The EMBO Journal vol.15 no.6 pp.1203–1210, 1996

Exploring the functional robustness of an enzyme by *in vitro* evolution

Miguel Angel Martinez[1], Valérie Pezo, Philippe Marlière[2] and Simon Wain-Hobson[3]

Unité de Rétrovirologie Moléculaire and [2]Unité de Biochimie Cellulaire, Institut Pasteur, 28 rue du Dr Roux, 75724 Paris cedex 15, France

[1]Present address: Centro de Biologia Molecular 'Severo Ochoa', Universidad Autonoma de Madrid, Canto Blanco, E-28049 Madrid, Spain

[3]Corresponding author

The evolution of natural proteins is thought to have occurred by successive fixation of individual mutations. *In vitro* protein evolution seeks to accelerate this process. RNA hypermutagenesis, cDNA synthesis in the presence of biased dNTP concentrations, delivers elevated mutant and mutation frequencies. Here lineages of active enzymes descended from the homotetrameric 78 residue dihydrofolate reductase (DHFR) encoded by the *Escherichia coli* R67 plasmid were generated by iterative RNA hypermutagenesis, resulting in >20% amino acid replacement. The 22 residue N-terminus could be deleted yielding a minimum functional entity refractory to further changes, designating it as a determinant of R67 robustness. Complete substitution of the segment still allowed fixation of mutations. By the facile introduction of multiple mutations, RNA hypermutagenesis allows the generation of active proteins derived from extant genes through a mode unexplored by natural selection.

Keywords: dihydrofolate reductase/hypermutagenesis/*in vitro* protein evolution/protein robustness/reverse transcriptase

Introduction

Proteins evolved by living organisms are notoriously robust (Creighton, 1993). Not only can they function over a wide range of destabilizing conditions, they also appear largely impervious to structural alterations brought about by genetic changes. Comparison of natural sequences of enzymes descended from a common ancestor has amply shown that a near complete substitution of amino acid sequences can be effected while maintaining catalytic parameters (Doolittle *et al.*, 1989). Directed mutagenesis experiments have confirmed the elevated tolerance of proteins to genetic change (Shortle and Lin, 1985; Pakula *et al.*, 1986; Loeb *et al.*, 1989; Bowie *et al.*, 1990; Dao-Pin *et al.*, 1991; Rennell *et al.*, 1991; Shortle, 1992).

In vitro protein evolution seeks to accelerate the process of fixing mutations and holds considerable potential for biotechnology. Affinity maturation of immunoglobulin V region segments (Barbas *et al.*, 1992, 1994; Gram *et al.*, 1992) and enhanced bacterial resistance to antibiotics (Stemmer, 1994a,b) are but recent examples. Experimental protocols involve mutagenic PCR (Zhou *et al.*, 1991; Caldwell and Joyce, 1992), error prone reverse transcription of RNA (Lehotovaara *et al.*, 1988; Pjura *et al.*, 1993) or directed mutagenesis (Loeb *et al.*, 1989; Bowie *et al.*, 1990; Rennell *et al.*, 1991). Despite this, the stumbling block to extensive exploration of proteins by genetic means remains the inability to deliver elevated mutation rates. The introduction of multiple mutations at high frequencies would allow the exploration of sequence space in unprecedented ways by permitting compensating substitutions. This contrasts with the evolution of natural proteins, which is thought to have occurred by successive fixation of individual mutations (Maynard Smith, 1970). Furthermore, equivalents of millions of years of evolution would be simulated in a matter of days. How might this be achieved?

G→A hypermutation is a rather striking phenomenon in retrovirology whereby hundreds of G residues distributed over many kilobases may be substituted by A (Pathak and Temin, 1990; Vartanian *et al.*, 1991, 1994). It is a consequence of reverse transcription in the presence of highly biased intracellular dCTP and dTTP concentrations. G→A hypermutation has recently been reproduced *in vitro* with RNA, preferably that of human immunodeficiency virus type 1 (HIV-1), dNTPs and reverse transcriptase (Martinez *et al.*, 1994, 1995). The average G→A substitution frequency may be modulated as a function of the [dCTP]:[dTTP] ratio. The most elevated mutation and mutant frequencies achieved to date are of the order of 0.2 and 0.9 respectively. By exploiting different dNTPs biases, as well as using complementary RNA, a sequence could be hypermutagenized to an unprecedented degree in six different ways, i.e. G→A, U→C, A→G, C→U, G→A plus U→C and A→G plus C→U (Martinez *et al.*, 1994).

The model protein chosen for iterative hypermutagenesis was type II dihydrofolate reductase (DHFR), encoded by plasmid R67 isolated from *Escherichia coli* (Pattishal *et al.*, 1977), which has four advantageous traits. First, R67 DHFR confers resistance to the antibiotic trimethoprim (tmp), unlike its *E.coli* genomic counterpart, ensuring simple and rapid selection of functional clones. Secondly, it is specified by a mere 78 codons (Brisson and Hohn, 1984), so reducing the sequence space to be explored and rendering the study more informative. In addition it facilitates rapid sequencing of the variants. Thirdly, a large fraction of the residues are involved in quaternary and protein–cofactor–substrate interactions (Matthews *et al.*, 1986), as it functions as a tetramer requiring NADPH as cofactor. In other words, it should constitute a stringent test for robustness. Finally, a crystal structure of the dimer is available (Matthews *et al.*, 1986), while a 1.7 Å resolution crystal structure of the functional

Table I. Trimethoprim-resistant mutant frequencies following G→A and U→C hypermutagenesis of R67 DHFR plus and minus strand RNA

dNTP/μM				Hypermutation	R67 plus strand tmp^R:amp^R		R67 minus strand tmp^R:amp^R	
C	T	A	G		Ratio	%	Ratio	%
0.1	440	40	20	G→A	521:833	62.5	3132:3250	88.9
0.03	440	40	20	G→A	111:328	33.8	231:441	52.3
0.01	440	40	20	G→A	268:1964	13.6	224:1690	13.2
10	44	0.03	200	U→C	1660:2524	65.7	956:1171	80.6
0.1	440	0.1	200	G→A, U→C	556:908	61.2	380:503	75.5
0.03	440	0.03	200	G→A, U→C	20:69	28.9	182:308	59.1

tmp^R and amp^R, trimethoprim- and ampicilin-resistant colonies. The ratio tmp^R:amp^R yields the proportion of functional variants following hypermutagenesis. The type of hypermutation is given with respect to the RNA strand modified.

