J. theor. Biol. (1982) **98**, 563–574

RNA Splicing: Advantages of Parallel Processing

GUNNAR VON HEIJNE

Research Group for Theoretical Biophysics, Department of Theoretical Physics, Royal Institute of Technology, S-100 44 Stockholm, Sweden

AND

MICHAEL A. SAVAGEAU

Department of Microbiology and Immunology, The University of Michigan Medical School, Ann Arbor, Michigan 48109, *U.S.A.*

(*Received* 15 *March* 1982)

Parallel and sequential modes of RNA processing are systematically compared by an analysis of the relevant kinetic reaction schemes. The parallel mode is shown to be superior in the sense that it allows molecules to be processed with larger numbers of introns, smaller losses of immature intermediates, and shorter processing times. It also is more sensitive to variations in the rate constants for individual splice-reactions, and hence more amenable to evolutionary refinements. Quantitatively, the parallel mode agrees well with published experimental data.

Introduction

Processing of ribonucleic acid provides an example of a general class of biochemical reactions in which a large molecule is modified at a number of different sites during its maturation. If the sites to be modified are far apart in the molecular structure, one may reasonably expect them to be attacked independently, either by the same of different enzymes. Alternatively, a strictly sequential mode of processing is possible, in which site A must be modified before site B can be attacked, and so on. Of course, various hybrid schemes also can be envisioned.

Splicing of eukaryotic RNA, in which the introns to be removed are usually separated by exons a few hundred nucleotides long, would seem *a priori* likely to conform to the first processing mode, and some fairly recent experimental data indicate that this might be true in many cases (Chow, Broker & Lewis, 1979; Ryffel *et al.*, 1980; Tsai *et al.*, 1980). However, no detailed analysis of the parallel reaction scheme has yet been applied to these data, and their interpretation has remained rather qualitative.

0022-5193/82/200563+12 $03.00/0

In this paper we develop a formalism describing parallel reaction schemes and apply it to the problem of RNA splicing. In addition to accounting for the experimental data, this analysis also has some interesting functional and evolutionary implications for the splicing process.

Parallel Processing

Although it has not been demonstrated in all systems, parallel processing along multiple pathways seems to be the predominant mode of RNA splicing (Sharp, 1981). Moreover, splicing is an irreversible, nuclear event (Nevins, 1979; Piper, Wardale & Crew, 1979), and the RNA molecules are somehow exported from the nucleus into the cytoplasm where no further processing is possible. In some cases, incompletely spliced specimens have been found in the cytoplasm (Chow *et al.*, 1979; Klessig & Chow, 1980; Ghosh *et al.*, 1981), indicating that complete processing is not an obligatory prerequisite for export. The fate of these exported "intermediates" is unclear: some may be degraded rapidly, whereas others appear to be functional mRNAs (Ghosh *et al.*, 1981).

Based on these general characteristics, a kinetic model describing the nuclear RNA splicing process can be constructed. We restrict ourselves to the case with non-overlapping introns, but the extension to cases with mutually excluding splices and multiple products is straight-forward. Consider an irreversible scheme in which a molecule is to be modified at a total of N sites, each modification reaction characterized by a rate constant k_i that is assumed to be independent of whether or not other sites have

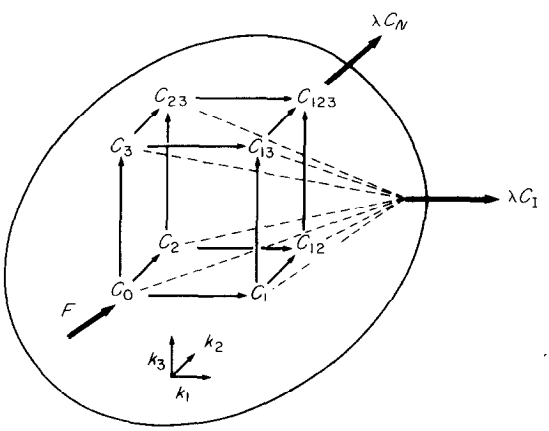

FIG. 1. A graphic representation of the parallel processing scheme with $N = 3$, c.f. equation (1).

