ($m = 10$ in Figure 14). This is in qualitative agreement with Gorini and Kataja's (1964a) observation that the optimal growth rate of OTC-defective strains in the presence of streptomycin was only half of that achieved by removing the selection pressure (addition of essential amino acid to medium).

Another observation made in these experiments is somewhat puzzling. While the growth rate was close to its optimum with 5 μg/ml of streptomycin, it stayed at this value with antibiotic concentrations as high as 500 μg/ml. The curves in Figure 15 indicate that increasing translational errors beyond the optimal value would also decrease selective advantage. This seeming discrepancy may be explained by the sophistication of the present day ribosome. Since misreading has been shown to be influenced by the reading context (Davies *et al.*, 1966), it was postulated that a missense suppressor, *Su* of OTC$^-$ was capable of specifically misreading the mutated codon by recognizing that it is 'out of context', and the ribosomal alteration mutant *ram* is thought to abolish this ribosomal specificity. Biswas and Gorini (1972) observed severe inhibition of growth in strains that contain mutant *ram* as well as a missense suppressor locus. Growth of such strains can be completely inhibited by 50–200 μg/ml of streptomycin. The higher levels of translational error in the presence of *ram* was confirmed by the production of up to 60% faulty proteins, as determined by measurement of immunologically cross-reacting material replacing particular enzymes. If the wild-type ribosome can indeed select the location of suppressor action, then the *ram* mutant which abolishes this sophisticated specificity would probably be more representative of the primitive system considered in the model. Moreoever, there is evidence for the decrease in selective value at higher error rates even in the absense of *ram*. Gorini and Kataja (1964b) have isolated mutants which are OTC$^-$, and have the suppressor locus *Su*. When the effect of streptomycin is compounded to this suppressor mutation, decrease in growth rates and eventually arrest of growth with 12–50 μg/ml of antibiotic is observed. Simultaneous action of the suppressor locus and streptomycin cause more than additive increase in OTC production, but several other metabolic enzymes are severly decreased.

These results are in good qualitative agreement with the behavior predicted by the model: a mutation at a critical site of an essential component (OTC$^-$) drastically reduces selective advantage if the translation system is very accurate, but the introduction of translational errors (streptomycin) improves selective value through suppression. Further

increase in random translational errors (*ram* or *Su* mutation plus streptomycin) will eventually take the selective value beyond its optimum. It is important to note that lethality is not a necessary consequence even with significant levels of translational error, indicating that much of the Q_x surface in Figure 14 may correspond to viable systems capable of competition and evolution.

13. Possible Experimental Verification of the Model

The experimental evidence discussed in the previous section demonstrates the ability of the model to account for the effect of translational ambiguity in *E. coli* of various genotypes, and indicates that the basic assumptions of the model are compatible with the organization of simple living systems. Although further data could be obtained on the relation of growth rates of wild type and auxotropic bacteria to translational errors introduced by ribosomal alterations, tRNA suppressors, or external agents, this line of evidence is of limited value in establishing whether or not the proposed model could indeed be the basis for the origin of the genetic code. Similar limitations exist for the approach developed by Gallant and coworkers (Gallant and Palmer, 1979; Gallant and Prothero, 1980) for testing models of error propagation and error catastrophe.

Although studies on translational errors in living cells are bound to be very useful, direct evidence concerning the origin of the genetic code can best be obtained through abiotic synthesis experiments and cell-free systems. The assumption in our model that enzymatic activity and selectivity for assignments can be observed in relatively short polypeptides made of only a few amino acids needs special attention. One experimental approach to this important issue is to use only three amino acids (neutral, charged, and hydrophobic), and construct thermal proteinoids or purer polypeptides (e.g. copolymers of short peptides of the amino acids). These polypeptides can then be tested for assignment catalyst activity in an *in vitro* translation system without elongation factors (Spirin, 1976). Such a system would have a very low level of peptide formation and any significant catalytic activity by the test polypeptides could be detected. If such a component is found its specificity could be tested with the double-label method of Govrilova *et al.* (1981): PolyA or polyU is used as the 'message', in each case supplying the system with aminoacyl-tRNAs with a radiolabel on the amino acid. The correct aminoacyl-tRNA would be labeled with H^3, and incorrect ones with C^{14}. The ratio of C^{14}/H^3 counts in polypeptides produced would be a measure of the specificity of the test polypeptide. Alternatively, the synthetic polypeptides could be tested in a system with aminoacyladenylates and tRNAs for possible aminoacyl-tRNA synthetase activity and specificity. Such activity would support our premise that primitive assignment catalysts could be found in a population of random polypeptides. It would also allow determination of the effect of number of amino acids, chain length, and primary sequence on the observed activities and specificities. Such an approach would test both a basic assumption of the model (that a small amino acid alphabet could be an adequate starting point), and the prediction that the number of critical sites required for function could not be too large.

