Biophysical Chemistry 16 (1982) 329–345
Elsevier Biomedical Press

SELF-REPLICATION WITH ERRORS

A MODEL FOR POLYNUCLEOTIDE REPLICATION **

Jörg SWETINA and Peter SCHUSTER *

Institut für Theoretische Chemie und Strahlenchemie der Universität, Währingerstraße 17, A-1090 Wien, Austria

Received 4th June 1982
Revised manuscript received 23rd August 1982
Accepted 30th August 1982

Key words: Polynucleotide replication; Quasi-species; Point mutation; Mutant class; Stochastic replication

A model for polynucleotide replication is presented and analyzed by means of perturbation theory. Two basic assumptions allow handling of sequences up to a chain length of $\nu \approx 80$ explicitly: point mutations are restricted to a two-digit model and individual sequences are subsumed into mutant classes. Perturbation theory is in excellent agreement with the exact results for long enough sequences ($\nu > 20$).

1. Introduction

Eigen [8] proposed a formal kinetic equation (eq. 1) which describes self-replication under the constraint of constant total population size:

$$\frac{\mathrm{d}x_i}{\mathrm{d}t} = \dot{x}_i = \sum_j w_{ij} x_j - \frac{x_i}{c}\phi;\ i = 1,\ldots,n^{\dagger} \tag{1}$$

By x_i we denote the population number or concentration of the self-replicating element I_i, i.e., $x_i = [I_i]$. The total population size or total concentration $c = \Sigma_i x_i$ is kept constant by proper adjustment of the constraint ϕ: $\phi = \Sigma_i \Sigma_j w_{ij} x_j$. Characteristically, this constraint has been called 'constant organization'. The relative values of diagonal (w_{ii}) and off-diagonal (w_{ij}, $i \neq j$) rates, as we shall see in detail in section 2, are related to the accuracy of the replication process. The specific properties of eq. 1 are essentially based on the fact that it leads to exponential growth in the absence of constraints ($\phi = 0$) and competitors ($n = 1$).

The non-linear differential equation, eq. 1 – the non-linearity is introduced by the definition of ϕ at constant organization – shows a remarkable feature: it leads to selection of a defined ensemble of self-replicating elements above a certain accuracy threshold. This ensemble of a master and its most frequent mutants is a so-called 'quasi-species' [9]. Below this threshold, however, no selection takes place and the frequencies of the individual elements are determined exclusively by their statistical weights.

Rigorous mathematical analysis has been performed on eq. 1 [7,15,24,26]. In particular, it was shown that the non-linearity of eq. 1 can be removed by an appropriate transformation. The eigenvalue problem of the linear differential equation obtained thereby may be solved approximately by the conventional perturbation technique

* Dedicated to the late Professor B.L. Jones who was among the first to do rigorous mathematical analysis on the problems described here.

** This paper is considered as part II of Model Studies on RNA replication. Part I is by Gassner and Schuster [14].

† All summations throughout this paper run from 1 to n unless specified differently: $\Sigma_i = \Sigma_{i=1}^{n}$ and $\Sigma_{i,i\neq j} = \Sigma_{i=1}^{j-1} + \Sigma_{i=j+1}^{n}$, respectively.

of linear algebra (see section 2).

The main, although not exclusive, field of application for eq. 1 is polynucleotide replication. Therefore, the formal treatment was encouraged enormously by the results of experimental work on template-directed RNA synthesis (for reviews, see refs. 12,17,21 and 22; as well as refs. 3 and 20). These studies presented, among many other interesting aspects, experimental verifications of the predictions of the analysis of eq. 1: the concentrations of polynucleotides grow exponentially under proper conditions – at concentrations below enzyme saturation – and the individual growth properties are evaluated by selection.

A second series of studies was performed by Weissmann and co-workers [2,5,6] on the simple bacteriophage Qß. They were able to present experimental proof for the existence of a mutant distribution in the naturally occurring phage populations. This distribution fits perfectly into the concept of a quasi-species. Later on, nucleotide sequence heterogeneities were also found in foot-and-mouth disease virus [4] and influenza virus populations [13,19].

In this paper, we investigate the nature of the approximations used in the perturbational analysis of the linear differential equation mentioned above and show its validity for long polynucleotide sequences by means of a properly chosen model. This model is based on experimental frequencies of mutations and seems to be suitable also for other studies on polynucleotide replication in general.

2. The origin of the problem

In order to illustrate the application of perturbation theory to eq. 1 we have to specify the rates w_{ij} in more detail. Differing slightly from the original version [8], we treat diagonal and off-diagonal rates by the same equation:

$$w_{ij} = A_j Q_{ij} - D_j \delta_{ij} \tag{2}$$

where A_j is the rate of polymer synthesis on the template I_j, i.e., the number of newly synthesized molecules per unit time and unit template concentration, irrespective of whether they are correct copies or mutants. D_j is the rate of degradation of polymers I_j (by δ_{ij} we denote the Kronecker symbol: $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ if $i \neq j$). The accuracy of the replication process is described by the matrix $Q = \{Q_{ij}\}$. The elements of Q are dimensionless and Q_{ij} is the probability of obtaining molecule I_i through a replication of I_j. Evidently, the diagonal elements Q_{ii} are identical with the quality factors Q_i of Eigen [8]. The off-diagonal elements Q_{ij} $(i \neq j)$ are simply the probabilities of mutations. From the definition of probabilities follows immediately

$$\sum_i Q_{ij} = 1 \tag{3}$$

(a synthesized copy is either correct or erroneous). Making use of eq. 3, we can rewrite the constraint ϕ:

$$\phi = \sum_i \sum_j w_{ij} x_j = \sum_i (A_i - D_i) x_i = \sum_i E_i x_i = c\bar{E} \tag{4}$$

We introduced an excess productivity $E_i = A_i - D_i$ of template I_i as well as its ensemble average: $\bar{E} = \sum_i E_i x_i / c$.

Without losing generality, we can restrict ourselves to the case $c = \sum x_i = 1$, since the solution curves of eq. 1 do not depend on total concentration c after transformation to internal coordinates: $\xi_i = x_i / c$ [10]. Eq. 1 now is of the form

$$\dot{x}_i = (w_{ii} - \bar{E}) x_i + \sum_{j, j \neq i} w_{ij} x_j \tag{5}$$

Intuitively, eq. 5 describes a selection process. Due to elimination of the less efficiently replicating polynucleotides – for these molecules $w_{ii} - \bar{E}$ is negative – $\bar{E}$ increases monotonically in time until it reaches an optimum at the stationary state:

$$\lim_{t \to \infty} \bar{E} = \bar{E}_{\mathrm{opt}} = w_{\mathrm{mm}};\ w_{\mathrm{mm}} = \max(w_{ii};\ i = 1, \ldots, n) \tag{6}$$

This selection process may be disturbed by mutations in case the w_{ij} values are large. As we shall see this intuition is essentially correct. We find selection if the mutation terms in eq. 5 are sufficiently small. Nevertheless, we summarize the rigorous analysis of eq. 5 first.

Following Thompson and McBride [24] or Jones et al. [15], we may remove the non-linearity in eq.

4 by a time-dependent transformation of variables

$$x_i(t) = \exp\left\{-\int_0^t \bar{E}(\tau)\mathrm{d}\tau\right\} z_i(t) \tag{7}$$

and are left with the linear equation for which we properly use vector notation (z is a column vector with the component z_i, $i = 1, \ldots, n$):

$$\dot{z} = Wz \tag{8}$$

$W = (w_{ij})$ is the matrix of rates. This linear differential equation can be solved by standard numerical techniques (provided n is not too large). Solution curves are of the general form

$$z_i(t) = \sum_j a_{ij} \exp\{\lambda_j t\}$$

with

$$a_{ij} = u_{ij} \sum_k (U^{-1})_{jk} z_k(0) \tag{9}$$

The negative reciprocal time constants λ_j are the eigenvalues of the matrix W. For the sake of simplicity, we consider them as being ordered $\lambda_1 > \lambda_2 > \ldots > \lambda_n$. Similarly, we may assume $w_{11} > w_{22} > \ldots > w_{nn}$. Thus, λ_1 is the largest eigenvalue and w_{11} the largest diagonal element of matrix W. The coefficients u_{ij} build up the eigenvectors of $W(U = \{u_{ij}\}$ and $\Lambda = \{\lambda_{ij} = \lambda_j \delta_{ij}\}$ is a diagonal matrix):

$$WU = U\Lambda \text{ or } \Lambda = U^{-1}WU \tag{10}$$

The solution curves $x_i(t)$ can be obtained from known functions $z_i(t)$ by means of a Bernoulli equation (For further details, the reader is referred to refs. 15 and 24.)

In the following we shall be interested in stationary mutant distributions exclusively. For this purpose, a knowledge of the functions $z_i(t)$ is sufficient. We notice that the problem formulated in internal coordinates is invariant under the transformation (eq. 6):

$$\xi_i(t) = \frac{x_i(t)}{\sum_j x_j(t)} = \frac{z_i(t)}{\sum_j z_j(t)} \tag{11}$$

For the moment we assume $\lambda_1 > \lambda_2$ (we shall come back to the degenerate case $\lambda_1 = \lambda_2$ in section 3). In approaching the steady state, all contributions in eq. 8 except those of the eigenvector corresponding to the largest eigenvalue (principal eigenvector) will vanish:

$$z_i(t) \overset{t \text{ large}}{\rightarrow} a_{i1} \exp(\lambda_1 t)$$

In internal coordinates we find

$$\bar{\xi}_i = \lim_{t \to \infty} \frac{z_i(t)}{\sum_j z_j(t)} = \frac{a_{i1}}{\sum_j a_{j1}} = \frac{u_{i1} \sum_k (U^{-1})_{1k} z_k(0)}{\sum_j \sum_k u_{j1} (U^{-1})_{1k} z_k(0)} \tag{12}$$

Eq. 9 thus describes the selection process mentioned above: out of a superposition of eigenvectors, that with the largest eigenvalue is selected. This eigenvector in our problem (W is a positive definite matrix) is characterized by only positive components, $u_{i1} > 0$ [7].