already been modified (k_i includes the concentration of free enzyme, assumed to be constant). Let c_0 denote the concentration of substrate (no modified sites), c_N the concentration of product (all sites modified), and $c_{ij\ldots r}$ the concentration of the partly processed intermediate with sites i, $j, \ldots, r$ modified (Figure 1). The kinetic behavior of the process is described by the following equations:

$$\begin{aligned}
&\frac{dc_0}{dt} = F - \left(\lambda_0 + \sum_{l=1}^{N} k_l\right) c_0 \\
&\vdots \\
&\frac{dc_i}{dt} = k_i c_0 - \left(\lambda_i + \sum_{\substack{l=1 \\ l \neq i}}^{N} k_l\right) c_i; \qquad i = 1, .., N \qquad (1) \\
&\vdots \\
&\frac{dc_{ij}}{dt} = k_j c_i + k_i c_j - \left(\lambda_{ij} + \sum_{\substack{l=1 \\ l \neq i,j}}^{N} k_l\right) c_{ij}; \qquad i = 1, .., N-1; \quad j = i+1, .., N
\end{aligned}$$

where F is the total flow through the system (e.g. rate of precursor-RNA synthesis in the nucleus), and the λs are rate constants for removal of molecules from the system (e.g. through degradation in the nucleus, transport of RNA across the nuclear membrane into the cytoplasm, or through dilution by growth).

The steady-state concentrations are obtained by putting all time-derivatives equal to zero and solving for the cs:

$$\begin{aligned}
&c_0 = F \Big/ \left(\lambda_0 + \sum_{l=1}^{N} k_l\right) \\
&\vdots \\
&c_i = k_i c_0 \Big/ \left(\lambda_i + \sum_{\substack{l=1 \\ l \neq i}}^{N} k_l\right); \qquad i = 1, .., N \qquad (2) \\
&\vdots \\
&c_{ij} = (k_j c_i + k_i c_j) \Big/ \left(\lambda_{ij} + \sum_{\substack{l=1 \\ l \neq i,j}}^{N} k_l\right); \qquad i = 1, \ldots, N-1; \quad l = i+1, \ldots, N
\end{aligned}$$

When the rate constants are known, all concentrations can be calculated recursively from equation (2).

The overall characteristics of the system can best be studied by making some simplifying assumptions. Indeed, the need for simplification becomes

obvious when one considers the number of different intermediates in a scheme with N processing sites: there are 2^N different molecular species in the scheme, and $\binom{N}{n}$ different intermediates with n out of N modified sites.

At the present state of the art, it seems reasonable to put all λs equal (i.e. to assume that the machinery responsible for exporting RNAs across the nuclear membrane cannot differentiate between incompletely and fully processed molecules, but see Campos, Jovanovich & Villarreal (1981)). In what follows, two choices of the ks are considered: (i) all ks are equal, and (ii) $k_1 = k$, $k_i = \alpha k$ for $i = 2, \ldots, N$. These choices serve to illustrate the most important properties of the system.

CASE I: ALL kS AND ALL λS EQUAL

In this case, all intermediates with n out of N modified sites will have the same concentration. If C_n is the *total* concentration of such intermediates, equation (1) can be simplified:

$$\begin{aligned}
\frac{dC_0}{dt} &= F - (\lambda + Nk)C_0 \\
&\vdots \\
\frac{dC_n}{dt} &= (N-n+1)kC_{n-1} - [\lambda + (N-n)k]C_n \\
&\vdots \\
\frac{dC_N}{dt} &= kC_{N-1} - \lambda C_N
\end{aligned} \tag{3}$$

The full solution to equation (3) depends on the initial concentrations at $t = 0$. As an illustration, consider the case when the total flow is suddenly increased from zero to F at $t = 0$ (e.g. the amount of transcribed RNA is suddenly changed in response to a regulatory signal). The solution then becomes:

$$\begin{aligned}
C_n(t) = {} & \frac{F}{k(N-n)} \prod_{i=0}^{n} \left(\frac{N-i}{(\lambda/k + N - i)} \right) \\
& \times \left[1 - \sum_{i=0}^{n} e^{-(\lambda/k+N-i)kt} \prod_{\substack{j=0 \\ j \neq i}}^{n} \left(1 - \frac{\lambda/k + N - i}{\lambda/k + N - j} \right)^{-1} \right]
\end{aligned} \tag{4}$$

where the factor in front of the bracket is the final steady-state concentration at $t = \infty$.