Table II. Base substitution frequencies for hypermutagenized R67 DHFR RNA

dNTP/μM				tmp sensitivity[a]	Colonies sequenced	No./type[b] of substitutions	Fidelity[c] ($\times10^{-2}$)	% Non-synonymous substitutions[d]
C	T	A	G					
Plus strand								
0.03	440	40	20	R	38	40 G→A	1.7	60
				S	18	48 G→A	4.2	77
0.01	440	40	20	R	68	256 G→A	5.9	90
				S	13	87 G→A	11	76
10	44	0.03	200	R	17	9 U→C	1.3	34
				S	20	35 U→C	3.2	83
0.03	44	0.03	200	R	49	45 G→A	1.4	80
						12 U→C	0.5	
				S	17	36 G→A	3.3	82
						2 U→C	0.2	
Minus strand								
0.03	440	40	20	R	35	20 C→U	1.7	50
0.01	440	40	20	R	48	127 C→U	4.5	39
				S	5	7 C→U	10	75
10	44	0.03	200	R	36	28 A→G	1.4	64

[a]R, tmp-resistant; S, tmp-sensitive.
[b]The type of transition is given with respect to the plus or mRNA strand.
[c]The fidelity of reverse transcription was calculated for each type of transition as follows: $f_{G\rightarrow A}$ = number of G→A transitions/(number of target G×number of clones sequenced). $f_{U\rightarrow C}$, $f_{C\rightarrow U}$ and $f_{A\rightarrow G}$ were calculated in an analogous manner except that the latter two were derived from G→A and U→C hypermutation of minus strand RNA. The number of target U, C, G and A in the plus strand was 52, 60, 65 and 54 respectively.
[d]Proportion of total number of mutations resulting in non-synonymous amino acid substitutions. For the plus strand the proportions expected from a random distribution of G→A and U→C transitions are, excluding the initiator methionine, 83 and 58%; for the minus strand the values are 57 and 62% respectively.

tetramer (Narayana *et al.*, 1995) enables precise structure–function analyses.

Results

Experimental strategy

The R67 gene was initially amplified from the parent plasmid, pSUR67, with primers which altered the flanking sequences. The 18 bases 5′ of AUG were modified to include A+C only, except for a pair of G residues as part of the Shine–Dalgarno sequence. The 3′ flanks were changed to G+T only, while the opal stop codon was changed to ochre. Consequently, these regions, used subsequently for PCR amplification of G→A and U→C hypermutated messenger (+) and complementary (–) RNA, would hardly be modified by hypermutagenesis. The final amplification primers used maintained the Shine–Dalgarno sequence and the initiator and stop codons, attenuating differences in transcription between different clones. PCR-amplified hypermutated products were cloned into the amp^R pTrc99A expression vector 3′ of the inducible *lac* promoter. Scoring the proportion of functional DHFR genes required simple plating on ampicilin plus trimethoprim and ampicilin only plates respectively. No mutant selection was employed, as previous work showed that mutant frequencies of >80–90% were usually achieved, depending upon the [dCTP]:[dTTP] ratio (Martinez *et al.*, 1994, 1995). This proved also to be the case for R67 DHFR RNA.

The proportion of trimethoprim-resistant (tmp^R) clones was proportional to the [dCTP]:[dTTP] ratio (Table I), indicative of the degree of hypermutation. Sequencing of both tmp^R and tmp^S (sensitive) clones revealed that the requirements for DHFR function restricted the total number of substitutions by only a factor of 2–2.5, no matter the type of reaction or strand hypermutated (Table II, fidelity). Most of the substitutions within tmp^R clones were non-synonymous, although the propor-

