

# *In vitro* selection of RNA molecules that bind specific ligands

Andrew D. Ellington & Jack W. Szostak\*

Department of Molecular Biology, Massachusetts General Hospital, Boston, Massachusetts 02114, USA

Subpopulations of RNA molecules that bind specifically to a variety of organic dyes have been isolated from a population of random sequence RNA molecules. Roughly one in  $10^{10}$  random sequence RNA molecules folds in such a way as to create a specific binding site for small ligands.

THE probability that a given random sequence polynucleotide or peptide chain will fold to form a stable three-dimensional structure with a given ligand binding or catalytic activity is unknown, but is generally thought to be very low. A related problem, which has been equally difficult to address theoretically, is to estimate the number of fundamentally different classes of structures capable of carrying out a given binding or catalytic function (for example, hundreds of different proteases have been identified, but these fall into only five independent structural classes). These questions are important for theories of the origin and early evolution of life, as the first biological catalysts presumably arose from random sequence polymers<sup>1-3</sup>. We have started to address these questions by developing methods for the synthesis of large numbers of random sequence RNA molecules, and for the isolation of molecules with specific ligand binding properties from such populations.

## Synthesis of random sequence RNAs

We began by chemically synthesizing a large pool of DNA molecules made up of an average of 100 bases of random sequence flanked by defined regions at the 5' and 3' ends. The defined sequences were necessary for primer hybridization, and also contained restriction sites to facilitate cloning. The 155-nucleotide (nt) long DNAs were synthesized by solid-phase phosphoramidite chemistry on a Biosearch 8700 automated synthesizer, using an equimolar mixture of the four bases for the random portion of the sequence. On the basis of the yield of synthesized oligonucleotides (100 µg), the complexity of this pool corresponds to approximately  $10^{15}$  individual sequences.

Preliminary experiments indicated that these synthetic DNA molecules were not efficiently transcribed into RNA by T7 RNA polymerase, perhaps because of incomplete deprotection. To obtain a larger amount of fully active DNA suitable for repeated use, the initial 'random-mers' were amplified 20-fold by large-scale polymerase chain reaction (PCR) to generate a sequence library. Transcription of this pool allowed the reproducible synthesis of high yields of RNA. The quality of the pool was evaluated by sequencing 14 clones derived from the initial pool of DNA, and an additional 14 clones derived from the PCR-amplified DNA. The base composition is roughly random, with a slight excess of G. There is no bias in the distribution of di- and tri-nucleotides (data not shown). As expected from a pool of such large complexity, all 28 sequences were different, and showed no similarity to each other.

A considerable decrease in pool complexity is likely to have occurred during the PCR amplification because of inability to replicate past lesions on the chemically synthesized DNA and because of poor replication of a very GC-rich portion of the 5'

defined sequence. Primer extension experiments have shown that only 4–5% of the synthetic DNA template can be completely copied, leading to a 20–25-fold decrease in pool size. In addition, sequencing showed that about two thirds of the amplified molecules had mutations in the GC-rich region after PCR; if these were preferentially amplified from rare mutants, the pool complexity could have been reduced by an additional factor of three or more. We estimate that the effective complexity of the final amplified pool corresponds to  $\sim 10^{13}$  different sequences.

## *In vitro* selection

The double-stranded DNA molecules generated by PCR have a T7 RNA polymerase promoter at one end; *in vitro* transcription with T7 RNA polymerase results in the formation of a pool of RNA random-mers. This RNA pool was purified by denaturing polyacrylamide gel electrophoresis to ensure size homogeneity, then subjected to a brief incubation at 70 °C to allow conformational equilibration. RNA molecules in this population that were capable of binding to specific ligands were then enriched by affinity column chromatography. About 30 µg of this RNA ( $\sim 4 \times 10^{14}$  molecules, or an average of 40 copies of each of the  $10^{13}$  different sequences in the pool) was applied to the affinity column in a high-salt buffer (0.5 M LiCl) to allow the negatively charged RNAs to approach negatively charged ligands. Non-binding species were washed off the column with three column volumes of high salt buffer, and specifically bound material was then eluted with water.

Eluted RNAs were treated with reverse transcriptase to generate complementary DNAs, which were then amplified by PCR (Fig. 1). In the first cycle, only one third of the eluted RNA was converted to cDNA. Transcription of the resulting double-stranded DNA regenerates RNA, which can then be used for another cycle of enrichment. Multiple cycles of affinity chromatography and *in vitro* amplification result in the continued purification of binding species, until virtually the entire population is capable of binding to the column.

