Phenotypic Traits and Regulatory Role of RNA Folding in Molecular Selection

Ariel Fernández

Department of Chemistry, University of Miami, Coral Gables, Fla. 33124, U.S.A. and
Department of Biochemistry and Molecular Biology, Medical School,
P.O. Box 016129, Miami, Fla. 33101, U.S.A.

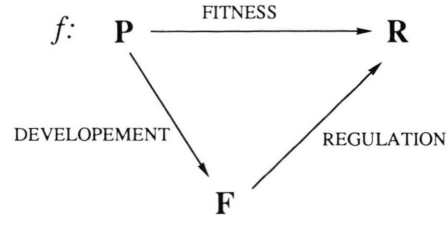
Z. Naturforsch. **46c,** 656–662 (1991); received March 5/April 3, 1991

Biochemical Kinetics, RNA Replication, Statistical Mechanics

We concentrate on instances in which the phenotypic expression of information encoded in an RNA primary sequence might be revealed by the folding of the RNA itself. We have discovered that this situation finds concrete realization in the design of RNA molecules capable of maximizing the rate of autocatalytic synthesis when incubated with viral Qβ-replicase. This requires that we introduce the notion of phenotypic traits at the molecular level. Thus, the problem of finding RNA sequences whose phenotype favorably influences propagation amounts to finding RNA sequences which fold so as to optimize enzymatic performance and are in addition endowed with the proper recognition sites. The proof that these two problems are indeed equivalent has two steps: First we predict the metastable folded structures formed as a template RNA chain grows by sequential incorporation of nucleotides. The transient folded states appear to be involved in the regulation of the enzyme activity and they occur in a manner which is "oblivious" of thermodynamic time scales. Secondly, we compute the *time-dependent* activation energy for relaxation of each intermediate structure. This is done to establish *constraints* necessary for optimization of the regulatory role of RNA folding. The search for prospective template sequences is subject to such constraints. Our results aim at elucidating an optimization process realized by molecular selection in *de novo* (template-free) RNA synthesis by Qβ-replicase. We argue that the phenotype which mediates selection is given by metastable folding which emerges *together with* the printing of the genotype, that is, within the time span of a replication turnover.

Introduction

More than a decade ago, it was demonstrated that viral Qβ-replicase can assemble an RNA species able to subsequently direct its own synthesis in an autocatalytic fashion [1]. The assembling of the replicating molecule appears to take place in complete absence of traces of template, thus, the term *de novo* synthesis has been coined [2]. The resulting species might differ from run to run, but in each case a single species is produced whose population grows to detectable amounts. In order to circumvent the apparent breakdown of the central dogma of biology, it has been proposed [2] that the assembling of an RNA which serves as template is the result of an evolutionary process whereby selection pressure is imposed by the selectivity of Qβ-replicase. If that were so, a question which naturally arises is: What is the phenotype expression of the RNA primary sequence mediating molecular selection? In order to reply we must first demonstrate that molecular evolution can be perceived as an optimization of the regulatory role of RNA. That is, the target is an RNA sequence, obviously endowed with primers and adequate internal recognition sites, which optimizes the regulation of the enzyme activity. In view of this, our problem may be casted in more concrete terms: How is the fitness parameter f of a particular RNA sequence dependent on its phenotypic expression? If R denotes real numbers, P, sequence space and F the set of phenotypes, we are focusing on the possibility of factorizing f: P → R through F, as shown in the diagram below. The two paths indicated by the arrows must be equivalent:



Thus, as Dawkins has stated [3], the key difficulty is to specify the developmental steps which,

Reprint requests to Prof. A. Fernández.

Verlag der Zeitschrift für Naturforschung, D-7400 Tübingen
0939–5075/91/0700–0656 $ 01.30/0

given a primary sequence and the replicative apparatus, lead to the phenotypic expression. Our task is twofold: a) Conveniently define a phenotype at the molecular level for the Qβ-replicator and b) specify the *algorithm* which yields the phenotypic traits mediating molecular selection.