The largest eigenvalue and the principal eigenvector of W were determined by perturbation theory [24]. To this end, the matrix W is split into its diagonal and off-diagonal elements:

$$W = \{w_{ij}\} = \begin{pmatrix} w_{11} & 0 & \ldots & 0 \\ 0 & w_{22} & \ldots & 0 \\ \vdots & \vdots & & \vdots \\ 0 & 0 & \ldots & w_{nn} \end{pmatrix}$$

$$+ \begin{pmatrix} 0 & w_{12} & \ldots & w_{1n} \\ w_{21} & 0 & \ldots & w_{2n} \\ \vdots & \vdots & & \vdots \\ w_{n1} & w_{n2} & \ldots & 0 \end{pmatrix} = W_0 + W'$$

The diagonal part W_0 is considered as the unperturbed system, W' is the perturbation.

The results are obtained at different orders of the off-diagonal elements $w_{ij}, j \neq i$. For the sake of convenience, we present the sums of all contributions up to a given order, e.g., $\lambda_1^{(2)} = \lambda_1^{(0)} + \Delta\lambda_1^{(1)} + \Delta\lambda_1^{(2)}$, wherein $\Delta\lambda_1^{(k)}$ are the proper contributions of perturbation theory at order k. At zeroth order, off-diagonal elements do not enter the computation explicitly. Hence, we find [8]:

$$\lambda_1^{(0)} = w_{11} \tag{13a}$$

By means of eq. 6, we have $w_{11} = \bar{E}$ at the steady state, from which we derive

$$\xi_1^{(0)} = \frac{w_{11} - \bar{E}_{-1}}{E_1 - \bar{E}_{-1}} = 1 - \frac{A_1(1 - Q_{11})}{E_1 - \bar{E}_{-1}} = \frac{Q_{11} - \sigma_1^{-1}}{1 - \sigma_1^{-1}} \tag{13b}$$

Herein, the mean excess productivity of all sequences except the master sequence I_1 is defined by

$$\bar{E}_{-1} = \frac{\sum_{i=2}^{n} E_i x_i}{\sum_{i=2}^{n} x_i} \tag{14}$$

and the superiority of I_1 by

$$\sigma_1 = \frac{A_1}{D_1 + \bar{E}_{-1}} \tag{15}$$

The evaluation of eq. 13b requires some care, since it is not self-evident how to apply it. The computation of the 'mean except the master' excess productivity $\bar{E}_{-1}$ according to eq. 14 presupposes knowledge of the stationary mutant distribution ($\bar{x}_i$; $i = 2, \ldots, n$). Strictly speaking, these concentrations are not accessible without making explicit use of the off-diagonal elements w_{ij}. There are ways to get out of this problem; we mention two alternatives:

(1) In the phenomenological approach [8,9], we consider the superiority σ to be an empirical parameter which in principle can be determined experimentally through systematic studies on wild type and mutants. Examples are the work on Qß-RNA [6] and a low molecular weight variant of it [16].

(2) Alternatively, we could make some assumptions on the mutant distribution or 'borrow' the corresponding expression for the relative frequencies of mutants form the first-order results (see eq. 16b). From these frequencies we obtain $\bar{E}_{-1}$ easily.

In some of our test cases to be discussed in detail in the forthcoming sections we shall circumvent this problem by assuming equal rate constants for all mutants: $E_2 = E_3 = \ldots E_n = \bar{E}_{-1}$.

In first order we obtain no contribution to the eigenvalue λ_1. The eigenvectors can now be expressed in the form of ratios:

$$\lambda_1^{(1)} = \lambda_1^{(0)} = w_{11} \tag{16a}$$

$$\xi_i^{(1)} = \xi_1^{(1)} \frac{w_{i1}}{w_{11} - w_{ii}}; \; i = 2, \ldots, n \tag{16b}$$

The relative concentration of the master sequence can be obtained from the conservation relation $\Sigma_i \bar{\xi}_i = 1$ [24]:

$$\bar{\xi}_1^{(1)} = \frac{1}{1 + \sum_{i=1} \frac{w_{i1}}{w_{11} - w_{ii}}} = 1 - \sum_{i=1} \frac{w_{i1}}{w_{11} - w_{ii}}$$

Again we may introduce a mean except the master for the selective values w_{ii} and find

$$\bar{\xi}_1^{(1)} = 1 - (w_{11} - \bar{w}_{-1})^{-1} \sum_{i=1} w_{i1} = 1 - \frac{A_1(1 - Q_{11})}{w_{11} - \bar{W}_{-1}} \tag{16c}$$

Eq. 16c is the first-order analogue of eq. 13b. Note that here the excess productivity is replaced by the selective values.

In second order we obtain the following results:

$$\lambda_1^{(2)} = w_{11} + \sum_{i=2}^{n} \frac{w_{i1} w_{1i}}{w_{11} - w_{ii}} \tag{17a}$$

and

$$\bar{\xi}_i^{(2)} = \bar{\xi}_1^{(2)} \left\{ \frac{w_{i1}}{w_{11} - w_{ii}} + \sum_{j=2, j=i}^{n} \frac{w_{ij} w_{ji}}{(w_{11} - w_{ii})(w_{11} - w_{jj})} \right\} \tag{17b}$$

The zeroth-order equation turned out to be particularly useful: Using the condition $\bar{\xi}^{(0)} > 0$, Eigen [8] was able to define an accuracy threshold below which a given sequence is inevitably lost during multiple replication. This error threshold was applied to various replication processes like enzyme-free template-instructed RNA replication [9]. The error limit provides an explanation for naturally occurring genome lengths. Despite the success mentioned above, we found it somewhat unsatisfactorily to rely on zeroth-order perturbation theory, particularly in the present case where rigorous numerical tests are not accessible for reasonably long sequences. The number of possible different sequences of chain length ν equals 4^{ν}, an expression which soon goes to 'superastronomic' numbers with increasing ν. In the following sections we shall analyse the validity of the perturbational approach by means of an appropriate model.

3. An exactly solvable test case ($n = 2$)

For the purpose of illustration we consider the exceedingly simple example of two sequences ($n =$

2). They can be understood as a master sequence and a mutant. This case has the advantage that analytical expressions of the exact solutions are easily available. From the two-dimensional eigenvalue problem, we obtain ($w_{11} > w_{22}$):

$$\lambda_1 = \tfrac{1}{2}\left\{ w_{11} + w_{22} + (w_{11} - w_{22})\sqrt{1 + \frac{4w_{12}w_{21}}{(w_{11} - w_{22})^2}} \right\}$$

and

$$\bar{\xi}_1 = \frac{w_{12}}{\lambda_1 - w_{11} + w_{12}}.$$

Perturbation theory yields the following approximations

$$\lambda_1^{(0)} = w_{11},\ \lambda_1^{(2)} = w_{11} + \frac{w_{12}w_{21}}{w_{11} - w_{22}}$$

for the eigenvalue. We recognize that these expressions follow immediately from an expansion of the square root in the exact eigenvalue

$$\sqrt{1 + \frac{4w_{12}w_{21}}{(w_{11} - w_{22})^2}} = 1 + \frac{2w_{12}w_{21}}{(w_{11} - w_{22})^2} + \dots$$

As expected the results of perturbation theory converge to the exact solution when the quotient $w_{12}w_{21}/(w_{11} - w_{22})^2$ approaches zero.

For the eigenvector, first-order perturbation theory leads to the expression

$$\bar{\xi}_1^{(1)} = \frac{w_{11} - w_{22}}{w_{11} - w_{22} + w_{21}}$$

There is no second-order contribution to $\bar{\xi}_1$ in this simple example.

Perturbation theory is not applicable to the case of equal or almost equal selective values, w_{11} and w_{22}. In this case we obtain for $w_{11} = w_{22}$

$$\lambda_1 = w_{11} = \sqrt{w_{12}\cdot w_{21}} \text{ and } \bar{\xi}_1 = \frac{\sqrt{w_{12}}}{\sqrt{w_{12}} + \sqrt{w_{21}}}$$

from the exact expressions. The two sequences thus are present in relative amounts proportional to the square roots of the off-diagonal $\sqrt{w_{12}}$ and $\sqrt{w_{21}}$. The steady state of the system is determined by the mutation terms exclusively. In the case of equal selective values, the master sequence is replaced by an ensemble of two or eventually more 'masters'. The mutant distribution belonging to such an ensemble can be obtained from perturbation theory of degenerate states [24].