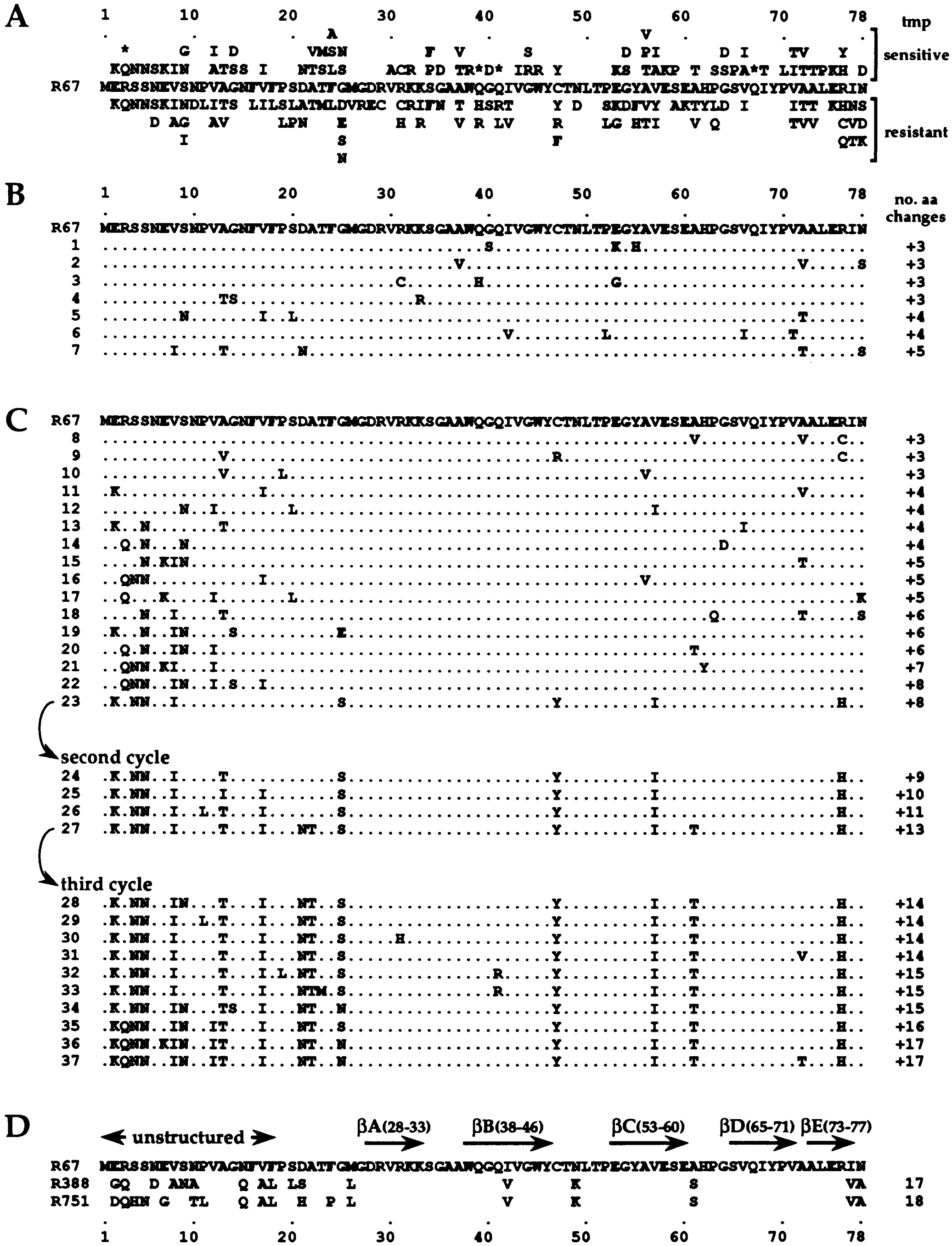

Fig. 1. Collection of R67 hypermutants. (**A**) Amino acid replacements among trimethoprim (tmp)-sensitive and -resistant clones located on the R67 DHFR sequence. An asterisk defines an in-phase stop codon. Due to multiple substitutions the same substitutions may appear in both resistant and sensitive variants. The sequence of the R67 clone used differs from that published at position 21, where aspartic acid replaces asparagine (Brisson and Hohn, 1984). (**B**) Selection of functional multisubstituted DHFR variants derived from the 30 nM dCTP or 30 nM dATP hypermutagenesis reactions using a single cDNA primer (Tables I and II). Only sequence differences are given, a dot indicating identity. To the left is the clone designation, to the right the number of amino acid substitutions per clone. (**C**) Amino acid sequences derived from multiple primer hypermutagenesis reactions using 10 nM dCTP (Tables I and II and Materials and methods). Three cycles of hypermutation are shown. The second cycle sequences were derived from clone 23, while the third cycle sequences were from clone 27. (**D**) Wild-type R67 DHFR amino acid sequence aligned with two other bacterial type II DHFRs (Swift *et al.*, 1981; Zolg and Hänggi, 1981; Flensberg and Steen, 1986). Arrows above indicate the five strands in the β-barrel and the unstructured N-terminus found in the crystal structure (Matthews *et al.*, 1986).

tions could not be simply related to the base composition of the target sequence. A compilation of amino acid replacements from a collection of 57 tmp^R and 31 tmp^S clones is given in Figure 1A. Initiator methionine apart, which was derived from the forward PCR primer, all but six residues could be substituted, demonstrating that there were few sites refractory to hypermutation. Further hypermutation reduced the number to two (see below, Figure 2B). Among the tmp^R mutants, 18 of 77 (23%) residues mapping mainly to regions 43–51 and 65–70