The Molecular Phenotype

The rate of polymerization of MDV-1 RNA (the natural template of Qβ-replicase) in template-instructed replication is variable and presents well-defined pauses at specific positions along the RNA chain [4]. This was experimentally established by Mills and coworkers making use of pulse-chase gel electrophoresis, trapping the replication intermediate which corresponds to each pause. These pauses have been shown to be the responses to signals defined by transient RNA structures formed concurrently with the assembling of the RNA molecule. Specifically, it has been proven making use of Monte-Carlo simulations [4, 5] that each pause is induced by a refolding event upstream of the replication fork. In order to understand how this is established, we need to point out that the simulations mimick a Markov process made up of chain elongation (polymerization) and refolding events taking place in the growing chain. Thus, as new possibilities for folding arise, previously-existing metastable structures are dismantled to allow for the formation of the emerging ones. A choice is made at each stage between a refolding event or a polymerization event based on the respective rates of the events. Thus, unless a refolding event may occur faster than a polymerization event, the latter will take place until progressive polymerization leads to better folding possibilities. At this point, a pause will be induced by the favored refolding event. In this way, the author has established that each pause site coincides exactly with the locus along the sequence where not further polymerization is favored until a refolding event has taken place. Based on this evidence, the author has conjectured that each intra-chain refolding event is responsible for a partial relaxation of the enzyme-template interaction, which is required in order for Qβ-replicase to move forward along the replication fork: The template-replicase interaction would be replaced by an intra-chain interaction in the template. Thus, the scenario of intra-chain folding which occurs to avoid the enzyme environment is somewhat justified and it is precisely in this sense that we understand the regulatory role of RNA structure. In other words, the pauses may be correlated with the modulation of the footprint of the enzyme: contraction occurs at the pause sites and progressive expansion elsewhere.

To summarize the preliminary work, one may say that by studying how an RNA molecule explores sequentially configuration space, searching for structures as it being synthesized, we may establish the regulatory role of RNA structure in replication [4, 5]. Such results reveal the paramount importance of refolding events in the regulation of the enzyme performance and will enable us to address the problems a) and b) stated above.

The accurate prediction of the pause sites leads one to think that an RNA sequence which optimizes the regulatory role is one which realizes certain refolding events as it is being synthesized. Thus, the phenotypic expression is the set of transient or metastable folded structures whose formation is kinetically-governed and takes place as the genotype is being printed (or the replica is growing). The following comparative table might clarify the problem of defining a suitable phenotype at the molecular level for the Qβ-system:

Table I. Contrast between a general evolutionary scenario (left column) and an instance in which evolution operates at the molecular level (right column).

Selection pressure	Selectivity of the replicase
Molecular evolution	*de novo* RNA synthesis
Phenotypic expression	Metastable RNA folding occurring during printing of genotype
Embryology	Refolding events concomitant with chain growth

For the sake of completion, we shall now sketch the description of the algorithm for the development of the phenotype. *It is essential to emphasize that our simulation accounts for a scenario in which the phenotype emerges as the genotype is being printed.*

The Markov process is comprised of three different kinds of *kinetically-governed* elementary events: i) intra-chain partial helix formation, ii) intra-chain helix decay and iii) chain growth by incorporation of a single nucleotide. The transi-

tion time for each event in the Markov process is a Poissonian random variable. If an admissible helix formation happens to be the event favored, the inverse of the mean time for the transition will be given by:

$$t^{-1} = f\, n \exp(-\Delta G_{\text{loop}}/RT) \tag{1}$$

where f is the kinetic constant for a single base-pair formation (estimated at $10^7\ s^{-1}$, *cf.* [6]), n is the number of base pairs comprising the helix and ΔG_{loop} is the change in free energy of the set of all loops due to the folding which leads to the new intra-chain stem formation. The formation of new helices should always be topologically compatible with the pattern of existing ones in the sense that no knots can be allowed. This condition has been given proper combinatorial form and as such is incorporated in the algorithm in a standard manner.

If intra-chain helix decay is the chosen event, the inverse mean time can be obtained from an improved version of the expression for the kinetics for helix decay, obtained by Anshelevich *et al.* [6]. These authors give the equation:

$$t^{-1} = f\, n\, S(\text{eq.})^{-n} \tag{2}$$

where S(eq.) = equilibrium constant for base-pair formation. However, their treatment does not properly distinguish between stacking and initiation of the base-pairing process. Thus, we shall use instead the improved equation:

$$t^{-1} = f\, n\, [KS^{n-1}]^{-1} \tag{3}$$

where S = geometrical mean of the base-stacking equilibrium constants (adequate for a random uncorrelated primary sequence) and K = equilibrium constant for base-pairing initiation (nucleation equilibrium constant); $K(\text{A}-\text{U}) \approx 4 \times 10^{-5}\ \text{M}^{-1}$, $K(\text{G}-\text{C}) \approx 2.5 \times 10^{-4}\ \text{M}^{-1}$.