As an initial test of this method, we selected RNAs that would bind to an oligo(dT)-Sepharose (OT) column. Less than 1% of the applied RNA bound to the column in the first two cycles, but on the fourth cycle over 50% of the applied RNA bound to the oligo(dT) column (Table 1). We cloned and sequenced seven oligo(dT)-binding RNAs and, as expected, all of these contained oligo(A) stretches of from 10–24 nucleotides in length.

## Selection on dye columns

We then used the *in vitro* selection system to isolate RNAs that bind to several dyes that appear to mimic metabolic cofactors. For example, Cibacron Blue (Fig. 2) binds tightly to the NAD-binding site of many dehydrogenases and Cibacron Blue columns have been used for purification of these proteins by affinity chromatography<sup>4</sup>. The experiments described here used Cibacron Blue 3GA (CB), Reactive Red 120 (R), Reactive Yellow 86 (Y), Reactive Brown 10 (BR), Reactive Green 19 (GR) and Reactive Blue 4 (B4) attached to cross-linked, beaded agarose. We chose these molecules because they have many possible hydrogen-bond donor and acceptor groups as well as planar surfaces for stacking interactions.

Each of the six organic dye columns initially bound less than 0.1% of the applied RNA, but after four to five cycles of selection

\* To whom correspondence should be addressed.

TABLE 1 Multiple selection cycles yield dye-binding aptamers

Cycle number	Dye column							
	CB	R	Y	BR	OT	GR	B4	
1	0.08	0.1	0.08	0.07	0.7	0.02	0.1	
2	2.0	0.2	1.0	0.4	0.8	0.9	0.4	
3	0.06	0.1	0.08	0.05	1.7	0.4	0.1	
4	1.6	3.7	2.1	2.7	55	65	32	
5	49	42	57	69	85	69	59	
6	63	55	57	78	(82)	(69)	(61)	
Amplification factor	$2 \times 10^{10}$	$4 \times 10^{10}$	$2 \times 10^{10}$	$5 \times 10^{10}$	$2 \times 10^6$	$3 \times 10^8$	$1 \times 10^9$	

The fraction of RNA which bound to a given column in each cycle (% RNA retained per cycle) was calculated by dividing the number of counts eluted by the number of counts originally loaded. The 'Amplification factor' is the inverse of the product of the fraction bound in each cycle. The sixth round values for OT, GR, and B4 (in parentheses) are replicates of the fifth round.

and amplification, every column bound over 50% of the applied RNA. The cumulative amplification factor after five cycles ranged from  $3 \times 10^8$  for GR to  $5 \times 10^{10}$  for BR. Assuming that all of the molecules capable of specific binding were in fact retained in each column cycle, the amplification factor for each cycle represents the degree of enrichment of the specific binding sequences. By dividing the estimated initial pool size by the cumulative amplification factor, we calculate that each of the dye-binding RNA pools is a complex mixture of  $10^2$ – $10^5$  different sequences. The apparent complexity of the selected pools implies that there are many independent sequence 'solutions' to each ligand-binding 'problem.'

### Binding of selected RNAs is specific

The binding specificity of the dye-selected RNA pools was tested by measuring the binding of each pool to each column. Three of the six pools were found to bind only to their cognate ligands (CB, GR and B4). The R pool bound preferentially to the R column but to a lesser extent to the other dye columns, and two of the pools (Y and BR) contained species that bound to all columns, either because they failed to discriminate between the different organic dyes or because they bound to the matrix itself (Table 2). The structural similarity between CB and B4 (Fig. 2) implies that the RNAs in these pools have binding sites capable of discriminating between functional groups on ligands.

We examined the binding properties of individual sequences from the CB and B4 pools by cloning PCR-amplified DNA from both pools. Individual RNA species were generated from each clone by using PCR to reintroduce the T7 promoter and primer binding sequences, followed by T7 transcription of the PCR DNA. We have termed these individual RNA sequences 'aptamers', from the Latin '*aptus*', to fit. RNAs from the cloned aptamers were examined for binding to their cognate ligands and exhibited a range of affinities (Table 3), as measured by column retention. Eight of the CB-binding aptamers were examined for binding to the Y column. Only one bound, thus seven out of eight were specific for CB binding.