Returning to the general case $w_{11} \neq w_{22}$, we consider concrete examples. For the sake of simplicity we assume the same quality factors for both sequences: $Q_{11} = Q_{22} = Q$. This assumption is well justified in the case where both molecules are replicated by the same molecular machinery which is characterized by a given accuracy of replication. The matrix W, then, is of the form

$$W = \begin{pmatrix} A_1Q - D_1 & A_2(1-Q) \\ A_1(1-Q) & A_2Q - D_2 \end{pmatrix}.$$

For a given set of rate constants A_1, A_2, D_1 and D_2, we calculate eigenvalue and eigenvector as a function of Q. Often it is appropriate to assume equal rate constants of the decomposition process: $D_1 = D_2 = D$. A straightforward computation

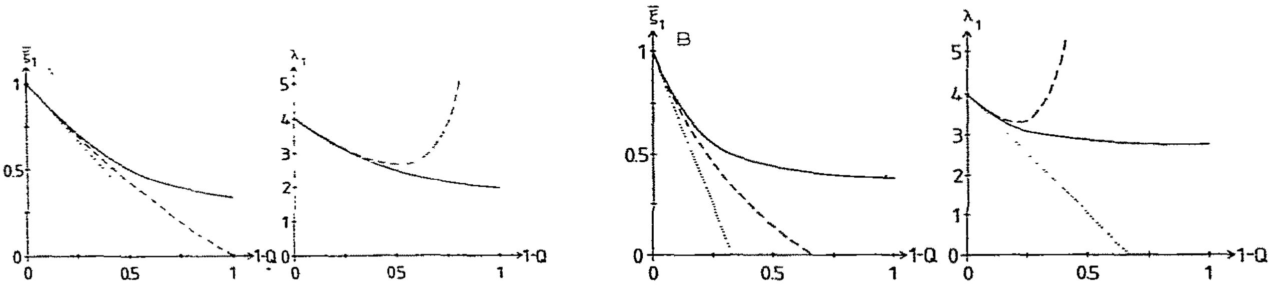

Fig. 1. Largest eigenvalue (λ_1) and eigenvector ($\bar{\xi}_1$) for the two-dimensional case ($n = 2$). We present exact solutions (———) together with the results of zeroth-order (.....) and second-order (- - - - -) perturbation theory. Numerical values chosen are (A) $A_1 = 4$, $A_2 = 1$, $D_1 = D_2 = 0$ and (B) $A_1 = 6$, $A_2 = 3$, $D_1 = 2$ and $D_2 = 1$ in arbitrary reciprocal time and concentration units.

shows that the eigenvector is independent of D in this case.

Graphs for two concrete numerical examples are shown in fig. 1. We observe the expected behaviour of the approximative solutions when Q deviates gradually from 1. The zeroth-order approximation diverges first from the exact solution. At a certain critical Q value, generally denoted by $Q_{\min}$, the component $\tilde{\xi}_1^{(0)}$ of the zeroth-order eigenvector vanishes. This indicates ultimate breakdown of the zeroth-order approach. As we shall see later on $Q_{\min}$ is of great importance for long sequences where the zeroth-order approximation turns out to be very close to the exact solution, provided $Q > Q_{\min}$.

The first- or second-order expressions are better approximations to the exact solutions and remain useful down to smaller values of Q. We note that $\xi_1^{(2)}$ diverges when w_{11} approaches w_{22}. This may happen at $Q \to 0$ in the case $D_1 = D_2$ (fig. 1a) or at some finite Q value in the case $D_1 > D_2$ (fig. 1b).

We would like to draw the reader's attention to a treatment of the wild type and mutant of bacteriophage Qß by means of a two-dimensional model system [1].

4. A simplifying model for replication with errors

In order to conceive a model for polynucleotide replication with errors we have to consider the nature of mutations in some detail. The available experimental data (for a recent kinetic study on RNA replication, see ref. 3) establish that template-induced replication proceeds digit per digit from the 5′ end to the 3′ end of the newly synthesized molecule. Point mutations are single-digit errors of the replication process. For the sake of simplicity we dispense with a discussion of other, usually rare sources of replication errors except point mutations. But, in principle, deletions or insertions occur as well. They could be incorporated into a modified model. Accordingly, we can assign a single-digit accuracy to every propagation step of the growing chain. In the case of a polymer with ν segments the quality factor Q can be written down as a product of ν individual factors (q_i; $i = 1, \ldots, \nu$). This assumption does not necessarily mean that a single-digit accuracy of the replication process is independent of the preceding base-pairs. The influence of these base-pairs is implicitly contained in the corresponding, position-dependent q factor. For the correct replication of a given polynucleotide, e.g., I_k, we obtain:

$$Q_{kk} = q_1^{(k)} \cdot q_2^{(k)} \ldots q_\nu^{(k)} = \bar{q}_k^\nu \tag{18}$$

where q_1 is the single-digit accuracy of the incorporation of the first base, q_2 that of the incorporation of the second base, etc. This single-digit accuracy depends on the nature of the base to be incorporated, on the preceding base-pairs, and on the mechanism of replication as well as on environmental factors. Following eq. 18, we define a mean single-digit accuracy $\bar{q}_k$ which accounts implicitly for all these influences, and which is characteristic for a given sequence I_k. For long enough, naturally occurring, particularly non-repetitive polynucleotides with similar mean base compositions, these mean single-digit accuracies will mainly depend on the mechanisms of replication, since specific neighbour effects cancel out in long sequences.

With slight modifications, eq. 18 is also useful as a quantitative measure for the frequencies of the various point mutations. To this end, we have to assign single-digit mutation frequencies to the different base exchange processes. Proceeding systematically we distinguish three classes of base exchanges: (1) purine-purine and pyrimidine-pyrimidine, an AU pair is replaced by a GC pair and vice versa; (2) intrapair exchange, A is replaced by U or G is replaced by C and vice versa; and finally (3) interpair purine-pyrimidine exchange, A is replaced by C, U by G, G by U or C by A. All possibilities are summarized in table 1. Again we can account implicitly for the influences of the preceding base-pairs on the mutation frequencies and assign position-dependent factors for base exchange (table 1 and fig. 2). We denote the three mutation frequencies at position 'i' of the polynucleotide I_k by $\alpha_i^{(k)}$, $\beta_i^{(k)}$ and $\gamma_i^{(k)}$. In absence of insertions and deletions we have the conservation relation

$$q_i^{(k)} + \alpha_i^{(k)} + \beta_i^{(k)} + \gamma_i^{(k)} = 1 \tag{19}$$

which simply expresses the fact that we have one

Table 1

Systematics of point mutations

Base in the original sequence	Class of point mutation					
	Purine-purine pyrimidine-pyrimidine exchange		Intrapair exchange		Interpair purine-pyrimidine Pyrimidine-purine exchange	
	Base in the mutant	Frequency	Base in the mutant	Frequency	Base in the mutant	Frequency
G	A	$\alpha^{(G)}$	C	$\beta^{(G)}$	U	$\gamma^{(G)}$
A	G	$\alpha^{(A)}$	U	$\beta^{(A)}$	C	$\gamma^{(A)}$
C	U	$\alpha^{(C)}$	G	$\beta^{(C)}$	A	$\gamma^{(C)}$
U	C	$\alpha^{(U)}$	A	$\beta^{(U)}$	G	$\gamma^{(U)}$

of the four bases at position i in the replica. For these four possibilities, the correct copy and the three mutations at position i, $I_k \to I_l$, $I_k \to I_m$, $I_k \to I_p$, we thus obtain the following four elements of the Q matrix:

$$I_k \to I_k : Q_{kk} = q_1^{(k)} \cdot q_2^{(k)} \ldots q_i^{(k)} \ldots q_\nu^{(k)} \qquad (18)$$

$$I_k \to I_l : Q_{lk} = q_1^{(k)} \cdot q_2^{(k)} \ldots \alpha_i^{(k)} \ldots q_\nu^{(k)} \qquad (20a)$$

$$I_k \to I_m : Q_{mk} = q_1^{(k)} \cdot q_2^{(k)} \ldots \beta_i^{(k)} \ldots q_\nu^{(k)}. \qquad (20b)$$

$$I_k \to I_p : Q_{pk} = q_1^{(k)} \cdot q_2^{(k)} \ldots \gamma_i^{(k)} \ldots q_\nu^{(k)} \qquad (20c)$$

All these expressions are exact insofar as neighbouring effects are taken into account implicitly. As we see from fig. 2, the pattern of possible mutations becomes exceedingly complicated for larger values of ν and hence we have to look for physically meaningful simplifications.