Summing $C_n(t)$ from $n=0$ to $N-1$ gives the total concentration of immature RNA in the nucleus at time t, $C_I(t)$, and the fraction of immature RNA in the total flow of molecules from the nucleus into the cytoplasm, $p(t)$, can be calculated:

$$C_I(t) = \sum_{n=0}^{N-1} C_n(t) \tag{5}$$

$$p(t) = C_I(t)/(C_I(t)+C_N(t)) \tag{6}$$

If $\lambda \ll k$, these expressions can be further simplified. In the steady-state, we get the particularly simple and important result that p depends logarithmically on N:

$$p(\infty) = \lambda C_I(\infty)\Big/F \approx \frac{\lambda}{k}\sum_{n=0}^{N-1}\frac{1}{N-n} \approx \frac{\lambda}{k}\left(\ln N + C_e + \frac{1}{2N}\right) \tag{7}$$

where $C_e \approx 0{\cdot}5772$ is Euler's constant.

Similarly, it can be shown that the response time of the system, i.e. the time it will take C_n to change halfway to its new steady-state value after a sudden change in the input flux, is approximately given by:

$$\tau \approx \frac{1}{k}\sum_{n=0}^{N-1}\frac{1}{N-n} + \frac{1}{\lambda} \approx \frac{1}{k}\left(\ln N + C_e + \frac{1}{2N}\right) + \frac{1}{\lambda} \tag{8}$$

CASE II: $k_1 = k$; $k_i = \alpha k$ FOR $i = 2, \ldots, N$; ALL λS EQUAL

The steady-state equations (2) can be solved to give a recursive expression for C_n (the total concentration of intermediates with n out of N modified sites):

$$C_n = C_0\left[a_n\binom{N-1}{n-1} + b_n\binom{N-1}{n}\right] \tag{9}$$

where

$$C_0 = \frac{F}{k}\frac{1}{1+\lambda/k+\alpha(N-1)};$$

$$a_n = \frac{\alpha}{\lambda/k+\alpha(N-n)}\left[(n-1)a_{n-1} + \frac{1}{\alpha}b_{n-1}\right];\ a_0 \text{ finite}$$

$$b_n = \frac{\alpha n}{1+\lambda/k+\alpha(N-n-1)}b_{n-1}; \qquad b_0 = 1$$

Sequential Processing

In this section, we briefly analyze a simple sequential process in which the sites have to be modified in an obligatory order, starting with site one and progressing toward site N. In this case, the equivalent of equation (3) is:

$$\begin{aligned}
\frac{dC_0}{dt} &= F-(\lambda+k)C_0 \\
&\vdots \\
\frac{dC_n}{dt} &= kC_{n-1}-(\lambda+k)C_n \\
&\vdots \\
\frac{dC_N}{dt} &= kC_{N-1}-\lambda C_N
\end{aligned} \tag{10}$$

with the solution:

$$C_n(t)=\frac{F}{k+\lambda}\left(\frac{k}{k+\lambda}\right)^n\left[1-e^{-(\lambda+k)t}\sum_{j=0}^{n}\frac{[(k+\lambda)t]^j}{j!}\right]; \qquad n=0,\ldots,N-1 \tag{11}$$

$$C_N(t)=\frac{F}{\lambda}\left(\frac{k}{k+\lambda}\right)^N \times\left[1-\left(\frac{k+\lambda}{k}\right)^N e^{-\lambda t}-e^{-(k+\lambda)t}\sum_{j=0}^{N-1}\frac{[(k+\lambda)t]^j-((k+\lambda)/k)^N(kt)^j}{j!}\right]$$

$p(\infty)$ is easily calculated:

$$p(\infty)=1-\left(\frac{k}{k+\lambda}\right)^N\approx N\frac{\lambda}{k} \tag{12}$$

i.e. it depends approximately linearily on N, as does the response time:

$$\tau\approx\sum_{n=0}^{N-1}\frac{1}{k}+\frac{1}{\lambda}=\frac{N}{k}+\frac{1}{\lambda} \tag{13}$$

If we put $k_j=k$, $k_i=\alpha k$ for $i\neq j$, and all λs equal (see Case II above), p turns out to be independent of j and is given by:

$$p(\infty)=1-\frac{1}{k+\lambda}\left(\frac{\alpha k}{\alpha k+\lambda}\right)^{N-1}\approx\frac{\lambda}{k}\left(1+\frac{N-1}{\alpha}\right) \tag{14}$$

Parallel vs Sequential Processing

Two quantities seem especially important when one compares the parallel and sequential modes of processing: the steady-state value of p (i.e. the

fraction of immature RNA in the total flow from the nucleus into the cytoplasm), and the response time τ. Neither of these should be allowed to increase beyond certain limits, p_{max} and τ_{max}, set by the amount of waste and temporal sluggishness that the cell can tolerate: as p increases so does the amount of energy wasted in the synthesis of exported, immature RNA (and possibly in the synthesis of junk protein from these mRNA's), and large τs mean that there will be a long lag-time between a change in the level of transcription and the corresponding change in the concentration of cytoplasmic mRNA.