We sequenced 17 clones from the CB pool, three of which were identical, and 14 clones from the B4 pool, all of which were different. We also sequenced 14 clones from the R pool and found two identical clones, while 10 clones from the GR pool were all different; this level of duplication suggests that the total complexity of these pools is in the range of 100–1,000 sequences. Most of the clones had no detectable similarity to any of the other clones, confirming the idea that there are many completely different ways of making specific binding sites for any given ligand. Careful sequence comparison did reveal short regions of similarity between several pairs of aptamers from each pool, suggesting that these regions might be involved in binding.

FIG. 1 Schematic diagram of the *in vitro* selection cycle. DNA random-mers (top) consist of 100 bases of random sequence flanked by defined regions. PCR-amplified DNA is transcribed into RNA, and RNA molecules of the desired binding properties are selected by affinity chromatography. Bound RNA is eluted, converted to cDNA, and amplified by PCR, at which point the cycle can be repeated.

**METHODS.** The initial RNA pools were synthesized by T7 RNA polymerase transcription of the PCR-amplified random-mers (500  $\mu$ g DNA in a 5 ml reaction), subsequent RNA pools were synthesized by transcription of 1–2  $\mu$ g of PCR DNA in an 80  $\mu$ l reaction mix containing 3.2% PEG 8000, 5 mM each ATP, GTP, UTP and CTP, 5 mM DTT, 40 mM Tris-HCl, pH7.9, 26 mM MgCl<sub>2</sub>, 0.01% triton X-100, 1 mM spermidine, 60 U RNAsin (Promega), 300 U T7 RNA polymerase (New England Biolabs) and 20  $\mu$ Ci of [ $\alpha$ -<sup>32</sup>P]GTP. After 16 h at 37 °C, samples were treated with 2U of RQ1 DNase for 1 h at 37 °C to remove all DNA template. RNA transcripts were purified on 6% acrylamide, 7 M urea gels as previously described<sup>6</sup>. About 30  $\mu$ g of RNA in 0.6 ml of 0.5 M LiCl, 20 mM Tris-HCl, pH 7.6, 1 mM MgCl<sub>2</sub> (column buffer) was heated to 70 °C for 5 min, allowed to cool to room temperature and applied to a 3.5 ml column of immobilized ligand (dyes linked to 4% cross-linked agarose, Sigma; oligo(dT)<sub>12–18</sub> on cellulose, Pharmacia). The column was then washed with 13 ml of column buffer and bound species were eluted with 10 ml of water. In cycles 4, 5 and 6, the water elution was followed by a 2 mM EDTA elution to remove tightly bound species. Eluted material was precipitated using 200  $\mu$ g of glycogen as carrier, and 1/3 to 1/60 of the total eluted RNA was converted to cDNA as described<sup>7</sup> using an 18-mer primer complementary to the 3' end (5'-AAGCTTCCCGGGCTGCAG-3'). The cDNA was amplified to 3–6  $\mu$ g of double-stranded DNA by PCR<sup>8</sup> with a 34-mer primer complementary to the 5' defined region; (5'-TAATACGACTCACTATAGGGGAG-AATTCGCCGGC-3'). A shorter version of this primer (a 29-mer) was used to amplify the initial DNA pool. PCR DNA was purified by chloroform extraction and a Sephadex-G50 spin column before transcription for subsequent cycles.