Systematic studies on base replacements in bacteriophage RNA replication [2,5,16] revealed

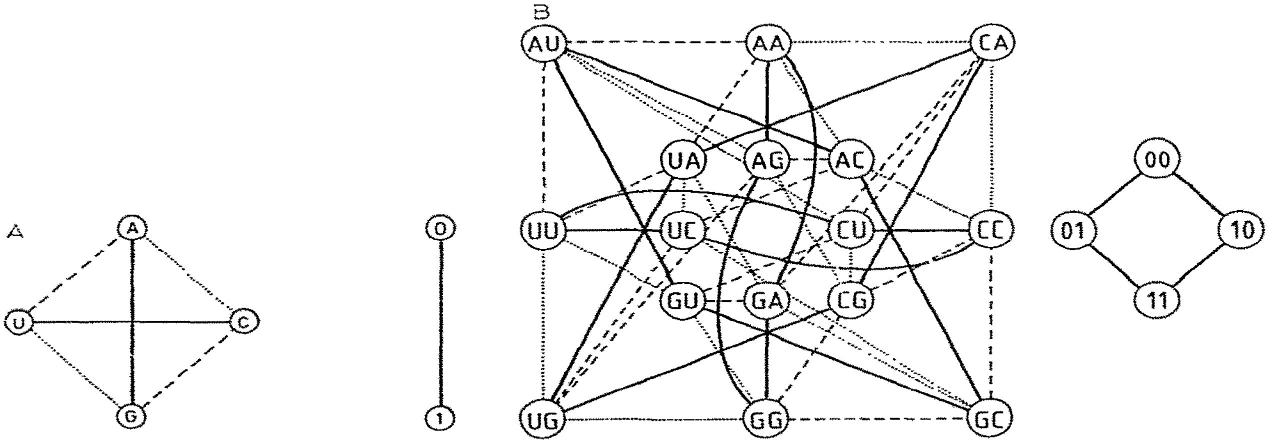

Fig. 2. Systematics of point mutations. We distinguish three classes of base exchange processes: (1) purine-purine, pyrimidine-pyrimidine exchange (————), (2) intrapair exchange (— — —) and (3) interpair purine-pyrimidine exchange (.....). We present the two most simple examples, $\nu = 1$ (A) and $\nu = 2$ (B). In the case where we restrict our analysis to base exchange processes of one class only, the set of 4^ν different sequences can be partitioned into 2^ν disjoint subsets of 2^ν elements. The 2^ν elements form a ν cube – for the sake of simplicity we use 0 and 1 as digits after restriction. In the case of $\nu = 1$ we obtain a straight line (A), in the case of $\nu = 2$ a square (B). The rationale behind this simplification is to be seen in the experimental fact that base exchange processes of class 1 occur much more readily than those of the other two classes (cf. table 1).

that purine-purine, pyrimidine-pyrimidine exchange is the most frequently occurring type of point mutations, at least in the particular cases investigated. The physics behind this finding seems to lie in the relative stability of the GU or UG wobble pairs. Presumably, this preference is very generai and not an idiosyncrasy of some peripheral bacteriophages. The other two classes of point mutations then are rare events and we have:

$$\beta, \gamma \ll \alpha = 1 - q \qquad (21)$$

Our simplifying model is based on the validity of eq. 21 and the 'averaging out' of neighbouring effects in long sequences. We assume the same mean single-digit accuracy for all polynucleotides under consideration, $\bar{q}_k = q$, and we restrict our model to mutations of class 1: $\alpha = 1 - q$. Then, the diagram of possible mutations reduces to a set of equivalent cubes of dimension ν (figs. 2 and 3; these 'ν cubes' are straight lines for $\nu = 1$, squares for $\nu = 2$, cubes for $\nu = 3$, etc.). The Q matrix is of a fairly simple form, since the frequency factors Q_{jk} for mutations of this class, $I_k \to I_j$, depend on the Hamming distances of the two sequences, D_{jk}, only. The Hamming distance is the smallest number of base exchanges which convert a given sequence (I_k) into another (I_j):

$$I_k \to I_j\colon\ D_{jk} = d;\ Q_{jk} = q^{\nu-d}(1-q)^d \qquad (22)$$

The rare mutations which we have excluded by means of eq. 21 can be visualized easily by jumps from a given ν cube to an equivalent one within the complete mutation diagram (fig. 2; to give examples such jumps lead from the A-G line to the U-C line in the $\nu = 1$ case or from the (AA,AG,GG,GA) square to one of the three other squares, e.g., the (UA,UG,CG,CA) square for $\nu =$ 2).

5. Some low-dimensional test cases ($\nu = 3$)

The system with $\nu = 3$ consists of eight different sequences ($n = 8$): the scheme of mutations is shown in fig. 3. The Q matrix is mapped onto a '3 cube', every edge of the cube corresponds to a factor $1 - q$. The matrix element Q_{ij} thus is of the form $q^{3-d}(1-q)^d$ where d is the minimum num-

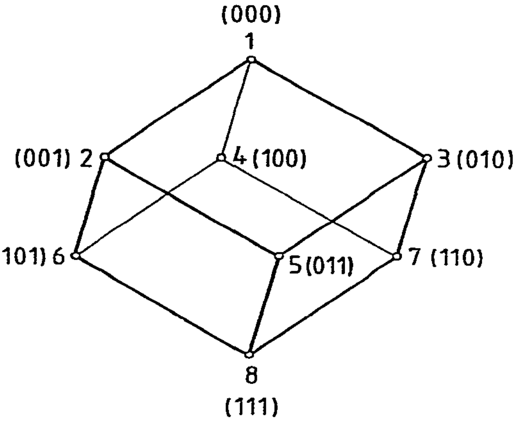

Fig. 3. The eight sequences of the two-digit system with $\nu = 3$.

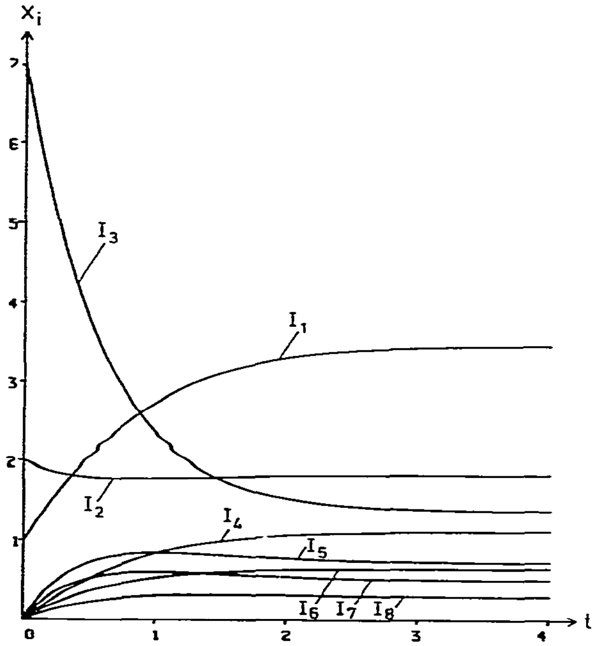

Fig. 4. Numerical integration of eq. 1. the rate constants chosen are: $A_1 = 5$, $A_2 = 3$, $A_3 = 2$, $A_4 = \ldots = A_8 = 1.1$ and $D_1 = \ldots = D_8 = 1.0$ in arbitrary reciprocal time and concentration units. The single-digit accuracy of replication was $q = 0.8$. Initial concentrations: $x_1(0) = 1$, $x_2(0) = 2$, $x_3(0) = 7$ and $x_4(0) = \ldots = x_8(0) = 0$ were applied.

Table 2

The steady state of the system described in fig. 4 ($\nu = 3$; $q = 0.8$; the A_i values are 5,3,2,1.1,1.1,1.1,1.1 and 1.1 for $i = 1, \ldots, 8$, respectively)

We present the exact eigenvector together with the results of first- and second-order perturbation theory according to eqs. 16b and 17b.

$\bar{\xi}_i$	Perturbation theory		Exact values
	First order	Second order	
$i = 1$	0.381	0.305	0.347
2	0.238	0.209	0.184
3	0.159	0.144	0.139
4	0.122	0.115	0.113
5	0.031	0.079	0.073
6	0.031	0.070	0.064
7	0.031	0.051	0.051
8	0.008	0.026	0.031

ber of edges separating the two corners corresponding to the sequences I_i and I_j. In order to make it easier to follow the forthcoming discussion it is worthwhile to present the matrix W for this example (as introduced in fig. 2 the two digits are denoted by 0 and 1):

$$\begin{array}{ll|llllllll}
 & W & 000 & 100 & 010 & 001 & 110 & 101 & 011 & 111 \\
 & & 1 & 2 & 3 & 4 & 5 & 6 & 7 & 8 \\
\hline
000 & 1 & A_1q^3 & A_2q^2(1-q) & A_3q^2(1-q) & A_4q^2(1-q) & A_5q(1-q)^2 & A_6q(1-q)^2 & A_7q(1-q)^2 & A_8(1-q)^3 \\
100 & 2 & A_1q^2(1-q) & A_2q^3 & A_3q(1-q)^2 & A_4q(1-q)^2 & A_5q^2(1-q) & A_6q^2(1-q) & A_7(1-q)^3 & A_8q(1-q)^2 \\
010 & 3 & A_1q^2(1-q) & A_2q(1-q)^2 & A_3q^3 & A_4q(1-q)^2 & A_5q^2(1-q) & A_6(1-q)^3 & A_7q^2(1-q) & A_8q(1-q)^2 \\
001 & 4 & A_1q^2(1-q) & A_2q(1-q)^2 & A_3q(1-q)^2 & A_4q^3 & A_5(1-q)^3 & A_6q^2(1-q) & A_7q^2(1-q) & A_8q(1-q)^2 \\
110 & 5 & A_1q(1-q)^2 & A_2q^2(1-q) & A_3q^2(1-q) & A_4(1-q)^3 & A_5q^3 & A_6q(1-q)^2 & A_7q(1-q)^2 & A_8q^2(1-q) \\
101 & 6 & A_1q(1-q)^2 & A_2q^2(1-q) & A_3(1-q)^3 & A_4q^2(1-q) & A_5q(1-q)^2 & A_6q^3 & A_7q(1-q)^2 & A_8q^2(1-q) \\
011 & 7 & A_1q(1-q)^2 & A_2(1-q)^3 & A_3q^2(1-q) & A_4q^2(1-q) & A_5q(1-q)^2 & A_6q(1-q)^2 & A_7q^3 & A_8q^2(1\text{-}q) \\
111 & 8 & A_1(1-q)^3 & A_2q(1-q)^2 & A_3q(1-q)^2 & A_4q(1-q)^2 & A_5q^2(1-q) & A_6q^2(1-q) & A_7q^2(1-q) & A_8q^3
\end{array}$$

At first, we consider an integration of the differential equation (eq. 1) with a properly chosen set of numerical values for rate constants and initial concentrations. The resulting solution curves are presented in fig. 4. It takes about three time units for the system to reach the steady state. A comparison of the stationary eigenvectors obtained by first- and second-order perturbation theory and by exact calculation is shown in table 2. We realize the expected improvement of the first-order results by the second-order terms. This particular case study provides also an opportunity to check the reliability of the zeroth-order approximation. As we have outlined before, a computation of the superiority (σ_1) of the master sequence (I_1), if it is not determined experimentally, requires a knowledge of the quasi-species distribution. For the purpose of comparison, we may use the three eigenvectors summarized in table 2 and obtain from first-order perturbation theory

$\sigma_1 = 2.424$ and $\bar{\xi}_1^{(0)} = 0.169$

from second-order perturbation theory

$\sigma_1 = 2.690$ and $\bar{\xi}_1^{(0)} = 0.233$ and

from the exact quasi-species distribution

$\sigma_1 = 2.739$ and $\bar{\xi}_1^{(0)} = 0.231$.