Finally, if a polymerization event happens to be favored, the rate constant for phosphodiester linkage formation, $t^{-1} \approx 50\ s^{-1}$ [4], should be adopted as transition rate.

The Markov process is simulated by selecting one of the three possible elementary events at each stage. The effective transition time for the chosen elementary event is a Poissonian random variable with mean k^{-1} where the effective rate constant k is given by:

$$k = \sum_{i=1}^{F} k_1(i) + \sum_{j=1}^{D} k_2(j) + k_3. \tag{4}$$

The subindices 1, 2, 3 correspond to events of type I, II and III respectively. The indices $i = 1, \ldots, F$ label helices that can be formed so that they are topologically compatible with the pattern of existing ones. The latter ones are labelled by the dummy index $j = 1, \ldots, D$. In order to implement the simulation, we shall relabel the rate constants as follows:

$$\begin{aligned} &k = \sum_{m=1}^{M} k'_m,\ M = F + D + 1 \\ &k'_1 = k_1(1), \ldots, k'_F = k_1(F), k'_{F+1} = k_2(1), \ldots \\ &k'_{F+D} = k_2(D), k'_{F+D+1} = k_3. \end{aligned} \tag{5}$$

This is done in order to find the transition index m at each stage of the process. Thus, we consider a uniformly distributed random variable R, $0 \le R \le k$, so that if the value r of R lies in the interval

$$\sum_{m=1}^{m'-1} k'_m \le r \le \sum_{m=1}^{m'} k'_m \tag{6}$$

then, the index m' has been chosen.

The key quantity accessible from the simulation described is $p(n, t)$, the probability of a certain structure n at time t. This probability is given by:

$$\begin{aligned} &k(n-1 \rightarrow n, t) \geq \Sigma_\beta\, k(n \rightarrow \beta, t) \\ &p(n, t) = \{\Sigma_\mu\, k(n-1 \rightarrow \mu, t)\}^{-1} \times \\ &\{k(n-1 \rightarrow n, t) - \Sigma_\beta\, k(n \rightarrow \beta, t)\} \end{aligned} \tag{7}$$

where α, β, μ, n, $n+1$ denote folding patterns occurring during replication and $k(\alpha \rightarrow \beta, t)$ is the time-dependent rate of refolding of structure α to yield structure β. These rates depend solely on the transition rates for whichever elementary events are required to refold the first structure into the second one. Obviously, all rates $k(\alpha \rightarrow \beta, t)$ will vary as more nucleotides are incorporated to the growing chain and, in this implicit sense, the rates are time-dependent. The probabilities $p(n, t)$ are kinetically determined and obviously path-dependent since they are defined inductively. Thus, the final most probable structure for MDV-1 RNA, a natural template for Qβ-replicase, would not have been the biologically active structure depicted in Fig. 1 if it weren't that we have chosen the strong hairpin between nucleotides 1–22 as our initial point for the induction [4]. This choice is