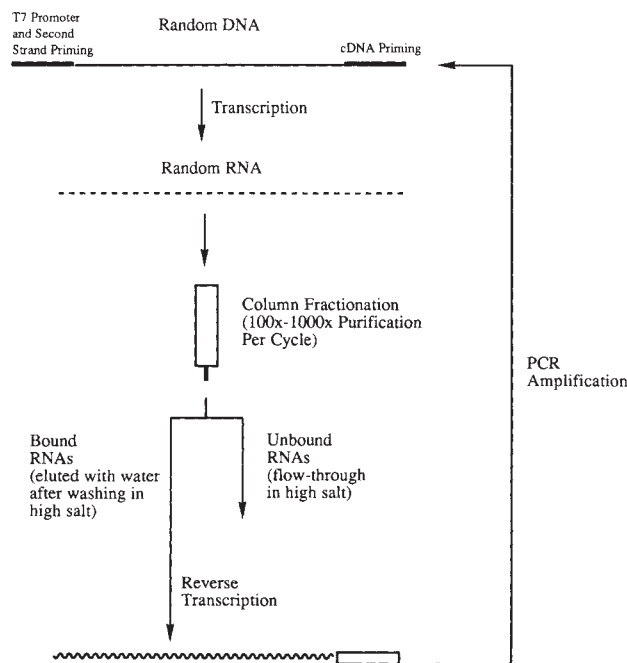




TABLE 3 Individual aptamers bind their cognate ligands

Name of cloned aptamer	Retention on cognate column (% binding)	Retention on Y column (%)	Mutational variant (cloned from mutagenized monoclonal)	Retention (%)
<b>Cibacron blue binding RNAs</b>				
CB-29	49	<1		
CB-30	9	3		
CB-31	65	80		
CB-37	9	<1		
CB-38	25	<1		
CB-41	8	—		
CB-42	71	1	CB-42 wt	82
CB-44	0	—	CB-42 No. 1	76
CB-45	64	2	CB-42 No. 9	69
CB-46	16	<1	CB-42 No. 10	73
<hr/>				
Average of all CB sequences	32			
CB pool	52			
<b>Reactive blue 4 binding RNAs</b>				
B4-19	40			
B4-21	3			
B4-22	44			
B4-25	50		B4-25 wt	65
B4-26	13		B4-25 No. 1	64
B4-27	29		B4-25 No. 4	55
B4-44	25		B4-25 No. 5	69
B4-45	20			
B4-49	42			
B4-51	32			
B4-52	53			
<hr/>				
Average of all B4 sequences	32			
B4 Pool	32			

PCR DNAs from the sixth selection cycle of CB or B4 (or from the second selection cycle of mutagenized CB-42 or B4-25) were cloned into a derivative of pMLC28 (a gift of Dr Brian Seed) and sequenced. 'Monoclonal' DNA oligomers were produced by PCR amplification of individual plasmids and these oligomers were transcribed as described in Fig. 1 to produce single aptamers. After isolation, 2–5  $\mu\text{g}$  of RNA (in 150  $\mu\text{l}$  column buffer) was applied to small dye-ligand columns (1.25 ml bed volume) and these columns were developed as before (with appropriately smaller volumes), except that two column-volumes of 2 mM EDTA were used for elution. Fractions were counted and retention was determined by dividing the number of counts in the EDTA wash by the total number of counts applied. 'Retention on Y column' refers to cross-binding experiments between CB aptamers and Y-agarose; —, value not determined.

sequences capable of specific binding to a variety of small ligands. Because the total possible number of sequences 100-bases in length is  $10^{60}$ , even a pool of  $10^{13}$  molecules contains only a very small fraction of all possible sequences ( $10^{-47}$ ), and therefore the total number of different sequences that can bind to a given ligand must be in the range of  $10^{49}$ – $10^{52}$ . The question

of how many of these sequences represent truly independent solutions (that is, structurally different binding modes) is difficult to answer at this stage, but a rough estimate can be made. It is clear from our sequence data on RNAs from the CB and B4 pools that most of the binding RNAs bear no recognizable sequence similarity to each other. But we found one case from each pool of short regions of sequence similarity between two isolates, and binding site mapping by mutagenesis has implicated these sequences in the actual binding site. As the isolates that bear these conserved regions are otherwise entirely different, they must have arisen independently. This frequency of isolation of similar binding sites is consistent with the size of the two binding sites we mapped by mutagenesis: for example, CB-42 contains 22 conserved nucleotides. As there are only  $10^{13}$  different combinations of 22 nucleotides ( $4^{22}$ ), all such combinations should be present in a pool of  $10^{13}$  100-mers. The fact that the pool size is reduced to  $10^2$ – $10^5$  after selection, combined with the observed frequency of similar binding sites, implies that there may be no more than about 100–1,000 really different sequence solutions to the construction of specific binding sites for these ligands. It is important to note that there is no competition for binding under our selection system (ligand in excess), and that therefore our pools contain a variety of molecules with  $K_d$ s of less than about 500  $\mu\text{M}$ . Presumably the more tightly binding species are present at lower abundances.