Moreover, the splicing reactions must be carried out with very high accuracy because any errors are likely to cause frameshifts in the resulting mRNA, and ultimately nonsense protein products. This would call for an elaborate, and probably quite slow, enzymatic process, reflected in small values for the ks. In the few cases studied so far, single splices seem to take around 10–20 minutes to complete (Tsai *et al.*, 1980).

Assuming then that these times cannot be substantially reduced through further evolutionary refinements of the splicing enzymes, a cell has only two ways available to keep p below p_{max}: reduce λ or reduce N. The first choice leads to an increase in the nuclear RNA pool; the second strategy limits the maximum number of introns that can be utilized in any one gene.

However, these requirements are much less stringent in the parallel process because of its weak logarithmic dependence on N, as is clear from a comparison of equations (7) and (12): larger λs are allowed for a given N, or, conversely, larger Ns are allowed for a given λ. Likewise, equations (8) and (13) show that setting an upper limit on τ gives rise to an analogous situation: small λs and large Ns give long response times, and the sequential process depends more strongly on N than does the parallel one.

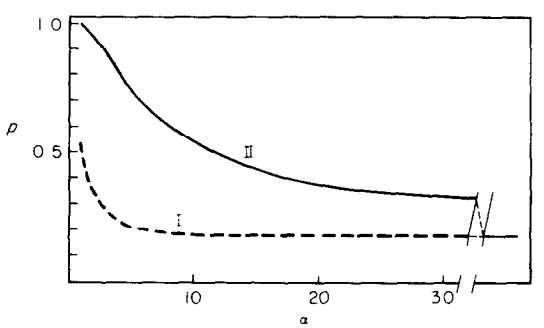

FIG. 2. Variation in p, i.e. the fraction of immature RNA in the total flow from the nucleus into the cytoplasm, with α, i.e. the quotient between k_i ($i \geq 2$) and k_1, for the parallel (I) and sequential (II) models calculated from equations (9) and (14), respectively. $N = 30$, $k = 5$ and $\lambda = 1$.

Case II and the corresponding equation (14) have been included to show the effect of unequal ks on p. Figure 2 shows how p varies with α for given k, λ and N. p decreases faster with α in the parallel case, reaching a value close to λ/k when $\alpha \approx \ln N$. A similar value for p is not reached until $\alpha \geq N$ in the sequential process.

To conclude, the parallel process allows a higher rate of RNA export from the nucleus and/or a larger maximum number of introns than does a sequential process when they are both constrained by a common p_{max} or τ_{max}. Furthermore, the improvements that result from increasing *some* but not all ks are greater in the parallel case.

Comparisons with Experimental Data

Experimental measurements of the concentrations of individual processing intermediates provides a stringent test of the parallel processing model, since in a scheme with N sites, the 2^N possible species have their steady-state concentrations determined by only $N+1$ independent variables: $k_1, \ldots, k_N$ and λ (c.f. equation (2)). The time-scale can always be chosen to make $F = 1$.

So far we have found only two experimental papers providing this kind of data, dealing with, respectively, processing of the leader sequence of the Adenovirus 2 fiber message (Chow & Broker, 1978), and intermediates of ovomucoid RNA processing (Tsai *et al.*, 1980). The method employed in both cases is that of counting R-loop hybrids on electron micrographs. The total number of molecules analysed is rather small, and the counts give at best a rough idea of the actual concentrations of the various species. Nevertheless, we feel that the data are reliable enough to yield a meaningful comparison between model and experiment.

ADENOVIRUS 2 FIBER RNA

In polysomal RNA coding for the Ad2 fiber protein, the common tripartite leader sequence is sometimes found spliced to as many as three extra leader segments, denoted x, y and z (Chow & Broker, 1978). Presumably, these transcripts are intermediates on the processing pathway that have escaped from the nucleus before processing is complete. The number of observed molecules with different combinations of the extra x, y and z leaders is listed in Table 1. In addition, the y leader is found in 25% of all cytoplasmic fiber RNA, i.e. the fully processed product, devoid of all extra leaders, represents about 75% of the total fiber RNA.