It is somewhat satisfactory to notice that the use of the exact results leads to that value of $\bar{\xi}_1^{(0)}$ which is closest to the correct numerical value ($\bar{\xi}_1$).

In the next example we consider the Q dependence of the steady state. In order to avoid any difficulty in the computation of the superiority σ_1, we assume equal rate constants for all mutants ($A_2 = A_3 = \ldots, = A_8 = 1$). All decomposition rate constants were chosen to be equal as well and hence do not enter into the calculation of the mutant distribution. Without losing generality, we put $D_1 = D_2 = \ldots = D_8 = D = 0$. For $A_1 = 4$, we thus have a superiority $\sigma_1 = 4$ independently of the final mutant distribution. The results are shown in

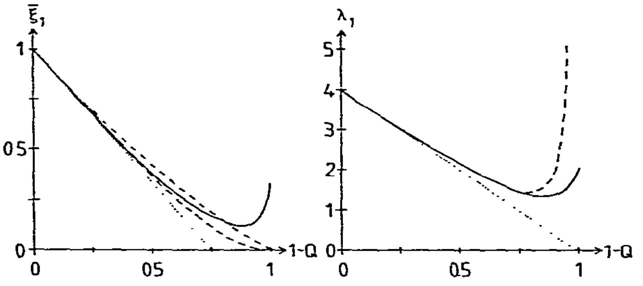

Fig. 5. Largest eigenvalue (λ_1) and the coefficient ($\bar{\xi}_1$) of the corresponding eigenvector for the eight-dimensional case ($\nu = 3$, $n = 8$). We present exact solutions (———) together with the results of zeroth-order (·····), first-order (— · — · —) and second-order (-----) perturbation theory. Numerical values chosen are: $A_1 = 4$, $A_2 = \ldots = A_8 = 1$ and $D_1 = \ldots = D_8 = 0$ in arbitrary reciprocal time and concentration units.

fig. 5. We notice subsequent deviations of the zeroth- (first) and second-order perturbational approach from the exact curves for eigenvalues and eigenvectors. As expected the second-order approximation to the eigenvalue diverges for $Q \to 0$. We observe also that the exact solutions for $\bar{\xi}_1$ and λ_1 pass through a minimum and increase again when $Q \to 0$. We shall draw our attention to this fact on the basis of the next example.

Now we consider the same dependence of the eigenvector on the accuracy of the replication process but we take a slightly different point of view. We plot our results as functions of the single-digit accuracy q (fig. 6). In principle, the change from Q to $q = \sqrt[3]{Q}$ means only a scaling of the abscissa. We shall, however, consider the physical interpre-

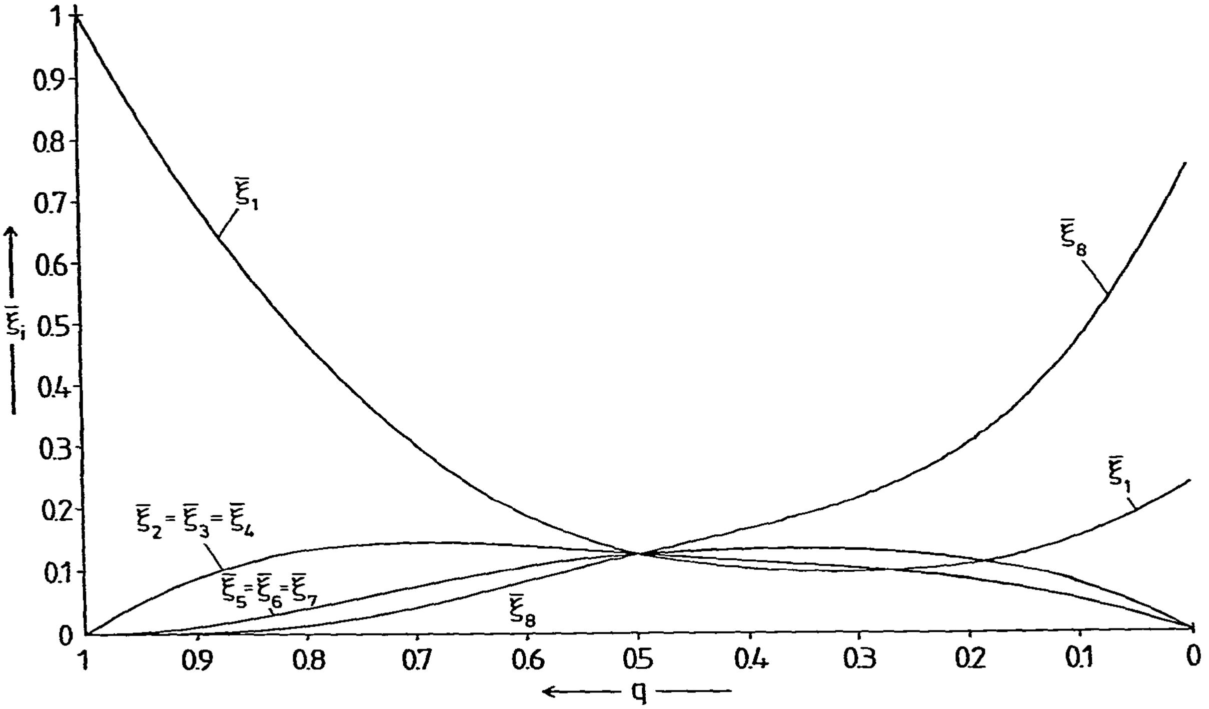

Fig. 6. The mutant distribution in the quasi-species ($\nu = 3$). The eigenvector ($\bar{\xi}_i$; $i = 1, \ldots, 8$) corresponding to the largest eigenvalue (λ_1) is presented as a function of the single-digit accuracy q. Note that all sequences are present in equal amounts at the point of stochastic replication ($q = 0.5$). For $q < 0.5$ we observe complementary replication: sequences are selected in pairs (I_1, I_8), (I_2, I_7), (I_3, I_6) and (I_4, I_5). These pairs are complementary sequences (cf. fig. 3). The numerical values chosen are: $A_1 = 10$, $A_2 = \ldots = A_8 = 1$ and all D values equal.

tation of the q dependence. At $q = 1$ we have accurate replication. Every digit is duplicated with ultimate precision. Single-digit accuracies $q < 1$ imply a certain frequency of errors. Once in a while the wrong digit is incorporated during the replication process.

Let us consider now the case $q = 0.5$. Correct and wrong digits are incorporated with equal frequencies. In this case we may speak of statistical replication, since there is no direction of the replication process by the template: all sequences, the correct copy and the mutants are formed with equal probabilities. Accordingly, even the most efficient template – the master sequence I_1 at large enough q values – does not benefit from its larger rate constant (A_1). Hence, as we see in fig. 6, all eight sequences are present in equal concentrations at $q = 0.5$.

Single-digit accuracies $q < 0.5$ mean that wrong digits are incorporated with higher frequencies than the correct ones. Thus, we find another kind of regularity with an altered logic of replication. At $q = 0$ we encounter again a completely determined situation: digits alternate, every '0' is replaced in the copy by a '1' and vice versa. What we observe is replication by means of complementary strands. This kind of a replication process has been studied in some detail before [8] (see also ref. 9). The matrix W at $q = 0$ can be factorized into four 2×2 diagonal blocks, each combining two complementary sequences, I_+ and I_-. All these blocks are of the same form:

	I_+	I_-
I_+	0	A_-
I_-	A_+	0

The positive eigenvalue and the corresponding eigenvector of this 2×2 matrix are $\lambda = \sqrt{A_+ A_-}$ and $\xi_+ = \sqrt{A_-} / \sqrt{A_+} + \sqrt{A_-}$. In our example we are dealing with four plus-minus ensembles: (000,111), (100,011), (010,101) and (001,110). These four plus-minus ensembles compete. The subsystem with the largest eigenvalue contains the plus-minus ensemble which is selected. In fig. 6 this is the combination (000,111). Note that the two sequences are present in the ratio $1/\sqrt{10}$ at $q = 0$, since we applied the following rate constants: $A_1 = 10$, $A_2 = A_3 = \ldots A_8 = 1$. In the case of two-digit replications we are in a position to describe direct and plus-minus replication by the same general type of equation. In terms of single-digit accuracies the former mechanism is restricted to the domain $1 \geq q > 0.5$. The latter case can be characterized by $0 \leq q < 0.5$. Finally, we would like to mention one point: the nice property eq. 1 receives through the incorporation of eq. 22, namely the ability to describe both, direct and plus-minus replication, holds only for true two-digit systems. We can apply the equation in both ranges to approximative treatments of the biological four-digit-(G,A,C,U) system, e.g., as we did by means of eq. 21 in the domain $1 \geq q > 0.5$. Then, the nature of the approximation, however, has to be different: from eq. 21 follows that A is the most likely alternative to G in mutations, whereas G, C complementarity is the basis for plus-minus replication in the range $0 \leq q < 0.5$.