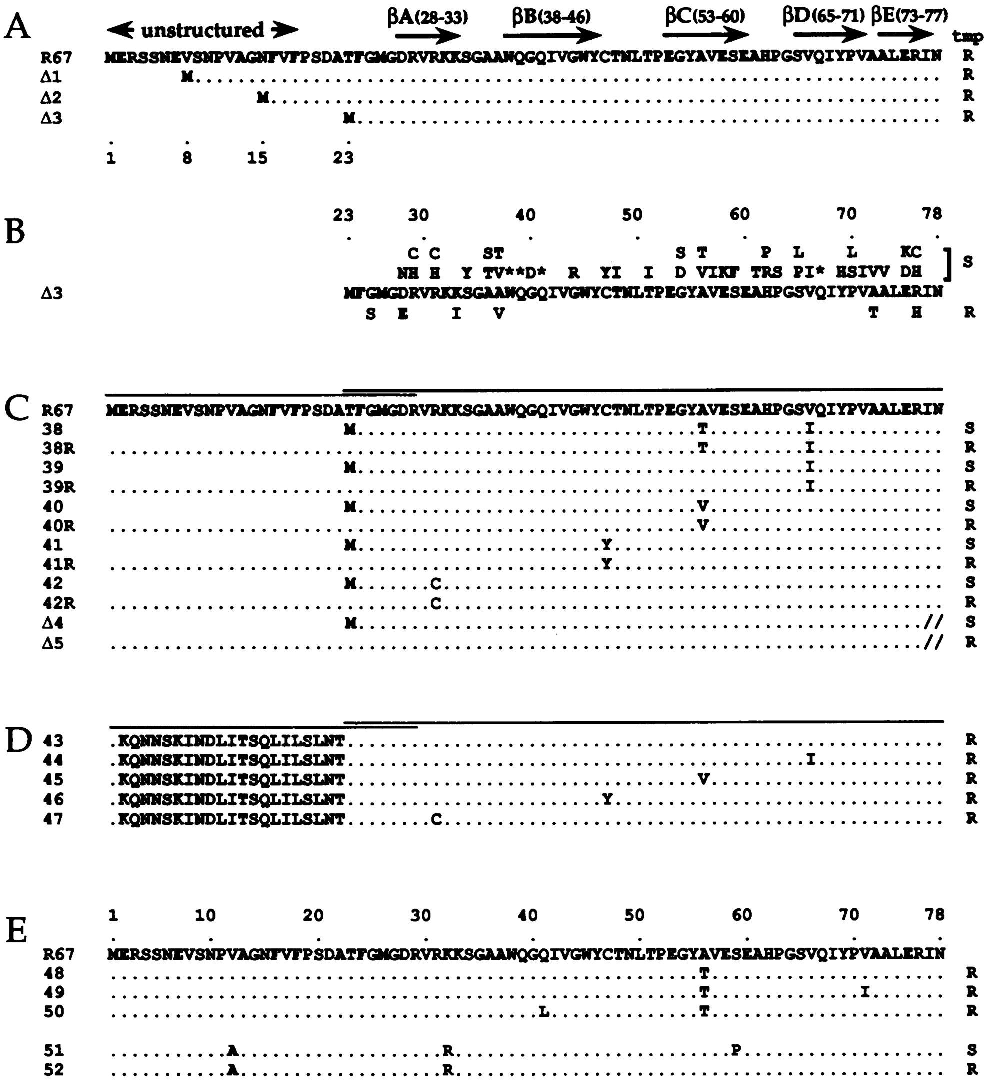

Fig. 2. Deletion analysis and hypermutation of a 56 residue form of R67 DHFR. To the right of the figure, R and S refer to tmp-resistant and -sensitive mutants respectively. (**A**) Sequence of deletion mutants at the N-terminus. Arrows above indicate the five strands in the β-barrel and the unstructured N-terminus found in the crystal structure (Matthews *et al.*, 1986). (**B**) Collection of amino acid substitutions found among 19 tmp^R (below) and six tmp^S (above) hypermutants of the minigene Δ3R67. An asterisk defines an in-phase stop codon. (**C**) Amino acid sequences of tmp^S Δ3R67 hypermutants along with the sequences following reconstruction with a DNA segment encoding the N-terminus. C-Terminal slashes were added to emphasize the dipeptide deletion. Bars above indicate the sizes of the DNA fragments used in PCR recombination. (**D**) Complete replacement of the wild-type N-terminus by a synthetic sequence chosen according to the substitutions noted in Figure 1A. DNA from defective Δ3R67 variants 38–42 (Figure 2C) were recombined with a DNA fragment by PCR. In all cases DHFR function could be recovered. Bars above indicate the sizes of the DNA fragments used in PCR recombination. (**E**) Recovery by forward mutation of efficient DHFR function from clone 50, which gave only very small plaques after 2 days growth. Clone 51 encoded a Val71→Ile substitution resulting from a G→A transition, while clone 52 encodes a Gln41→Leu replacement due to a spontaneous A→T transversion. The hypermutagenesis reaction yields ~5–10% of unanticipated mutations compared with those expected from the dNTP pool bias used (Martinez *et al.*, 1994). Finally, clone 54 represents reversion of the inactivating S59P replacement by G→A hypermutation of plus strand RNA from clone 53.

remained unchanged. That these segments are involved in extensive monomer–monomer and dimer–dimer interactions respectively may explain their conservation (Matthews *et al.*, 1986; Narayana *et al.*, 1995).

Cycling through sequence space

The main strength of hypermutagenesis is the elevated base mutation frequency (Martinez *et al.*, 1994, 1995). Functional variants with three to five non-synonymous substitutions were readily identified, representing ~5–10% of a total dominated by variants with one to two amino acid changes per gene. The multiple replacements were relatively uniformly distributed across the gene (Figure 1B). In an attempt to further increase the substitution frequency per cycle the gene was hypermutated in three segments, which allowed use of a 3-fold greater [dCTP]:[dTTP] bias. This follows from the finding that elongation efficiency dropped with increasing dNTP bias (Martinez *et al.*, 1994). The smaller the fragment amplified, the greater the dNTP bias that could be employed and, consequently, the degree of hypermutation. The complete R67 gene was reassembled from the fragments by PCR recombination (Meyerhans *et al.*, 1990). Now mutants with up to eight amino acid substitutions (eight out of 77,

Table III. Mutant frequencies of hypermutagenized tmp[R] Δ3R67 DHFR and reconstructed clones

Strand	Hypermutation	dNTP/μM				Δ3R67 No. colonies tmp[R]:amp[R]		Reconstructed R67 No. colonies tmp[R]:amp[R]	
		C	T	A	G	ratio	%	ratio	%
+	G→A	0.03	440	40	20	6:1684	0.4	436:944	46.2
+	G→A	0.01	440	40	20	22:3040	0.7	13:287	4.5
–	C→U	0.03	440	40	20	5:540	0.9	534:1032	53.6
–	C→U	0.01	440	40	20	3:779	0.4	52:714	7.3

Table IV. Base substitution frequencies for hypermutagenized Δ3R67 DHFR RNA

dNTP/μM				tmp sensitivity[a]	Colonies sequenced	No./type[b] of substitutions	Fidelity[c] ($\times 10^{-2}$)	% Non-synonymous substitutions[d]
C	T	A	G					
Plus strand								
0.03	440	40	20	R	5	2 G→A	1.3	50
0.01	440	40	20	R	28	8 G→A	0.6	12
				S	15	51 G→A	8.0	67
Minus strand								
0.03	440	40	20	R	5	3 C→U	1.5	33
0.01	440	40	20	R	4	1 C→U	0.6	–
				S	25	99 C→U	10	24

[a]R, tmp-resistant; S, tmp-sensitive.
[b]The type of transition is given with respect to the plus RNA strand.
[c]The fidelity of reverse transcription was calculated for each type of transition as follows: $f_{G\rightarrow A}$ = number of G→A transitions/[number of target G×number of clones sequenced]. $f_{C\rightarrow U}$ was calculated in an analogous manner except that the latter derived from G→A hypermutation of minus strand RNA. The number of target G on the plus and minus strands was 49 and 45 respectively.
[d]Proportion of non-synonymous amino acid substitutions. For the plus and minus strands the values expected from a random distribution of G→A transitions are 80 and 56% respectively.

or 10%) were found, many of which were concentrated in the N- and C-termini (Figure 1C). Once again, most of the replacements were non-synonymous (Table II, 10 nM dCTP reaction), reflecting the robustness of R67 DHFR.

Clone 23 was the most substituted of the variants sequenced (Figure 1C). It encoded eight amino acid replacements distributed throughout the protein, making it a good candidate for subsequent cycling. RNA was made and subjected to a second round of G→A hypermutagenesis followed by trimethoprim selection. The frequency of tmp[R] recombinants decreased only slightly with further cycling (data not shown). A maximum of five additional substitutions was accumulated, mainly in the N-terminus (Figure 1C). A third round of G→A hypermutagenesis using clone 27 RNA generated a cluster of mutants with up to five more changes (Figure 1C). Once again, the N-terminus was extensively modified. A total of 21 G→A transitions resulting in 18 sequential amino acid replacements characterized clones 36 and 37. However, as Gly25 was substituted twice via two consecutive G→A transitions, the final number of residues replaced was 17 out of 78, or 22%. The distribution and number of amino acid substitutions after three cycles were qualitatively comparable with the variation among naturally occurring *E.coli* type II DHFRs (Swift *et al.*, 1981; Zolg and Hänggi, 1981; Flensberg and Steen, 1986; Figure 1D). Many of the individual differences distinguishing them could be found in the entire collection (Figure 1A).