An alternative approach may provide a more direct test of the dynamics predicted by

the model. Elongation factor Tu has recently been shown to effect translational accuracy under some experimental conditions (Gavrilova *et al.*, 1981). If the structural gene for Tu is cloned, it could be used to enrich a total mRNA extract of *E. coli* for Tu message by hybridization and binding to hydroxyapatite. Such an enriched fraction could be used to drive the *in vitro* translation system with different initial concentrations of Tu, and the steady state concentration measured. Such a system, if realizable, could also be tested with controlled levels of translational error, introduced by the use of antibiotics or ribosomes from *ram E. coli*, while Tu message from different strains or mutagenized bacteria may provide a method of varying the template-product structural relationship. With both the structural and functional parameters, as well as initial concentrations of Tu, under experimental control, the bifurcations and multiple steady states predicted by the model can be tested.

14. Discussion

The model presented here is based on three major premises. One is the autocatalytic nature of polynucleotides, which is assumed to lead to their stable presence under prebiotic conditions; the second is realizability of catalytic functions by polypeptides with sufficient length and specification of primary structure; the third is colinearity within restricted subsets of the two classes of polymers, presumably established through the dominance of polynucleotide directed polypeptide formation. Under these conditions the assignment catalysts would form an ambiguous, non-coded system. Such systems would evolve towards coded translation, enabling them to achieve high efficiency (accuracy) of self-reproduction. Efficiency of monomer and energy utilization is the proposed selection criterion for this evolution.

The bifurcation plots presented in Sections 9 and 10 demonstrate that a rich variety of behaviors can be observed even in this simplified system of assignment catalysts competing under the constraint of constant total production rate of polypeptides. The system as a whole can approach a stable stationary point where x and y (representing the two alternative codes in the system) coexist in comparable amounts, but with low efficiency of self-reproduction for both. Although this state has been labeled noncoding, it is nevertheless autocatylitic. It corresponds to a primitive mechanism for producing relatively long polypeptides, but without much selection of particular primary sequences. As a first step in the nucleation of a code, protocells capable of such coupled polymerization may have been selected over those relying only on random polymerization or the external supply of polymers for growth and propagation. Stereochemical complementarity may have played a strong role in establishing these initial, ambiguous systems. Carter and Kraut (1974) have constructed a molecular model showing a very good fit between polypeptide and RNA double helices, and have proposed that such a fit was the basis of a primitive translation machinery.

Selectivity and self-reproductive efficiency are much higher when one of the codes dominates at a steady state. The latter behavior is possible when the template directed and catalytically controlled production of polypeptides is the dominant process (large K),

the total number of critical sites is not very large, and the s (activity) and m (structural) vectors of one *or both* components are coded. It is clear that equiprobability of all assignments does not necessarily lead to an ambiguous state. If the S and M matrices are coding for both components, two stable stationary points, one for each code, are possible. When multiple stable points are available, the fate of a particular unit depends on its initial conditions. In a population of protocells, those having more coded assignment catalysts would utilize raw materials more efficiently and have a selective advantage. With some variability in template structure, evolution to a highly coded translation machinery becomes an inevitable consequence of the competition for monomers and energy.

There is a close correspondence between results of our model and error catastrophe models of Hoffman (1974), Kirkwood and Holliday (1975), Goel and Ycas (1975, 1976) and Goel and Islam (1977). Error catastrophe (i.e., a non-coding state) is inevitable for nonspecific catalysts, but it may or may not exist as an option for more specific components. Thus a threshold effect with respect to specificities as well as initial conditions can be observed. However, similar effects can be observed with respect to the structural and the kinetic parameters.

The simple system presented here was developed for the purpose of elucidating the role of kinetic, functional and structural parameters in the evolution of a self-coding system of assignment catalysts. Many of the simplifying assumptions can be removed to yield generalized versions of the model, although their numerical investigation may become significantly more difficult and time consuming. Nevertheless, specific effects and processes can be investigated at the expense of other simplifying assumptions. Included in these effects are: relative abundance of amino acids and codons; reversibility of polymerization reactions; and explicit representation of polynucleotides, replicases and other functional components. Such components may very well be control elements for the overall translation process, and their own production will in turn depend on the translation system and its accuracy. In particular, the introduction of polymerases and polynucleotide dynamics would complete the hypercyclic coupling between the two polymer classes, and should lead to mutual stabilization of polymer populations (Eigen and Schuster, 1978a). The formalism can also be applied to the problem of code evolution through further inclusion or distinction of amino acids and codons. Preliminary results (Bedian, 1979) indicate that similar bifurcations would be involved in models of code evolution.

Acknowledgements

This work was part of the author's doctoral research, conducted with the advice and guidance of Dr H. H. Pattee and Dr R. A. Spangler. The conceptual and mathematical development has also benefited from discussions with Dr S. Colombano, Van Kurtz, and Dr S. Kauffman. Special thanks are due to Ms. Carla Hultman for typing the manuscript.