6. A model for cases of higher dimensions ($\nu > 3$)

In section 4 we made an attempt to simplify the analysis of eq. 1 by introducing a two-digit system. However, the number of possible sequences, $n = 2^\nu$, is still restrictive for a complete analysis of higher-dimensional cases. Nevertheless, we are looking for a test of the results from perturbation theory for longer sequences. Consequently, we need another kind of simplification which leads to a drastic reduction of the dimension of the eigenvalue problem to be solved. For this goal we form classes of sequences within a quasi-species distribution. These classes are defined by means of the Hamming distances between the master sequence and the sequence under consideration. Class 0 contains exclusively the master sequence, class 1 all ν one-error mutants, class 2 all $\binom{\nu}{2}$ two-error mutants, etc. In general, we have all $\binom{\nu}{k}$ k-error mutants in class k. Now we make two assumptions concerning the rate constants in order to prepare the system for a reduction of the 2^ν-dimensional differential equation for individual sequences to a $(\nu + 1)$-dimensional equation for individual classes: (1) all rate constants for the degradation process are assumed to be equal $D_1 = D_2 = \ldots = D_{2^\nu} = D$

and, hence, have no influence on the stationary mutant distribution; and (2) all formation rate constants are assumed to be equal within a given class, i.e., for

class 0, $A_0 = A'_0$; for
class 1, $A_1 = A_2 = \dots = A_\nu = A'_1$; for
class 2, $A_{\nu+1} = A_{\nu+2} = \dots$
$A_{\nu+\binom{\nu}{2}} = A'_2$; and, in general, for
class k, $A_{\binom{\nu}{1}+\binom{\nu}{2}+\dots+\binom{\nu}{k-1}+1} = \dots =$
$A_{\binom{\nu}{1}+\binom{\nu}{2}+\dots+\binom{\nu}{k}} = A'_k.$

New variables y_i are introduced for the concentrations of classes:

$$y_0 = \xi_0,\, y_1 = \sum_{i=1}^{\nu} \xi_i, \dots,\, y_k = \sum_{i=\binom{\nu}{1}+\binom{\nu}{2}+\dots+\binom{\nu}{k-1}}^{i=\binom{\nu}{1}+\binom{\nu}{2}+\dots+\binom{\nu}{k}} \xi_i \qquad (23)$$

Finally, there remains the calculations of the elements of the matrix Q' which describes mutations from an error class into another. A lengthy but straightforward calculation [23] yields for mutations from class l into class k:

$$Q'_{kl} = \sum_{j=0}^{m} q^{\nu-2j-|l-k|} \cdot (1-q)^{2j-|l-k|} \cdot \binom{\nu-l}{j+\frac{1}{2}(|l-k|-(l-k))} \binom{l}{j+\frac{1}{2}(|l-k|+(l-k))} \qquad (24)$$

where $m = [\frac{1}{2}(\min\{l+k, 2\nu-(l+k)\} - |l-k|)]$. In cases where the expression in square brackets is a half integer, the summation index j runs to the next smaller integer.

From eq. 24 we obtain the following expression for the diagonal elements of the matrix Q':

$$Q'_{kk} = \sum_{j=0}^{m} q^{\nu-2j}(1-q)^{2j} \binom{\nu-k}{j} \binom{k}{j} \qquad (25)$$

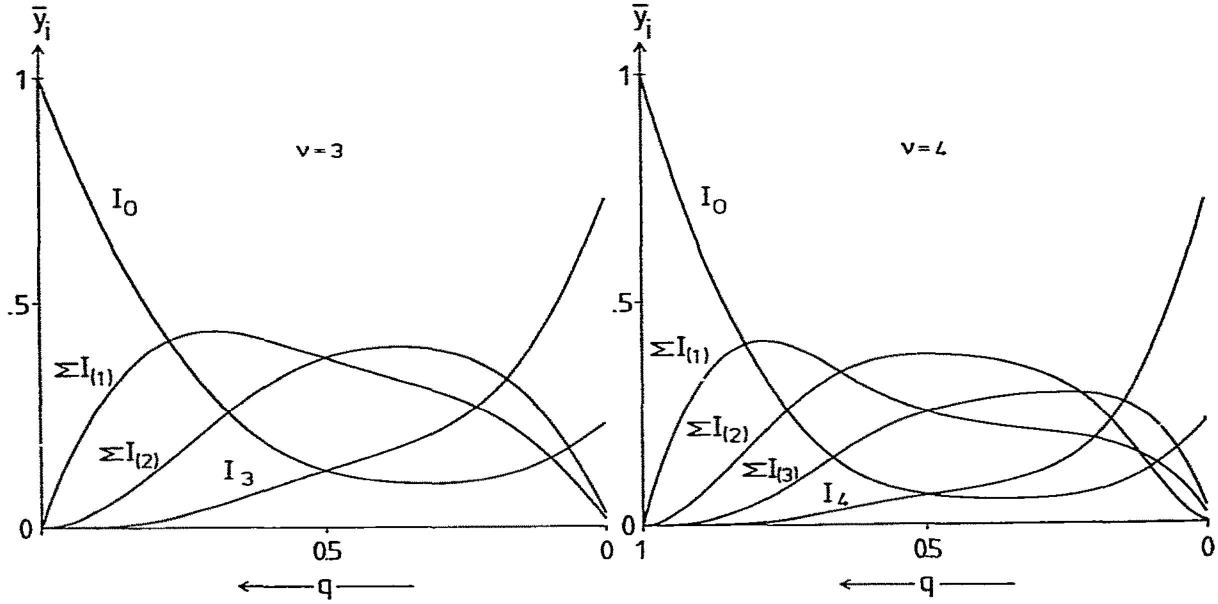

Fig. 7. Distribution of mutant classes as a function of single-digit accuracy q ($\nu = 3$ and 4). By $\Sigma I_{(m)}$ we denote the sum of all m-error mutants of the master sequence I_0. These are all mutants which are characterized by a Hamming distance $D = m$ from the master sequence. The corresponding sum of concentrations is denoted by $\bar{y}_m$. At the point of stochastic replication ($q = 0.5$) all sequences are present in equal amounts. The distribution of mutant classes then is given by the binomial coefficients: $\bar{y}_m = \binom{\nu}{m} \cdot 2^{-\nu}$. Numerical values chosen are the same as in fig. 6 ($A_0 = 10$; $A_i = 1$, $i \neq 0$, all D_i equal).

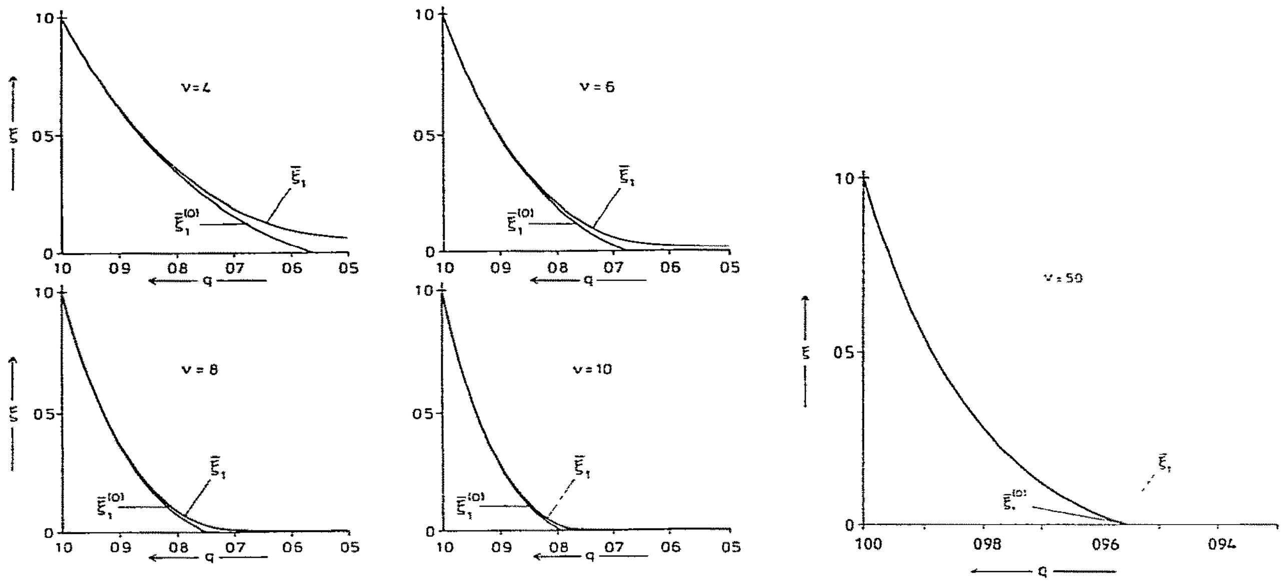

Fig. 8. The range of validity of perturbation theory in the calculation of quasi-species distributions as a function of the chain length ν. We present the coefficient of the master sequence in the stationary distribution: $\bar{\xi}_1$ is the exact solution (upper curve), $\bar{\xi}_1^{(0)}$ the zeroth-order approximation (lower curve). Note that the agreement gets better and better the longer the sequences are. For $\nu = 50$ both curves coincide for practical purposes. Numerical values chosen are the same as in fig. 6 ($A_0 = 10$; $A_i = 1$, $i \neq 0$; all D_i equal).

with $m = \min(k, \nu - k)$.