A dispensable N-terminus and tolerance to mutation

Such a variable N-terminus begged the question as to its role, the more so as chymotrypsin cleavage following Phe16 maintains normal reductase activity (Reece *et al.*, 1991), while the first 18 residues are completely unstructured (Matthews *et al.*, 1986; Narayana *et al.*, 1995). A series of deletions in the N-terminus up to Thr23 all yielded active DHFR genes, confirming functional seven and 18 residue deletions in the R388 type II DHFR N-terminus (Vermersch and Bennett, 1988). These findings contrast with the report that a 16 residue N-terminus deletion is inactive (Reece *et al.*, 1991). Further deletions were not made, as the crystal structure indicates the Phe24 side chain to be buried deep in a hydrophobic pocket (Matthews *et al.*, 1986; Narayana *et al.*, 1995), while cyanogen bromide cleavage at Met26 results in an inactive enzyme (Reece *et al.*, 1991). The resulting Δ3R67 construct of 56 residues constitutes one of the smallest polypeptide chains endowed with enzymatic activity and was ripe for further hypermutation. Surprisingly, the vast majority of clones were sensitive to trimethoprim (Table III, Δ3R67), down by a factor of 20–80 on experiments using wild-type R67 (Table I). This was not due to poor hypermutagenesis, as sequencing of 40 tmp[S] Δ3R67 variants revealed elevated substitution frequencies (Table IV, fidelity), comparable with those for full-length tmp[S] R67 genes (Table II, fidelity). Of the tmp[R] Δ3R67 variants only six unique sequences were found, all being singly substituted (Figure 2B). It is as though Δ3R67 is trapped in a cul de sac from which it cannot escape through point mutation.

It would seem therefore that the N-terminus conferred stability or folding efficiency on mutants within the structured β-barrel domain of the enzyme. To test this the hypermutated Δ3R67 PCR products were recombined by PCR (Meyerhans *et al.*, 1990) with an amplified segment spanning residues 1–28 of wild-type R67. Full-length products were recovered and cloned into the pTrc99 expression vector. Upon transformation tmp^R:amp^R ratios comparable with those of hypermutated wild-type R67 were found (Table III, reconstructed R67). Sequencing of a number of clones revealed hypermutated clones qualitatively and quantitatively indistinguishable from a hypermutated reaction using full-length R67 DHFR RNA (data not shown). Clearly the N-terminal 23 residues had the capacity to revitalize substituted proteins. To reinforce this finding, DNA from four defective Δ3R67 variants was individually recombined with the N-terminal fragment (see below). In all cases functional R67 variants were recovered (Figure 2C). The N-terminus was also able to rescue a two residue deletion at the C-terminus, which was otherwise lethal in a Δ3R67 background (cf. clones Δ4 and Δ5, Figure 2C).

In view of the extensive variability of the N-terminus, both among natural *E.coli* type II DHFRs and hypermutated variants (Figure 1C and D), two hypotheses presented themselves. Either the crucial residues had been missed or the precise sequence was unimportant. Of the first 23 residues, initiator methionine apart, all residues could be substituted excepting Asp15 (Figure 1A), although among the R388 and R751 sequences it was substituted by glutamine. A synthetic N-terminus was made by PCR recombination using 30–50mer synthetic oligomers encoding the replacements shown in Figure 1A, as well as the Asp15→Gln substitution. Recombination with the series of defective Δ3R67 variants shown in Figure 2C yielded viable full-length tmp^R clones (Figure 2D). Thus the primary sequence of the N-terminus could be totally changed yet retain the capacity to rescue substitutions lethal in a Δ3R67 background.

Amelioration by forward mutation

If, after transformation of hypermutated R67 DNA, the plates are left for 2 days a few very small colonies may be discerned. One such recombinant, clone 50 (Figure 2E), encoded a single Ala56→Thr substitution, as a result of a G→A transition. RNA was made from clone 50 and G→A hypermutated. Normal sized colonies (>1 mm) were found after overnight incubation. Sequencing of these clones revealed additional changes, notably Gln41→Leu or Val71→Ile, illustrating that compensating substitutions resulting from forward mutation could be identified. Relative colony size was reproducible upon transformation using plasmid DNA. A trivial explanation is lacking given that residues 41, 56 and 71 are not directly in contact in either the monomer or dimer (Matthews *et al.*, 1986; Narayana *et al.*, 1995).

Hypermutagenesis may be used to produce reversions. Clone 53 was tmp^S (Figure 2E), being produced by recombination of three segments; the first and third segments were U→C hypermutated plus strand RNA while the second was G→A hypermutated minus RNA. While G→A hypermutation of plus strand RNA yielded no colonies, that of minus strand RNA generated three of 2500 (~0.1%) tmp^R clones, all encoding reversion of the Ser59→Pro substitution.

Discussion

The several hundred tmp^R mutants sequenced should be considered as a representative sample of all the combinations produced in the multitude of hypermutagenesis reactions. Although a majority of sites were substituted, only a fraction of the combinations have been identified. Three cycles of G→A hypermutation were sufficient to reproduce qualitatively an unknown number of generations separating the R67, R388 and R751 type II DHFRs. Were the gene of eukaryotic origin, being replicated with much enhanced fidelity and long generation times, 22% amino acid substitution would represent tens of millions of years of evolution. As there was no sign that the DHFR gene was becoming saturated, it may be possible to continue cycling, simulating equivalents of hundreds of millions of years of eukaryotic evolution.

At no time were the variants in competition with one another and therefore hyperevolution occurred in the total absence of competitive Darwinian selection. Just what fraction of variants would survive the rigours of natural selection remains to be investigated. In this non-competitive setting the cost of maintaining DHFR function was only a factor of 2–3, i.e. on average one nucleotide change out of two or three was not lethal (Table II). Most of the variation was within the N-terminus, where the precise primary sequence was apparently unimportant. Even excluding from the analysis the first 18 residues, which were unstructured (Matthews *et al.*, 1986; Narayana *et al.*, 1995), up to 8 of 58 (14%) amino acids could be substituted, which still represents a substantial fraction, and this despite maintaining quaternary structure and substrate and cofactor binding sites.

Proteins can diverge considerably in a series of functional monosubstituted intermediates (Maynard Smith, 1970). Somatic hypermutation of rearranged immunoglobulin VDJ segments results in an average substitution of 10^{-3} bases/cycle, with up to 10 amino acids replaced (French *et al.*, 1989). While this is certainly accelerated evolution with respect to germline genes, it is still tantamount to substitution of a single residue per cycle and therefore little different to protein evolution in general. The introduction of multiple substitutions throughout a protein is not without precedent (Siderovski *et al.*, 1992). However, as the a R67 DHFR example demonstrates, it is possible to jump through sequence space five to eight residues at a time and land up on a viable peak. The probability of deriving functional lineages is, perhaps surprisingly, not trivially small bearing in mind the mutant frequencies and numbers of clones sequenced (Tables I–IV).