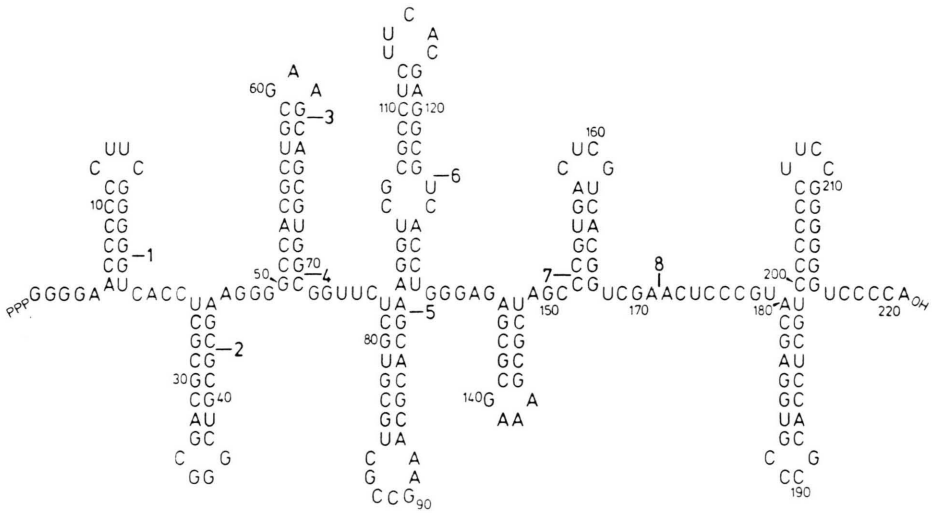


Fig. 1. Metastable most probable structure for MDV-1 RNA emerging immediately after a replication turnover. The pause sites for Qβ-replicase are labelled by digits.

a correct one for it accounts for the initiation signal for replication.

Had we chosen not to start by this structure, we would have ended up with the equilibrium structure with maximal degree of folding, schematically reproduced in Fig. 2. This structure is biologically inert since in it both primers are bind to each other.

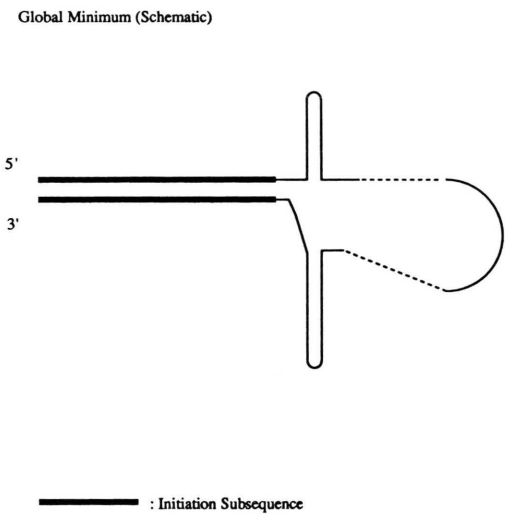


Fig. 2. Schematic representation of the global free energy minimum. This structure is biologically inert, for the two primers (represented by thick lines) bind to each other.

Of crucial importance to validate our developmental algorithm is the prediction of the major pause sites in replication, denoted by digits in Fig. 1. Each of these sites corresponds exactly to a time t^* in our simulation when two curves $p(n, t)$ and $p(n + 1, t)$ cross. That is: $p(n, t^*) = p(n + 1, t^*)$. This is revealing since it implies that a refolding event starts occurring precisely at the same time as the replicase starts idling at a particular position along the chain. Thus, we can conclude, upon examination of the distribution of pause sites in Fig. 1, that a sequence which optimizes the regulation of the enzymatic activity must realize at least one (kinetically governed) refolding event every 42 nucleotides. The phenotypic expression is then the set of metastable transient structures formed while the replica chain grows by sequential incorporation of nucleotides. However, for a particular phenotype to influence *favorably* the replication of a given strand, it must fulfill an additional restriction, as shown in the next section.

Constraints Imposed on the Phenotype

In order to find the necessary constraints the phenotypes are subject to, we shall attempt to design prospective RNA templates, inserting sequences on MDV-1 RNA. Consider, for instance, the insertion of species A and B shown in Fig. 3 following the nucleotide at position 46 in the original template (Fig. 1):