The differential equation we have to study now is analogous to eq. 5, after insertion of eq. 2:

$$\dot{y}_i = y_i\left(A'_i Q'_{ii} - D - \bar{E}\right) + \sum_{j \neq i} A'_j Q'_{ij};\ i, j = 0, 1, \ldots, \nu \qquad (26)$$

The obvious difference between eqs. 5 and 26 is to be seen in the structures of the matrices Q and Q': Q is symmetric, $Q_{ij} = Q_{ji}$, since the statistical probabilities of mutations and corresponding backward mutations are equal. The same rationale does not hold for classes of mutants

$$Q'_{ij} = Q'_{ji}.$$

In fig. 7 we present the distribution of the mutant classes for systems with $\nu = 3$ and $\nu = 4$ as functions of the single-digit accuracy q. At $q = 0.5$ the concentrations of the individual sequences become equal; since we have stochastic replication $q = 1 - q$, incorporation of correct and complementary digit occurs with the same probability. Then, the frequencies of the mutant classes are simply given by binomial coefficients. For the stationary concentration of the mutant class j we find

$$\bar{y}_j = \frac{1}{2^\nu}\binom{\nu}{j};\ j = 0, 1, \ldots, \nu + 1;\ q = 0.5 \qquad (27)$$

For $q < 0.5$ we observe complementary replication: a given sequence is selected together with the complementary strand.

The change in variables – from concentrations of individual sequences to concentrations of whole mutant classes – reduced the number of variables from 2^ν to $\nu + 1$ and we are now in a position to test the range of validity of the approximations derived by perturbation theory. The results obtained can be subsumed under two important statements:

(1) the longer the polynucleotide sequence is, the better is the agreement between perturbation theory and exact results (fig. 8). For long enough sequences, $\nu > 10$, the zeroth-order approximation

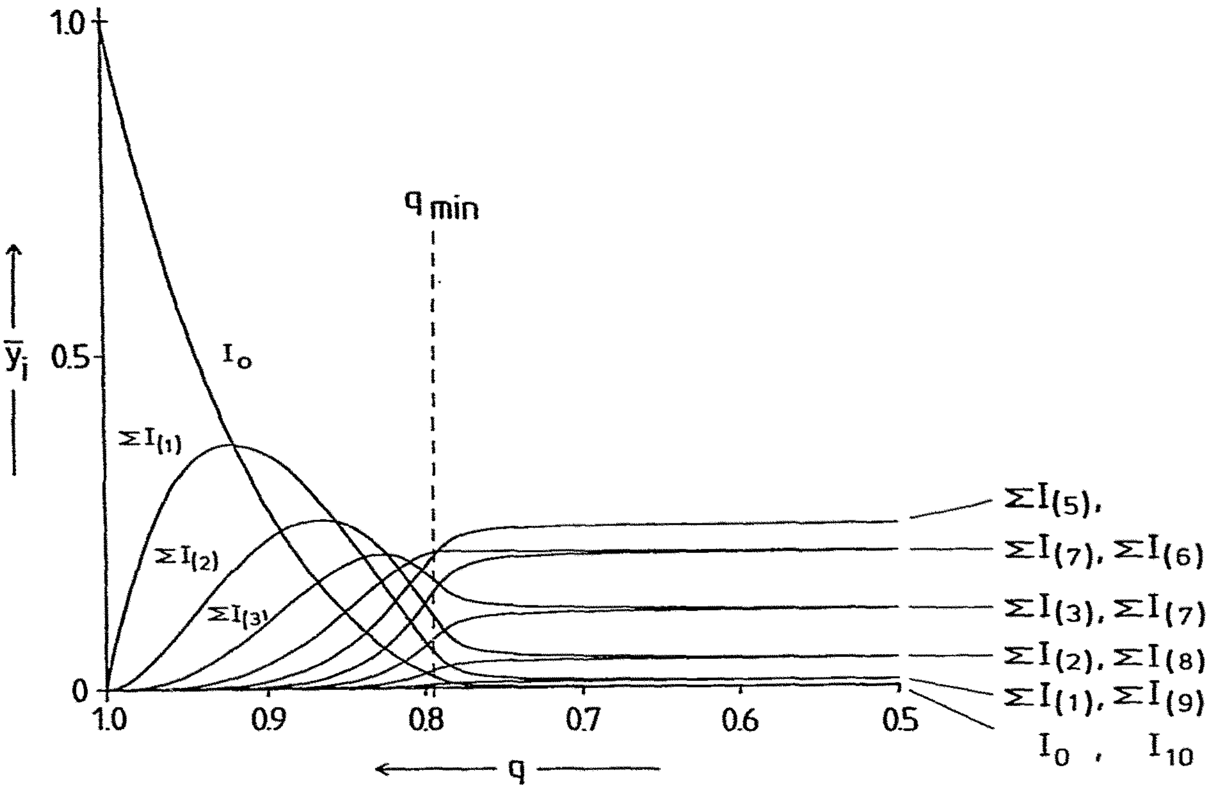

Fig. 9. Distribution of mutant classes as a function of the single-digit accuracy q for $\nu = 10$. Note the transition from direct to stochastic replication around q_{min}. In comparison to fig. 7 we realize the existence of a broad domain of stochastic replication which was restricted to the point $q = 0.5$ for the short sequences ($\nu = 3,4$). For basic definitions and numerical values see fig. 7.

practically coincides with the exact solution as long as the replication process is accurate enough. What 'accurate enough' means can be defined precisely by means of the zeroth-order approximation to the eigenvector of the largest eigenvalue (eq. 13b). A non-vanishing concentration of the master sequence (I_0) requires:

$$\bar{\xi}_0^{(0)} > 0 \rightarrow Q_{00} > Q_{\text{min}} = q_{\text{min}}^{\nu} = \sigma_0^{-1}\ *$$

The range of validity of the analysis by perturbation theory thus is given by $q > q_{\text{min}} = (\sigma_0)^{-1/\nu}$.

(2) The range of q values where stochastic replication occurs spreads enormously with increasing oligomer length ν. For $\nu = 3$ and $\nu = 4$ (fig. 7) we observe a uniform distribution, i.e., equal concentration of sequences, exclusively at the value $q = 0.5$ (In these two cases, the critical single-digit accuracies are $q_{\text{min}} = 0.464$ for $\nu = 3$ and $q_{min} = 0.562$ for $\nu = 4$; σ_0 was chosen to be 10 in all examples presented in figs. 7–10.) For $\nu = 10$ (fig. 9), the existence of a stable quasi-species is confined to the range $1 > q > q_{\text{min}} = 0.794$. Stochastic replication is observed for single-digit accuracies $q < 0.7$, since the concentrations of all sequences are equal for practical purposes beyond this value. For longer sequences, we show plots for $\nu = 50$ in fig. 10, the transition from the range of stable mutant distributions to stochastic replication sharpens enormously. We find a uniform se-

* For the sake of convenience, we use here '0' instead of '1' to denote the master sequence.

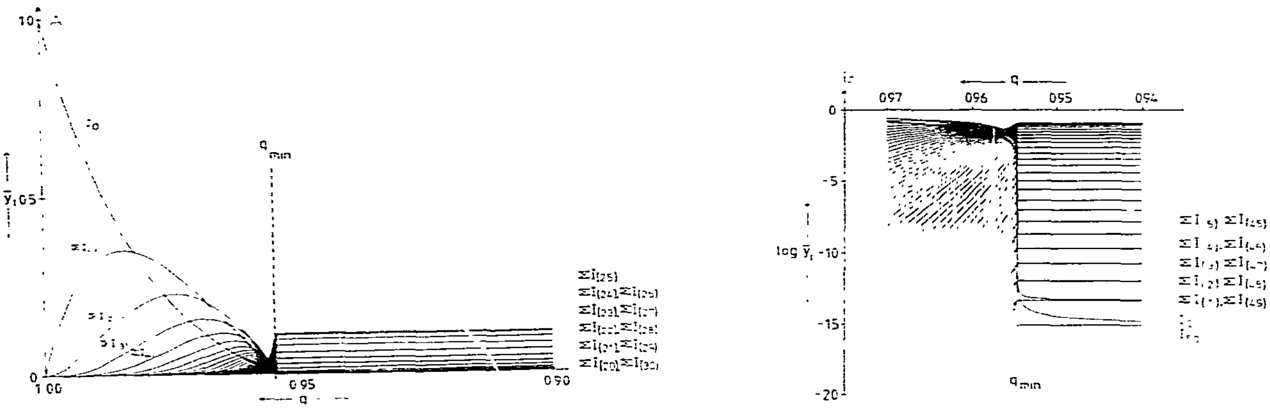

Fig. 10. Distribution of mutant classes as a function of the single-digit accuracy q for $\nu = 50$. Note the sharpness of the transition from direct to stochastic replication around q_{min}. This is seen best on the logarithmic plot. In the domain of stochastic replication individual concentrations become exceedingly small: $\bar{\xi}_i = 8.9 \times 10^{-16}$, $i = 0, \ldots, 2^{50} - 1$. For basic definitions and numerical values see fig. 7.

quence distribution at single-digit accuracies $q < q_{min}$ very close to the critical value already (The relative concentrations of individual sequences then become exceedingly small, $\bar{\xi}_i \approx 2^{-50} = 8.9 \times 10^{-16}$.) For polymer sequences ($\nu > 50$) this transition sharpens further (we extended our calculation up to the size of tRNA-like molecules with $\nu = 80$). The concept of a sharp error threshold as championed by Eigen and Schuster [9] is thus well justified.