The physical basis of the rescue by the N-terminal peptide of distal amino acid replacements in R67 DHFR can only be speculated upon for the moment. It might pertain to initial folding in the ribosome or to quaternary interactions. The βE strand of R67 DHFR makes multiple contacts with the βA strand (Matthews *et al.*, 1986). Consequently, it is possible that the N-terminus influences in some way the folding pathway of part of the β-sheet structure before the ribosome has completed translation. Some subtle structural differences are suggested by the finding that the free energy difference between the guanidinium hydrochloride-denatured states for the chymotrypsin-treated protein (residues 17–78) is increased ($\Delta\Delta G$ = 2.6 kcal/mol) with respect to full-length DHFR (Reece *et al.*, 1991). Nevertheless, the identification of a gene segment whose removal does not impair function but forbids further changes may be relevant to understanding the causes and consequences of robustness, i.e. the imperviousness to mutations, translational errors and chemical accidents, in protein evolution (Ninio and Bokor, 1986). It is as though R67 encodes an anticipatory segment dispensable for present function, but essential for accommodating change.

The overall amino acid composition of proteins is also amenable to gross biases. During the *in vitro* evolution of clone 37 protein from its R67 parent the proportion of asparagine (6.4–12.8%), isoleucine and threonine (both 3.8–9%) more than doubled, to the detriment of serine, valine and alanine respectively. Continued iterative G→A hypermutagenesis of clone 37 might drive the proportions of alanine, arginine, aspartic and glutamic acid, cysteine, glycine, methionine and valine down to low or near zero values, considerably simplifying the composition of the protein. Figure 1A shows that all but four of 36 of these amino acids have been substituted in at least one functional variant. If at all possible, the elaboration of simple proteins through regressive evolution *in vitro* would provide a top down approach complementary to the bottom up approach of protein design (DeGrado *et al.*, 1989).

It appears that hypermutagenesis should allow exploration of protein robustness in a way not possible by natural selection (Maynard Smith, 1970), perhaps offering some novel solutions to old problems.

Materials and methods

R67 subcloning and RNA synthesis

DNA from pSUR67 containing the 234 bp R67 DHFR gene was amplified by PCR using primers 1 (5′-GCACGG**GAGCT**CCACAACAAAGGAACCAAATGGAACGAAGTAGC) and 2 (5′-GCACCG**GGATCC**AACACCCAACCACCAACTTAGTTGATGCGTTC) containing *Sac*I and *Bam*HI restriction sites (bold type). The A+C rich sequences are underlined. PCR products were digested with *Sac*I and *Bam*HI and ligated into the pBluescript SK+ vector. One microgram of the resulting plasmid was digested with *Sac*I or *Bam*HI and used as substrate for *in vitro* transcription using T3 or T7 RNA polymerase. For subsequent cycles of hypermutagenesis (see Figure 1C) or in order to produce RNA from R67 DHFR-deleted clones (see Figure 2A and C) primers 3 (5′-GCACGG**GAGCTC**ATTAACCCTCACTAAAGGGACACAACAAAGGAACCAAATG) and 4 (5′-GCACCG**GGATCC**AATTTAATACGACTCACTATAGGGAAACACCCAACCACCAACTTA) were used. Primers 3 and 4 contain the T3 and T7 RNA polymerase promoter sequences respectively (underlined), allowing production of plus and minus strand R67 DHFR gene transcripts. The resulting PCR fragment was purified from a 2% low melting point agarose gel and used as template for *in vitro* transcription. Reaction conditions were 40 mM Tris–HCl, pH 8, 30 mM $MgCl_2$, 10 mM β-mercaptoethanol, 50 μg/ml RNase/DNase-free bovine serum albumin, 500 mM each NTP, 200 ng PCR template or 1 μg plasmid template, 0.3 U/ml RNase inhibitor (Pharmacia) and 2 U T3 or T7 RNA polymerase (Pharmacia) in a final volume of 100 μl. After 1 h incubation at 37°C the DNA template was digested with 0.075 U/μl RNase-free DNase I (Pharmacia) for 30 min at 37°C. RNA was phenol extracted and ethanol precipitated.

RNA hypermutagenesis

Two picomoles of primer 5 (5′-GCACGG**GAGCTC**CACAACAAAGGAACCAAATG, complementary to R67 minus strand) or primer 6 (5′-GCACCG**GGATCC**AACACCCAACCACCAACTTA, complementary to the R67 plus strand) was annealed to 0.5 pmol template RNA in a 50 μl reaction by first heating to 65°C for 1 min followed by incubation at 37°C for 1 min, after which 0.3 U/μl RNase inhibitor (Pharmacia) and 15 pmol (6.25 U) HIV-1 reverse transcriptase (Boehringer) were added. The reverse transcription reaction buffer was 50 mM HEPES, pH 7, 15 mM magnesium aspartate, 10 mM dithiothreitol, 130 mM potassium acetate, 15 mM NaCl and varying dNTP concentrations (see Table II). To recover sufficient material for subsequent cloning cDNA was amplified by PCR with primers 5 and 6, producing a DNA fragment with *Sac*I and *Bam*HI restriction sites at the extremities. PCR conditions were 2.5 mM $MgCl_2$, 50 mM KCl, 10 mM Tris–HCl, pH 8.3, 200 μM each dNTP, 10 μl reverse transcription reaction and 2.5 U Taq DNA polymerase (Cetus) in a final volume of 100 μl. Cycling parameters were: 37 (30 s), 72 (30 s) and 95°C (30 s) for two cycles and 55 (30 s), 72 (30 s) and 95°C (30 s) for 13 cycles.

Multiple primer hypermutagenesis reaction and cloning

Hypermutagenesis with a dCTP concentration of 10 nM (see Table II and Figure 1C) was carried out in six separate reactions using different primers. Plus strand primers were primer 13 (complementary to bases 78–93), primer 18 (5′-CTCGACGGCGTAGCCT, complementary to bases 159–174) and primer 6. Minus strand primers were primer 5, primer 19 (5′-GGACGCCACGTTTGGT, bases 60–75) and primer 20 (5′-ACAAATTTGACCCCCG, bases 142–157). Following the hypermutagenesis reaction the products were amplified separately by PCR using primer pairs 5 and 13, 19 and 18, and 20 and 6. Cycling parameters were: 37 (30 s), 72 (30 s) and 95°C (30s) for two cycles and 55 (30 s), 72 (30 s) and 95°C (30 s) for 10 cycles. To generate the 234 bp R67 DHFR gene from the above three overlapping PCR fragments 1 ng of each gel-purified fragment was resuspended in a standard PCR mixture. No additional primers were added at this point. Cycling parameters were 37 (30 s), 72 (30 s) and 95°C (30s) for two cycles and 55 (30 s), 72 (30 s) and 95°C (30 s) for 10 cycles. After this 50 pmol primers 5 and 6 were added and 10 additional PCR cycles [55 (30 s), 72 (30 s) and 95°C (30 s)] were performed. Appropriate PCR fragments were purified from a 2% agarose gel, digested with *Sac*I and *Bam*HI and ligated to the expression vector pTrc99A (Pharmacia). After transformation of XL-1 blue cells half of the transformation was plated on ampicilin plates and the other half on ampicilin plus trimethoprim plates (50 μg/ml; Sigma, St Louis, MO), both with IPTG. Plating efficiencies were strictly comparable. Sequences were determined by dideoxy DNA sequencing using primer 7 (5′-TCTGCGTTCTGATTTAATC) and Sequenase 2.0 (USB).