Our results suggest that it may be possible to isolate novel ribozymes from pools of random-sequence RNAs. Some fraction of RNA molecules selected for binding to transition state analogue affinity columns would probably catalyze the corresponding reactions, by analogy with the isolation of catalytic antibodies selected for binding to transition state analogues

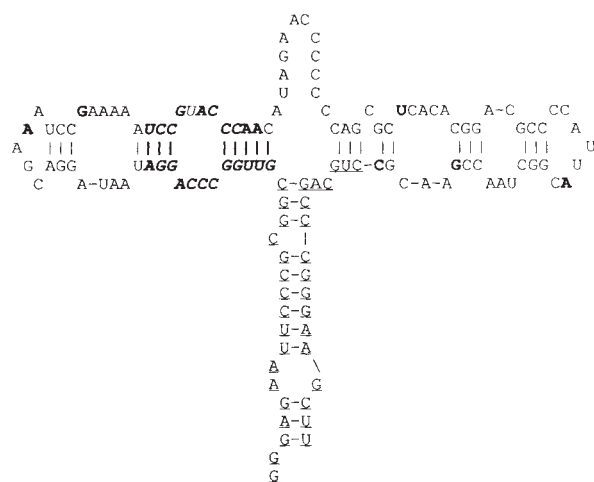


FIG. 4 Secondary structure diagram of the B4-25 aptamer generated by the FOLD program in the GCG sequence analysis package<sup>9</sup>. Primer binding sequences are underlined; the invariant residues that comprise the binding site are in bold, and those nucleotides in common with the B4-49 clone are in bold italics.

coupled to immunogenic substrates. Although it is difficult to extrapolate from substrate binding to catalytic function, the multiplicity of solutions seen for a given chemical problem

(ligand binding) suggests that complex catalysts such as the hypothesized primordial RNA replicase might also be accessible in even limited searches of sequence space. □

Received 29 May; accepted 20 July 1990.

1. Cech, T. R. *Proc. natn. Acad. Sci. U.S.A.* **83**, 4360–4363 (1986).
2. Orgel, L. E. *J. theor. Biol.* **123**, 127–149 (1986).
3. Benner, S. A., Ellington, A. D. & Tauer, A. *Proc. natn. Acad. Sci. U.S.A.* **86**, 7054–7058 (1989).
4. Thompson, R. T., Cass, K. H. & Stellwagen, E. *Proc. natn. Acad. Sci. U.S.A.* **72**, 669–672 (1975).
5. Arnold, F. H., Schofield, S. A. & Blanch, H. W. *J. Chromat.* **355**, 1–12 (1986).

6. Couture, S. *et al. J. molec. Biol.* (in the press).
7. Aruffo, A. & Seed, B. *Proc. natn. Acad. Sci. U.S.A.* **84**, 8573–8577 (1987).
8. Scharf, S. J., Horn, G. T. & Erlich, H. A. *Science* **233**, 1076–1078 (1986).
9. Devereux, J., Haeblerli, P. & Smithies, O. *Nucleic Acids Res.* **12**, 387–395 (1984).

ACKNOWLEDGEMENTS. This work was supported by Hoechst AG.

## LETTERS TO NATURE

## The Vela glitch of Christmas 1988

P. M. McCulloch, P. A. Hamilton, D. McConnell & E. A. King

Physics Department, University of Tasmania, GPO Box 252C, Hobart, Tasmania, Australia

In the past ten years, the Vela pulsar PSR 0833–45 has undergone several large, discontinuous changes—glitches—in its pulsation period. On 24 December 1988, we were making continuous radio measurements of the Vela pulsar with a 2-min time resolution when a glitch occurred. Here we report our observations, which on the day of the glitch extend from 12 h before the event to 6 h after and represent the first time a glitch has been caught as it happened. The period decrease occurred without warning and in less than 2 min; an exponential recovery similar to previously observed post-glitch behaviour began immediately. Observations were made at 635 MHz and 950 MHz, and the lower frequency signal showed an additional delay starting at the time of the glitch and continuing for about 35 days. This behaviour is consistent with a small increase in dispersion measure or a change in the pulsar's magnetic field configuration. After the glitch, the arrival times of pulses at both frequencies differed from the smoothly predicted times according to a damped sinusoidal oscillation with a period of about 25 days.

In a typical Vela glitch, an abrupt decrease in period of about a part in a million is followed by a much slower recovery. Period measurements<sup>1,2</sup> taken within hours of each of the three most recent events show that an initial recovery, characterized by a slowing down on a timescale of a few days is followed by further recovery on a timescale of ~100 days. These observations are well described by a model in which the pre-jump pulse frequency  $\nu$  is given by a power series in time with terms in  $\nu$ ,  $\dot{\nu}$  and  $\ddot{\nu}$ , and the post-jump recovery consists of a constant-frequency offset plus two exponential recoveries.