Taken together, these data allow an estimate of k_x, k_y, k_z and λ. A fairly good fit is obtained by putting $k_x = 0{\cdot}07$, $k_y = 0{\cdot}01$, $k_z = 0{\cdot}03$ and $\lambda = 0{\cdot}003$ (Table 1).

TABLE 1

Number of observed and calculated combinations of extra leader components in presumptive processing intermediates for Ad2 fiber mRNA. p, the fraction of immature RNA in the total flow from the nucleus into the cytoplasm, is 0·25 in both cases.

Leaders present in transcript	*xyz*	*xy*	*xz*	*yz*	*x*	*y*	*z*
Number of molecules observed	3	4	0	19	0	>33	3
Number of molecules calculated	8·8	3·2	0·9	14·4	0·8	50·5	6·2

OVOMUCOID RNA

The ovomucoid gene contains seven introns (A-G) and processing has been analysed by isolating nuclear RNA intermediates (Tsai *et al.*, 1980). Since there are a total of $2^7 = 128$ species on the processing pathway, and since only about 100 molecules have been analysed, a large number of possible intermediates have never been observed. Nevertheless, in pooled form, the data do conform quite well to a parallel mode of processing if an apparent anomaly in the original data is corrected for.

Gel electrophoresis of the nuclear ovomucoid RNAs reveals a monotonic increase in the concentration of the intermediates as one goes from unprocessed towards fully processed molecules. This type of behavior is actually indicative of parallel processing, since in this case the fastest reaction tends to occur before the second fastest one, etc., giving rise to a progressive pilling up of intermediates (i.e. increasing C_ns) as one approaches the fully processed product. This can be seen in the last column of Table 2.

However, as is clear from the column marked $\sum_{obs}$ in the Table, on the electron micrograph pictures the total number of molecules with a given number of splices per molecule at first increases but then *declines*, in contrast with the gel electrophoresis data. The reasons for this discrepancy are unclear, but the different intermediates seem to be affected to similar extents, and a tentative comparison between the observed counts and the concentrations calculated from equation (2) can be made by multiplying the latter by the factor $\sum_{obs}/\sum_{calc}$ so that the sum of each row becomes the same in both cases.

From the gel data one can further estimate that the fully processed molecule accounts for roughly 75% of total ovomucoid RNA in the nucleus. These experimental results can be well approximated by equation (2) with $k_A = k_B = k_C = k_G = 0{\cdot}005$, $k_D = 0{\cdot}01$, $k_E = k_F = 0{\cdot}025$ and $\lambda = 0{\cdot}0007$, Table 2.

TABLE 2

Observed (Tsai, M.-J., personal communication) *and calculated numbers of processing intermediates of nuclear ovomucoid RNA. The individual counts have been pooled according to the total number of splices in each molecule (e.g. 2 of the 26 molecules observed with two introns lacking had their A intron spliced out, etc.). The Table is an updated version of the one published by Tsai et al.* (1980). *The calculated value of p is* 0·25.

Intron spliced out:	A	B	C	D	E	F	G	$\sum_{obs}$	$\sum_{calc}$
Splices per molecule									
0	—	—	—	—	—	—	—	no data	12·4
1	1:0·9	2:0·9	1:0·9	0:2·0	6:6·2	7:6·2	1:0·9	18	16·2
2	2:2·9	2:2·9	1:2·9	4:6·2	17:17·2	20:17·2	6:2·9	26	22·8
3	5:3·5	6:3·5	3:3·5	8:7·9	12:14·6	11:14·6	6:3·5	17	33·1
4	8:8·5	8:8·5	8:8·5	22:16·5	17:22·8	19:22·8	14:8·5	24	49·0
5	6:5·4	8:5·4	7:5·4	6:8·6	9:9·9	8:9·9	6:5·4	10	76·3
6	5:6·1	8:6·1	7:6·1	8:7·7	8:8·0	8:8·0	4:6·1	8	143·6
7	—	—	—	—	—	—	—	no data	1075·2

Direct kinetic measurements also are consistent with a parallel process in which the largest rate constant (k_F) is approximately five times larger than the smallest one (k_A).

Discussion

In this paper, we have provided a first systematic comparison of the parallel vs the sequential mode of RNA splicing, and we have shown that the parallel scheme can account for the experimental results in the few well-studied cases available. However, in view of the limitations of R-loop analysis, more rigorous tests seem desirable. Specifically, one might suggest analyzing the processing of an RNA molecule with a fair but not too large number of introns (four or five, say) of sufficiently different lengths such that all intermediates can be separated and quantitated by some suitable method, e.g. gel electrophoresis.