The numerical finding that perturbation theory and exact solution yield converging results for increasing chain lengths ν in the range of accurate enough replication can be confirmed analytically [23]. The relative importance of second-order contributions to the eigenvalue vanishes for large ν and $w_{11} > \bar{E}_{-1}$:

$$\lim_{\nu \to \infty} \frac{\Delta\lambda_1^{(2)}}{\lambda_1^{(0)}} = 0.$$

Inspection of higher-order terms leads to the strong conjecture that these contributions vanish as well and, thus, the zeroth-order result coincides with the exact solution.

7. Conclusion

The model we applied to polynucleotide replication is based essentially on two assumptions: (1) replication errors are restricted to point mutations which are described by a two-digit model, and (2) the kinetic parameters of all sequences in a certain mutation class are assumed to be equal. Both assumptions, at first glance, may appear somewhat unrealistic. The first assumption can be justified more easily because of the unequal frequencies for different types of base exchange processes (see the data on bacteriophage Qß quoted in section 4; for the frequencies of point mutations in higher organisms and their interpretation by information theory see, e.g., ref. 25). The second assumption is more restrictive: Not all one-error mutants will have the same kinetic properties and the same will be the case for the sequences in the other mutant classes. When we study established populations, as we do here, and dispense with a consideration of

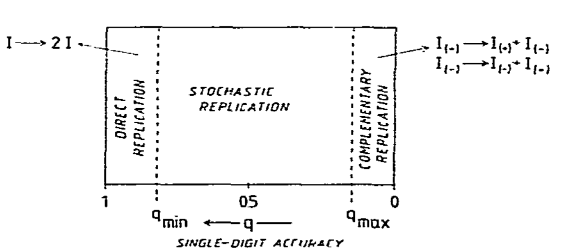
I ⟶ 2I
DIRECT REPLICATION
STOCHASTIC REPLICATION
COMPLEMENTARY REPLICATION
$I_{(+)} \to I_{(+)} + I_{(-)}$
$I_{(-)} \to I_{(-)} + I_{(+)}$
1
05
0
q_{min} ⟵ q ⟶ q_{max}
SINGLE-DIGIT ACCURACY

Fig. 11. Replication as a function of single-digit accuracy (schematic).

changes in the environment, the distribution of kinetic parameters within the mutant classes, however, has little influence on general results only. Through these two assumptions we are now able to handle up to 2^{80} and more individual sequences explicitly.

The model enabled us to derive three general results for the kinetics of self-replication with errors:

(1) In the range of sufficient replication accuracy, i.e., the range in which stable mutant distributions–quasi-species–are found, (zeroth-order) perturbation theory predicts the exact distributions of polynucleotide sequences ($\nu > 20$) very well. The differences between approximate and exact solutions decrease with increasing chain length.

(2) The kinetic model applied described both direct and complementary (plus-minus) replication in different domains of the single-digit accuracy q (fig. 11). Above the critical accuracy, $1 \geq q > q_{\min}$, we have direct replication and formation of a quasi-species. Below another critical value of q which we denote here by $q_{\max}$, i.e., in the domain $q_{\max} > q \geq 0$, we observe complementary replication and formation of a kind of quasi-species the elements of which are plus-minus pairs of polynucleotide sequences.

(3) the domains of direct and complementary replication are separated by an intermediate range of stochastic replication, $q_{\min} > q > q_{\max}$. In this range all sequences have equal probabilities. Replication is not accurate enough in order to sustain faithful transfer of sequences from one generation to the next. The kinetic parameters have no influence on the stationary mutant distribution. For very short sequences ($\nu \leq 4$) this intermediate range is more or less confined to the point $q = 0.5$. With increased chain length ν, however, $q_{\min}$ is readily shifted towards $q = 1$ whereas $q_{\max}$ moves towards $q = 0$. In the case of longer sequences. polynucleotides with $\nu > 50$, both domains of faithful replication are very small compared to the broad intermediate range of stochastic replication.

Stochastic replication has another important aspect which is not quite evident from the preceding discussion of the results derived from differential equations. The number of possible sequences is superastronomic 2^{ν} or 4^{ν}, respectively, already for polynucleotides of medium chain lengths. Thus, it exceeds by far the number of individuals in any realizable population. The range of stochastic replication, as we have shown by means of exact numerical solution of the eigenvalue problem, can be understood as an extension of the point $q = 0.5$. Here incorporations of correct and erroneous bases occur with equal probabilities. Hence, all sequences are present in equal amounts in the range of stochastic replication. The deterministic approach is no longer appropriate, since we really cannot have less than a single copy of a given sequence in the volume under consideration. We are dealing with a set of sequences which changes from generation to generation. New sequences appear due to copying errors and a certain percentage of the old sequences disappears as a consequence of degradation and dilution. The notion 'presence in equal amounts' can be replaced (at best) by 'equal probability of realization' in the course of a long-term experiment (The increase in accessible sequences due to the effect of multiple turnover is rather limited, since the time available is small compared to the largeness of numbers like 2^{ν} or 4^{ν} with $\nu > 100$.).

Faithful replication, direct as well as complementary, can be approximated appropriately by differential equations despite the breakdown of the deterministic approach in the intermediate range of accuracy. For this goal we restrict the variables of the differential equation to the master sequence and its most frequent mutants. Rare mutants are not accounted for explicitly. The appearance of a rare mutant is not described by the deterministic system. It can be considered as a stochastic event. Mutants less efficient than the master sequence will soon disappear after they have entered the ensemble of replicating molecules. More efficient mutants may replace the master sequence. Evolution in such an extended deterministic system of differential equations reflects the concerted action of chance represented by the stochastic event and necessity acting through the selection process.

Acknowledgements

The authors are indebted to Professors Manfred Eigen and Karl Sigmund for many stimulating discussions. Preparation of the manuscript by Mag. B. Gassner, Mrs. J. Jakubetz and Mr. J. König is gratefully acknowledged. Financial support was provided by the Austrian Fonds zur Förderung der wissenschaftlichen Forschung (Project Nos. 3502 and 4506), computer time by the Interuniversitaires Rechenzentrum Wien.

References

1 M. Arrigoni and A. Steiner, MNU Math. Naturwiss. Unterr. 33 (1980) 385.

2 E. Batschelet, E. Domingo and C. Weissman, Gene 1 (1976) 27.

3 C.K. Biebricher, M. Eigen and R. Luce, J. Mol. Biol. 148 (1981) 369.

4 E. Domingo, M. Davila and J. Ortin, Gene 11 (1980) 333.

5 E. Domingo, R.A. Flavell and C. Weissman, Gene 1 (1976) 3.

6 E. Domingo, D. Sabo, T. Taniguchi and C. Weissman, Cell 13 (1978) 735.

7 W. Ebeling and R. Feistel, Stud. Biophys. 46 (1974) 183.

8 M. Eigen, Naturwissenschaften 58 (1971) 465.

9 M. Eigen and P. Schuster, Naturwissenschaften 64 (1977) 541.

10 M. Eigen and P. Schuster Naturwissenschaften 65 (1978) 7.

11 M. Eigen and P. Schuster, Naturwissenschaften 65 (1978) 341.

12 M. Eigen, W. Gardiner, P. Schuster and R. Winkler-Oswatitsch, Sci. Am. 244 (1981) 88.

13 S. Fields and G. Winter, Gene 15, (1981) 207.

14 B. Gassner and P. Schuster, Mh. Chem. 113 (1982) 237.

15 B.L. Jones, R.H. Enns and S.S. Ragnekar, Bull. Math. Biol. 38 (1976) 12.

16 F.R. Kramer, D.R. Mills, P.E. Cole, T. Nishihara and S. Spiegelman, J. Mol. Biol. 89 (1974) 719.

17 B.O. Küppers, Naturwissenschaften 66 (1979) 228.

18 B.O. Küppers, Bull. Math. Biol. 41 (1979) 803.

19 J. Ortin, R. Najera, C. Lopez, M. Davila and E. Domingo, Gene 11 (1980) 319.

20 F.W. Schneider, D. Neuser and M. Heinrichs, in: Molecular mechanisms of biological recognition, ed. M. Balaban (North-Holland, Amsterdam, 1979) p. 241.

21 P. Schuster, in: Biochemical evolution, ed. H. Gutfreund (Cambridge University Press, Cambridge, 1981) p. 15.

22 S. Spiegelman, Q. Rev. Biophys. 4 (1971) 215.

23 J. Swetina, Diplomarbeit, Universität Wien (1981).

24 C.J. Thompson and J.L. McBride, Math. Biosci. 21 (1974) 127.

25 M.V. Volkenstein, J. Theor. Biol. 80 (1979) 155.

26 L. Weimin, Kexue Tongbao 27 (1982) 445.