N-terminal deletions

Deleted R67 DHFR genes were made by amplification from plasmid pSUR67 with five different primer pairs. Deletion 1 (Δ1) using primer 8 5′-CGG**GAGCTC**CACAACAAAGGAACCAAATGAGTAATCCAGTTGCTGG) and primer 6, Δ2 using primer 9 (5′-CGG**GAGCTC**CACAACAAAGGAACCAAATGGTATTCCCATCGAACGCCACGTTTG) and primer 6, Δ3 using primer 10 (5′-CGG**GAGCTC**CACAACAAAGGAACCAAATGTTTGGTATGGGAGATCGGTG) and primer 6, Δ4 using primers 10 and 11 (5′-CCG**GGATCC**AACACCCAACCACCAACTTAGCGTTCAAGCGCCGCAACAGG) and Δ5 with primers 5 and 11 (Figure 2A and C). The coding regions starting with the ATG are underlined. After digestion of the PCR products with *Sac*I and *Bam*HI and gel purification the PCR fragments were ligated to pTrc99A. Transformants were plated on amp or amp+tmp.

Reconstruction of full-length genes from Δ3R67

PCR products from individual Δ3R67 DHFRs or from a hypermutagenesis reaction were recombined with a DNA fragment that included the first 31 amino acid residues of the wild-type N-terminus by PCR in two steps (see Table II and Figure 2C and D). One nanogram of each purified fragment was mixed in a standard PCR mixture. No additional primers were added at this point. Cycling parameters were: 37 (30 s), 72 (30 s) and 95°C (30 s) for two cycles and 55 (30 s), 72 (30 s) and 95°C (30 s) for 5 cycles. Subsequently 50 pmol primers 5 and 6 were added and 15 additional PCR cycles [55 (30 s), 72 (30 s) and 95°C (30 s)] were carried out. The 93 bp DNA fragment (see Figure 2C) used in the above recombination experiment was constructed by PCR from the pSUR67 plasmid and primers 1 and 12 (5′-GCGCACGCGATCTCCC, complementary to positions 78–93). The completely mutated 93 bp N-terminus DNA fragment (see Figure 2D) was constructed by PCR from overlapping primers 13 (5′-CTGGCA**GAGCTC**CACAACAAAGGAACCAAATGAAACAAAATAATAGT), 14 (5′-TACAAGTTGACTAGTAATTAGATCATTAATTTTACTATTATTTTGTTTCAT), 15 (5′-ATTACTAGTCAACTTGTACTTAGTTTAAATACTACTTTTGGTATGGGAGATCGC) and 16 (5′-GCGATCTCCCATACCAAAAGT). One picomole of each oligonucleotide was added to a standard PCR mixture. Cycling parameters were: 55 (30 s), 72 (30 s) and 95°C (30 s) for 10 cycles. After 10 cycles 50 pmol primers 13 and 16 were added and 15 additional PCR cycles were performed.

Acknowledgements

We would like to thank Dr David Matthews for the co-ordinates of the 1.7 Å R67 DHFR structure prior to publication and Dr Thierry Rose for modelling and Dr Rupert Mützel for initial R67 constructs. Thanks to many colleagues for reading the manuscript, particularly John Holland, Esteban Domingo and Agnès Ullmann. This work was supported by grants from the Institut Pasteur and l'Agence Nationale pour la Recherche sur le SIDA (ANRS). M.A.M. was supported by fellowships from the European Community and the ANRS and V.P. by the Ministère de la Recherche et de l'Enseignement Supérieur.

References

Barbas,C.F., Bain,J.D., Hoekstra,D.M. and Lerner,R.A. (1992) Semisynthetic combinatorial antibody libraries: a chemical solution to the diversity problem. *Proc. Natl Acad. Sci. USA*, **89**, 4457–4461.

Barbas,C.F., Hu,D., Dunlop,N., Sawyer,L., Cababa,D., Hendry,R.M., Nara,P.L. and Burton,D.R. (1994) *In vitro* evolution of a neutralizing human antibody to HIV-1 to enhance affinity and broaden strain. *Proc. Natl Acad. Sci. USA*, **89**, 3809–3813.

Bowie,J.U., Reidhaar-Olson,J.F., Lim,W.A. and Sauer,R.T. (1990) Deciphering the message in protein sequences: tolerance to amino acid substitutions. *Science*, **247**, 1306–1310.

Brisson,N. and Hohn,T. (1984) Nucleotide sequence of the dihydrofolate-reductase gene borne by plasmid R67 and conferring methotrexate resistance. *Gene*, **28**, 271–275.

Caldwell,R.C. and Joyce,G.F. (1992) Randomization of genes by PCR mutagenesis. *PCR Methods Applic.*, **2**, 28–33.

Creighton,T.E. (1993) *Proteins: Structures and Molecular Properties.* W.H.Freeman and Co., New York.

Dao-Pin,S., Söderlind,E., Baase,W.A., Dahlquist,F.W. and Matthews, B.W. (1991) Cumulative site-directed charge–charge replacements in T4 lysozyme suggest that long-range electrostatic interactions contribute little to protein stability. *J. Mol. Biol.*, **221**, 873–887.

DeGrado,W.F., Wasserman,Z.R. and Lear,J.D. (1989) Protein design, a minimalist approch. *Science*, **243**, 622–628.

Doolittle,R.F., Feng,D.F., Johnson,M.S. and McClure,M.A. (1989) Origins and evolutionary relationships of retroviruses. *Q. Rev. Biol.*, **64**, 1–30.

Flensberg,J. and Steen,R. (1986) Nucleotide sequence analysis of the trimethoprim resistant dihydrofolate reductase encoded by R plasmid R751. *Nucleic Acids Res.*, **14**, 5933.

French,D.L., Laskov,R. and Scharff,M.D. (1989) The role of somatic hypermutation in the generation of antibody diversity. *Science*, **244**, 1152–1157.