References

Bedian, V.: 1979, Ph. D. dissertation, Dept. of Biophysical Sciences, SUNY at Buffalo.
Bedian, V. and Herman, G. T.: 1974, in *Proceedings of Workshop on Developmental Languages and Related Topics*, Springer-Verlag, Heidelberg.
Biswas, D. K. and Gorini, L.: 1972, *J. Mol. Biol.* **64**, 119.
Carter, C. W. and Kraut, J.: 1974, *Proc. Natl. Acad. Sci. U.S.A.* **71**, 283.
Crick, F. H. C.: 1968, *J. Mol. Biol.* **38**, 367.
Crick, F. H. C., Brenner, S., Klug, A., and Pieczenik, G.: 1976, *Origins of Life* **7**, 389.
Davies, J. E., Jones, D. S. and Khorana, H. G.: 1966, *J. Mol. Biol.* **18**, 48.
Eigen, M.: 1971, *Naturwissenschaften* **58**, 465.
Eigen, M. and Schuster, P.: 1977, *Naturwissenschaften* **64**, 541.
Eigen, M. and Schuster, P.: 1978a, *Naturwissenschaften* **65**, 7.
Eigen, M. and Schuster, P.: 1978b, *Naturwissenschaften* **65**, 341.
Fox, S. W.: 1974a, *Mol. Cell. Biochem*, **3**, 129.
Fox, S. W.: 1974b, in K. Dose, S. W. Fox, G. A. Deborin, T. E. Pavlovskaya (eds.) *The Origin of Life and Evolutionary Biochemistry*, Plenum Press, New York, 119.
Fox, S. W. and Dose, D.: 1977, *Molecular Evolution and the Origin of Life*, Marcel Dekker, Inc., New York, p. 173.
Fox, S. W., Jungek, J. R. and Nakashima, T.: 1974, *Origins of Life* **5**, 227.
Gado, I. and Horvath, I.: 1963, *Life Sciences* **10**, 741.
Gallant, J. A. and Palmer, L.: 1979, *Mechanisms of Aging and Development* **10**, 27.
Gallant, J. A. and Prothero, J.: 1980, *J. Theor. Biol.* **83**, 561.
Gavrilova, L. P., Perminova, I. N., and Spirin, A. S.: 1981, *J. Mol. Biol.* **149**, 69.
Goel, N. S. and Islam, S.: 1977, *J. Theor. Biol.* **58**, 167.
Goel, N. S. and Ycas, M.: 1975, *J. Theor. Biol.* **54**, 245.
Goel, N. S. and Ycas, M.: 1976, *J. Math. Biol.* **3**, 121.
Gorini, L.: 1974, in Nomura (ed.), *Ribosomes*, Cold Spring Harbor, 791.
Gorini, L. and Kataja, E.: 1964a, *Proc. Natl. Acad. Sci. U.S.A.* **51**, 487.
Gorini, L. and Kataja, E.: 1964b, *Proc. Natl. Acad. Sci. U.S.A.* **51**, 995.
Gorini, L. and Kataja, E.: 1965, *Biochem. Biophys. Res. Comm.* **18**, 656.
Hoffman, G. W.: 1974, *J. Mol. Biol.* **86**, 349.
Hoffman, G. W.: 1975, *Ann. Rev. Phys. Chem.* **26**, 123.
Jukes, T. H.: 1973, *Nature* **246**, 22.
Jukes, T. H.: 1974, *Origins of Life* **5**, 331.
Kirkwood, T. B. L. and Holliday, R.: 1975, *J. Mol. Biol.* **97**, 257.
Lacey, J. C., Weber, A. L., and White, W. E.: 1975, *Origins of Life* **6**, 273.
Minorsky, N.: 1962, *Nonlinear Oscillations*, D. Van Nostrand, Princeton, N.J.
Mizutani, M. and Ponnamperuma, C.: 1977, *Origins of Life* **8**, 183.
Oparin, A.: 1964, *Life, its Nature, Origin and Development*, Academic Press, New York.
Oparin, A. I.: 1966, *The Origin and Initial Development of Life*, Meditsina Publishing House, Moscow.
Orgel, L. E.: 1963, *Proc. Natl. Acad. Sci. U.S.A.* **49**, 517.
Orgel, L. E.: 1970, *Proc. Natl. Acad. Sci. U.S.A.* **67**, 1476.
Pelc, S. R. and Welton, M. G. E.: 1966, *Nature* **209**, 868.
Ralston, A.: 1965, *A First Course in Numerical Analysis*, McGraw-Hill, New York.
Saxinger, C. and Ponnamperuma, C.: 1971, *J. Molec. Evol.* **1**, 63.
Spirin, A. S.: 1976, *Origins of Life* **7**, 109.
Wagner, K. G. and Arav, R.: 1968, *Biochemistry* **7**, 109.
Walker, G. W. R.: 1974, *Origins of Life* **5**, 351.
Weber, A. L. and Lacey, J. C. Jr.: 1974, *Biochem. Biophys. Acta* **349**, 226.
Wong, J. T.: 1975, *Proc. Natl. Acad. Sci. U.S.A.* **72**, 1909.
Wong, J. T.: 1976, *Proc. Natl. Acad. Sci. U.S.A.* **73**, 2336.