From an evolutionary viewpoint, one might argue that a parallel process is simpler and requires relatively unsophisticated enzymatic machinery since the different introns are attacked independently. A sequential mode of processing, on the other hand, would require either an enzyme that would start at one end of the RNA and work its way to the other end (this may be how a *single* intron is spliced out, but does not seem to apply to the molecule as a whole (Sharp, 1981)), or RNA precursors with highly evolved secondary (or tertiary) structures that initially expose only one intron, the removal of which exposes the next one, etc., as has been suggested by Naora, Deacon & Buckle (1980).

In any case, the analysis presented here indicates that the parallel mode of processing allows for greater flexibility in the choice of rate constants and the maximum number of introns that the system can process without excessive waste. In fact, the appearance of genes such as those for vitellogenin and pro $\alpha 2$ collagen with about 35 and 50 introns, respectively (Wahli *et al.*, 1980; Wozney *et al.*, 1981), seems difficult to reconcile with a sequential process. Gene-duplications, evolution of multi-domain proteins from previously disparate elements, etc., would be much facilitated, and hence allow for greater evolutionary efficiency, if a parallel mode of processing is used. Thus, once a parallel mechanism has been established it would not seem likely to be replaced by a sequential one.

Our analysis further shows that the amount of waste in the process is largely determined by the quotient λ/k, i.e. the ratio between the rate of exit from the system and the rate of removal of individual introns. In eukaryotes, λ refers to the rate of export from the nucleus into the cytoplasm (since the rate of degradation is at least an order of magnitude less), a

quantity that conceivably could be relatively susceptible to evolutionary pressures. In prokaryotes, however, λ is given by the rate of RNA hydrolysis, which is on the order of 1 min^{-1}. The appearance of a splicing process that is not too wasteful would require splicing enzymes that are at least an order of magnitude faster with processing times around 0·1 min, i.e. at least 100-fold faster than present-day eukaryotic enzymes. It thus seems *a priori* unlikely that prokaryotic cells could evolve splicing mechanisms similar to those found in eukaryotes.

Finally, we have shown that the amount of incompletely processed molecules exported from the nucleus can be greatly reduced by increasing the ks. Since some ks will most likely be easier to increase than others (due to, e.g., differences in the primary and secondary structures of introns and splice sites), one should expect the ks for a given RNA precursor to be somewhat different. As was demonstrated above, for a parallel process a value of $k_{max}/k_{min} \approx \ln N$ will be all that is needed to bring the amount of exported, incompletely processed RNA close to its minimum value, and very little can be gained by further increases. We can thus predict that k_{max}/k_{min} should be $\lesssim 10$, which fits well with the values calculated for the Ad2 fiber and ovomucoid mRNAs.

REFERENCES

CAMPOS, R., JOVANOVICH, S. & VILLARREAL, L. P. (1981). *Nature (Lond.)* **291,** 344.
CHOW, L. T. & BROKER, T. R. (1978). *Cell* **15,** 497.
CHOW, L. T., BROKER, T. R. & LEWIS, J. B. (1979). *J. mol. Biol.* **134,** 265.
GHOSH, P. K., ROY, P., BARKAN, A., MERTZ, J. E., WEISSMAN, S. M. & LEBOWITZ, P. (1981). *Proc. natn. Acad. Sci. U.S.A.* **78,** 1386.
KLESSIG, D. F. & CHOW, L. T. (1980). *J. mol. Biol.* **139,** 221.
NAORA, H., DEACON, N. J. & BUCKLE, D. (1980). *J. theor. Biol.* **85,** 83.
NEVINS, J. R. (1979). *J. mol. Biol.* **130,** 439.
PIPER, P., WARDALE, J. & CREW, F. (1979). *Nature (Lond.)* **282,** 686.
RYFFEL, G. U., WYLER, T., MUELLENER, D. B. & WEBER, R. (1980). *Cell* **19,** 53.
SHARP, P. A. (1981). *Cell* **23,** 643.
TSAI, M-J., TING, A. C., NORDSTROM, J. L., ZIMMER, W. & O'MALLEY, B. W. (1980). *Cell* **22,** 219.
WAHLI, W., DAWID, I. B., WYLER, T., WEBER, R. & RYFFEL, G. U. (1980). *Cell* **20,** 107.
WOZNEY, J., HANAHAN, D., MORIMOTO, R., BOEDTKER, H. & DOTY, P. (1981). *Proc. natn. Acad. Sci. U.S.A.* **78,** 712.