The most recent period jump of the Vela pulsar occurred at ~19:00 UT on 24 December 1988 (Christmas Day at Hobart)<sup>3,4</sup>. Our observations on that day, part of a long-standing observational programme, were made simultaneously at 635 MHz and 950 MHz using a 14-m-diameter parabolic antenna with which the source is visible for 18 h a day. The receiver at each frequency is a full polarimeter, with dual-channel field-effect-transistor amplifiers receiving orthogonal linear polarizations which are combined to give the four Stokes parameters of the signal. The temperature corresponding to the system noise at each frequency is 60 K. The local oscillators are locked to a rubidium-vapour frequency standard which also drives the station clock. Receiver bandwidths of 250 kHz at 635 MHz and 800 kHz at 950 MHz limit the pulse broadening due to interstellar dispersion to <1% of the pulse period.

Integrated pulse profiles of 1,344 pulses from each polarimeter channel are recorded at intervals of ~2 min. The profiles are subsequently combined to give profiles of total intensity and

polarization information. The signal-to-noise ratio in each total intensity profile is typically 30:1, allowing a mean pulse arrival time to be determined to an accuracy of ~80  $\mu$ s at 635 MHz and ~50  $\mu$ s at 950 MHz. During every second integration a weak noise signal with known polarization properties is injected into the system to allow gain and phase calibration.

The data presented here were collected between 31 October 1988 and 27 March 1989. This interval was chosen because microglitches occurred immediately before and after these dates, complicating the analysis. The arrival time data have been reduced to arrival times at the Solar System barycentre using standard techniques. For these calculations the optical position of the Vela pulsar given by Manchester *et al.*<sup>5</sup>, precessed to the standard epoch of J2000.0 was used with data on the barycentre from the Jet Propulsion Laboratories ephemeris DE200.

For the interval from 31 October to 24 December, the pulse phase at time  $t$  for both frequencies is well described by

$$\phi(t) = \phi_0 + \nu(t - T_0) + \frac{1}{2}\dot{\nu}(t - T_0)^2 + \frac{1}{6}\ddot{\nu}(t - T_0)^3 \quad (1)$$

The parameters of this fit are given in Table 1a ( $\ddot{\nu}$  has a negative sign because the pulsar was recovering from a microglitch at JD 2447460). The mean value of the dispersion measure to the Vela pulsar derived from our data is 68.31443  $\text{cm}^{-3}$  pc. This is less than the value obtained some years ago, in accordance with a reported trend that the dispersion measure of this pulsar is decreasing with time<sup>6</sup>.

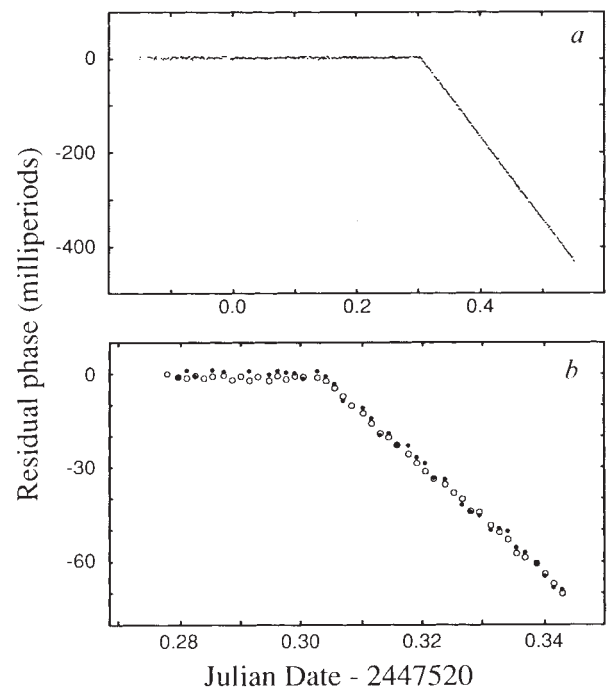


FIG. 1 The residuals from the fit using pre-jump parameters for the day of the glitch; each point represents a 2-min integration. a, 635-MHz data for the whole day. b, One hour's data including the glitch; ○, 950 MHz; ●, 635 MHz.