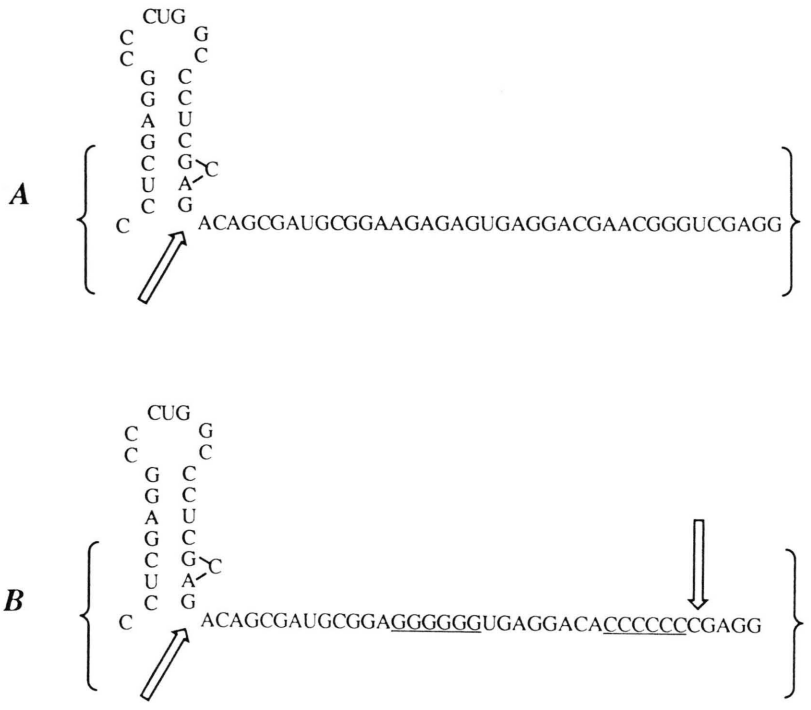


Fig. 3. Sequences to be inserted at position 46 in MDV-1 RNA. The additional pause sites resulting from the insertions are indicated by arrows. In the case of sequence B, the underlined subsequences form such a strong hairpin that the Qβ-replicase interprets it as a termination signal.

These modifications could be achieved preparing plasmids similar to psL 5, used to produce recombinant RNA by Mills, Kramers and coworkers [7], and then transcribe the results of such manipulations at the DNA level to the RNA level. None of the resulting RNA's would work as a template: In the first case (insertion A), too little secondary structure is feasible after the pause site indicated by the arrow. Thus, we predict that the replicase will not be able to copy the full inserted sequence since the footprint cannot be modulated properly. Case B is more subtle. Now the hairpin required for regulation is indeed formed, the basepairing involves the underlined sequences, but the regulatory signal is just too "strong", and thus, the final replication product will be the intermediate sequence up to the pause site at the end of the strong hairpin. This hairpin produces a termination signal, similar to that produced by the subsequence of nucleotides 200–215 in MDV-1 RNA.

The strength of a regulatory signal finds concrete meaning if we study the relaxation to equilibrium of biologically active structures. Fig. 4 displays the time dependence of the activation energy of relaxation in logarithmic scale for the abscissas [8].

The scaling constant for real time, T(nonerg.) is the minimum relaxation time. The solid thin lines correspond to random primary sequences with total length $N = 256$ and 512 respectively. The linearity of the plots indicates that the relaxation follows the so-called random energy model (REM), familiar from disordered condensed matter systems [8]. The activation energy must grow since in order to reach conformations of lower and lower energy, the chain must first unfold and then fold back into lower energy structures. Thus, since unfolded states have the same energy, the kinetic barrier becomes larger and larger as time progresses. The dashed line plot corresponds to MDV-1 RNA and the plot suggests that this biologically relevant molecule behaves far more like a random sequence than the mutant obtained by insertion B (thick solid line plot). The connection between disorder and relaxation properties is easily verified by the fact that insertion B, having very ordered subsequences (the ones underlined), may yield a highly stable hairpin, too stable to relax as random ma-