Gram,I.L.A.M., Barbas,C.F., Colet,T.A., Lerner,R.A. and Kang,A.S. (1992) *In vitro* selection and affinity maturation of antibodies from naive combinatorial immunoglobulin library. *Proc. Natl Acad. Sci. USA*, **89**, 3576–3580.

Lehotovaara,P.M., Koivula,A.K., Bamford,J. and Knowles,J.K.C. (1988) A new method for random mutagenesis of complete genes: enzymatic generation of mutant libraries *in vivo*. *Protein Engng*, **2**, 63–68.

Loeb,D.D., Swanstrom,R., Everitt,L., Manchester,M., Stamper,S.S. and Hutchison,C.A. (1989) Complete mutagenesis of the HIV-1 protease. *Nature*, **340**, 397–400.

Martinez,M.A., Vartanian,J.P. and Wain-Hobson,S. (1994) Hypermutagenesis of RNA using human immunodeficiency virus type 1 reverse transcriptase and biased dNTP concentrations. *Proc. Natl Acad. Sci. USA*, **91**, 11787–11791.

Martinez,M.A., Sala,M., Vartanian,J.P. and Wain-Hobson,S. (1995) Reverse transcriptase and substrate dependence of the RNA hypermutagenesis reaction. *Nucleic Acids Res.*, **14**, 2573–2578.

Matthews,D.A., Smith,L.S., Baccanari,D.P., Burchall,J.J., Oatley,J.S. and Kraut,J. (1986) Crystal structure of a novel trimethoprin-resistant dihydrofolate reductase specified in *Escherichia coli* by R-Plasmid R67. *Biochemistry*, **25**, 4194–4204.

Maynard Smith,J. (1970) Natural selection and the concept of a protein space. *Nature (Lond.)*, **225**, 563–564.

Meyerhans,A., Vartanian,J.P. and Wain-Hobson,S. (1990) DNA recombination during PCR. *Nucleic Acids Res.*, **18**, 1687–1691.

Narayana,N., Matthews,D.A., Howell,E.A. and Xuong,N.-H. (1995) A plasmid-encoded dihydrofolate reductase from trimethoprim-resistant bacteria has a novel D_2-symmetric active site. *Nature Struct. Biol.*, **2**, 1018–1025.

Ninio,J. and Bokor,V. (1986) Stratégies d'adaptation moléculaire. *Comptes Rendus Acad. Sci. (La Vie des Sciences)*, **3**, 121–136.

Pakula,A.A., Young,V.B. and Sauer,R.T. (1986) Bacteriophage lambda *cro* mutations: effects on activity and intracellular degradation. *Proc. Natl Acad. Sci. USA*, **83**, 8829–8833.

Pathak,V.K. and Temin,H.M. (1990) Broad spectrum of *in-vitro* forward mutations, hypermutations, and mutational hotspots in a retrovital shuttle vector after a single replication cycle: substitutions, frameshifts, and hypermutations. *Proc. Natl Acad. Sci. USA*, **87**, 6019–6023.

Pattishal,K.H., Acar,J., Burchal,J.J., Goldstein,F.W. and Harvey,R.J. (1977) Two distinct types of trimethoprim-resistant dihydrofolate reductase specified by R-plasmids of different compatibility groups. *J. Biochem. Chem.*, **252**, 2319–2323.

Pjura,P., Matsumura,M., Baase,W.A. and Matthews,B.W. (1993) Development of an *in vivo* method to identify mutants of phage T4 lysozyme of enhanced thermostability. *Protein Sci.*, **2**, 2217–2225.

Reece,L.J., Nichols,R., Ogden,R.C. and Howell,E.E. (1991) Construction of a synthetic gene for an R-plasmid-encoded dihydrofolate reductase and studies on the role of the N-terminus in the protein. *Biochemistry*, **30**, 10895–10904.

Rennell,D., Bouvier,S.E., Hardy,L.W. and Poteete,A.R. (1991) Systematic mutation of bacteriophage T4 lysozyme. *J. Mol. Biol.*, **222**, 67–87.

Shortle,D. (1992) Mutational studies of protein structures and their stabilities. *Q. Rev. Biophys.*, **25**, 205–250.

Shortle,D. and Lin,B. (1985) Genetic analysis of staphylococcal nuclease: identification of three intragenic 'global' suppressors of nuclease-minus mutations. *Genetics*, **110**, 539–555.

Siderovski,D.P., Matsuyama,T., Frigerio,E., Chui,S., Min,X., Erfle,H., Sumner-Smith,M., Barnett,R.W. and Mak,T.W. (1992) Random mutagenesis of the human immunodeficiency virus type-1 trans-activator of transcription (HIV-1 Tat). *Nucleic Acids Res.*, **20**, 5311–5320.

Stemmer,W.P.C. (1994a) DNA shuffling by random fragmentation and reassembly. *In vitro* recombination for molecular evolution. *Proc. Natl Acad. Sci. USA*, **91**, 10747–10751.

Stemmer,W.P.C. (1994b) Rapid evolution of a protein *in vitro* by DNA shuffling. *Nature*, **370**, 389–391.

Swift,G., McCarthy,B.J. and Heffron,F. (1981) DNA sequence of a plasmid-encoded dihydrofolate reductase. *Mol. Gen. Genet.*, **181**, 441–447.

Vartanian,J.P., Meyerhans,A., Åsjö,B. and Wain-Hobson,S. (1991) Selection, recombination and G→A hypermutation of human immunodeficiency virus type 1 genomes. *J. Virol.*, **65**, 1779–1788.

Vartanian,J.P., Meyerhans,A., Sala,M. and Wain-Hobson,S. (1994) G→A hypermutation of the HIV-1 genome: evidence for dCTP pool imbalance during reverse transcription. *Proc. Natl Acad. Sci. USA*, **91**, 3092–3096.

Vermersch,P.S. and Bennett,G.N. (1988) Synthesis and expression of a gene for a mini type II dihydrofolate reductase. *DNA*, **7**, 243–251.

Zhou,Y.Z., Zhang,X. and Ebright,R. (1991) Random mutagenesis of gene-sized DNA molecules by use of PCR with Taq DNA polymerase. *Nucleic Acids Res.*, **19**, 6052.

Zolg,J.W. and Hänggi,U.J. (1981) Characterization of an R plasmid-associated, trimethoprim resistant dihydrofolate reductase and determination of the nucleotide sequence of the reductase gene. *Nucleic Acids Res.*, **9**, 697–710.

Received on October 10, 1995; revised on November 2, 1995