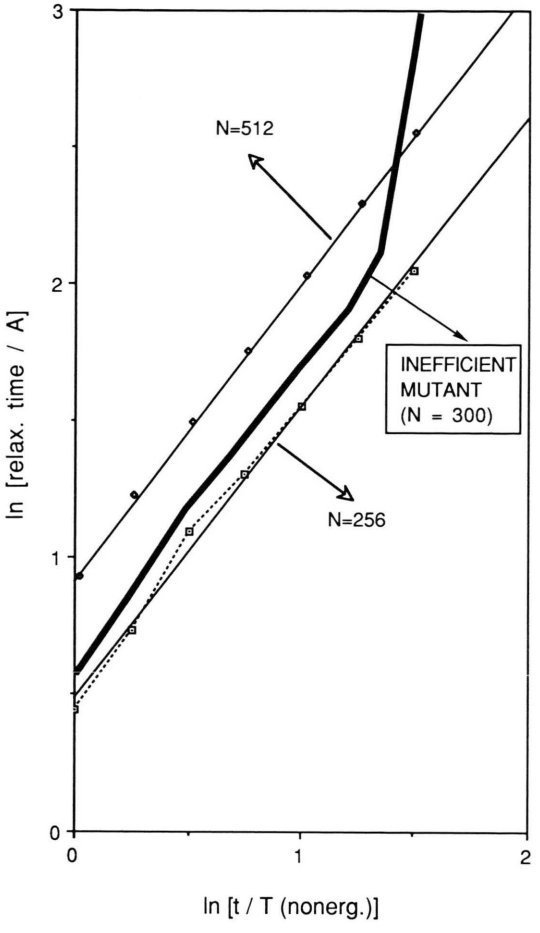


Fig. 4. Time-dependent behavior of the activation energy for relaxation. The ordinates are equal to $E(\text{act.})/RT$, the preexponential factor A being fixed at 1.12 s. The abscissas are given in logarithmic scale, with scaling factor for time $T(\text{nonerg.})$ = minimum relaxation time for any given structure. The logarithmic growth of the activation energy, revealed by the two thin solid-line plots ($N = 256$ and $N = 512$) is indicative of the validity of the random energy model (REM). The dashed-line plot corresponds to MDV-1 RNA and the thick solid line, exhibiting considerable departure from the REM, corresponds to the mutant obtained by insertion of fragment B (Fig. 3).

terials like spin glasses would do. Thus, the relaxation of the mutant due to insertion B departs considerably form the REM behavior which is a signature of the existence of additional termination signals.

At this point one could pose the question: Which sort of insertion might be favorable then? A relevant example of a favorable case would be the insertion of a subsequence of A comprised of the first 23 nucleotides, precisely up to the first pause site. The additional signal added is sufficiently soft to serve regulation: Its relaxation behavior (not shown in Fig. 4) realizes exactly the REM and almost overlaps with the solid plot for $N = 256$ in Fig. 4.

The connection between folding and fitness has already been investigated, weakening regulatory signals by site-directed mutagenesis [9]. For instance, a site mutation at position 121 in MDV-1 RNA (G → A) does not affect the internal recognition sequence but it destabilizes a stem in such a way that the relaxation of the mutant structure follows more closely the REM than the MDV-1 RNA. Again, the time-dependence of the activation energy for the mutant does not depart detectably from the linear plot for $N = 256$ in Fig. 4, whereas the relaxation of the wild type species MDV-1 RNA (dashed line) differs slightly. Moreover, when incubated together, mutant and MDV-1 RNA, with Qβ-replicase in the presence of free energy-rich nucleoside triphosphates, the mutant population overgrows that of the MDV-1 RNA [9].

Conclusion

The phenotypic expression of an RNA primary sequence in the Qβ-system is the set of refolding events which determine the regulation of the enzyme activity. The phenotype is formed as the genotype is printed, that is, as the replica is assembled. In order for a phenotype to act favorably upon replication of its genotype, it must fulfill certain restrictions. These restrictions concern the relaxation of each metastable intermediate structure: The regulatory signals cannot be too strong in the sense that their associated metastable structures should relax according to a random energy model.

Acknowledgements

The author would like to thank Profs. Manfred Eigen and C. Biebricher at the Max Planck-Institut for helpful discussions during the author's stay in Göttingen. Also, the kind hospitality of Dr. David Campbell at the Center for Nonlinear Studies, Los Alamos National Laboratory is gratefully acknowledged. This work was partly supported by an award from the Camille and Henry Dreyfus Foundation.

[1] M. Sumper and R. Luce, Proc. Natl. Acad. Sci. U.S.A. **72,** 162–166 (1975).
[2] C. K. Biebricher, M. Eigen, and R. Luce, J. Mol. Biol. **148,** 369–390 (1981).
[3] R. Dawkins, The Evolution of Evolvability, in: Artificial Life, Santa Fe Institute Studies in the Sciences of Complexity (C. Langton, ed.), Addison-Wesley Publishing Co. 1988.
[4] A. Fernández, Eur. J. Biochem. **182,** 161–163 (1989).
[5] A. Fernández, Arch. Biochem. Biophys. **280,** 421–424 (1990).
[6] V. V. Anshelevich, V. Vologodskii, A. V. Lukashin, and M. D. Frank-Kamenetskii, Biopolymers **23,** 39–58 (1984).
[7] S. A. la Flamme, F. R. Kramer, and D. R. Mills, Nuc. Acds. Res. **13,** 8425–8440 (1985).
[8] A. Fernández and E. I. Shakhnovich, Physical Review A – Rapid Communications **42,** 3657–3659 (1990).
[9] A. Fernández, Naturwissenschaften **76,** 525–526 (1989).