Proc. Natl. Acad. Sci. USA
Vol. 93, pp. 166–170, January 1996
Evolution

Evidence that two present-day components needed for the genetic code appeared after nucleated cells separated from eubacteria

(aminoacyl-tRNA synthetase/molecular evolution)

LLUÍS RIBAS DE POUPLANA*, MAGALÍ FRUGIER*, CHERYL L. QUINN†, AND PAUL SCHIMMEL*

*Department of Biology, Massachusetts Institute of Technology, Cambridge, MA 02139; and †Cubist Pharmaceuticals Inc., Cambridge, MA 02139

Contributed by Paul Schimmel, September 28, 1995

ABSTRACT **The trinucleotide/amino acid relationships of the present-day genetic code are established by the aminoacylation reactions of tRNA synthetases, whereby each of 20 specific amino acids is attached to its cognate tRNAs, which bear anticodon trinucleotides. Because of its universality, the appearance of the modern genetic code is thought to predate the separation of prokaryotic and eukaryotic organisms in the universal phylogenetic tree. In the light of new sequence information, we present here a phylogenetic analysis that shows an unusual picture for tyrosyl- and tryptophanyl-tRNA synthetases. In particular, the eukaryotic tyrosyl- and tryptophanyl-tRNA synthetases are more related to each other than to their respective prokaryotic counterparts. In contrast, each of the other 18 eukaryotic synthetases is more related to its prokaryotic counterpart than to any eukaryotic synthetase specific for a different amino acid. Our results raise the possibility that present day tyrosyl- and tryptophanyl-tRNA synthetases appeared after the separation of nucleated cells from eubacteria. The results have implications for the development of the genetic code.**

An open question concerning the origin of life is the mechanism of development of the genetic code and the role played by aminoacyl-tRNA synthetases (1). Because of their central role in linking amino acids with nucleotide triplets contained in transfer RNAs (2–6), aminoacyl-tRNA synthetases are thought to be among the first proteins that appeared in evolution. In agreement with this view of an ancient origin for the enzymes, all phylogenetic analyses carried out so far have determined that the sequences of tRNA synthetases cluster together by their amino acid specificities, and not by their positions in the phylogenetic tree (7–9). This observation clearly indicates that the enzymes appeared and evolved into their current types before the branching of the three Ur-kingdoms (eubacteria, archaebacteria, and eukaryotes). These previous analyses, however, were limited by the lack of (mainly) eukaryotic sequences for some of the enzymes. For instance, no eukaryotic tyrosyl-tRNA synthetase (YRS) or tryptophanyl-tRNA synthetase (WRS) had been analyzed. Here we present an analysis of the phylogenetic relationships between the prokaryote and eukaryote forms of this pair of enzymes.

Recent determination by Carter and coworkers (10) of the crystal structure of a eubacterial WRS (B-WRS) revealed a high degree of structural similarity between this enzyme and the eubacterial YRS (B-YRS) solved by Brick *et al.* (11), thus suggesting a surprisingly recent common ancestor not discernible through sequence analysis. This structural conservation is particularly remarkable considering the vast amount of time since both of these class I tRNA synthetases separated from their ancestor. No three-dimensional information is available for any eukaryotic WRS or YRS (E-WRS or E-YRS).

The publication costs of this article were defrayed in part by page charge payment. This article must therefore be hereby marked "*advertisement*" in accordance with 18 U.S.C. §1734 solely to indicate this fact.

During the process of analyzing E-YRS sequences recently obtained in our and other laboratories, we observed a striking distribution of sequence relationships between YRSs and WRSs. Data base search analysis with the BLAST algorithm (4) showed that E-YRS sequences had a higher degree of sequence identity with E-WRS—i.e., with synthetases specific for a different amino acid—than with sequences of prokaryotic enzymes specific for the same amino acid (B-YRS). Similarly, E-WRS showed higher identity to sequences of E-YRS than to the corresponding B-WRS sequences. This observation motivated us to pursue a more detailed analysis of the phylogenetic relationship of these two enzymes, which we describe below.

RESULTS

Analysis of Sequences for WRSs and YRSs. A first analysis of the above-mentioned sequence similarity pattern was carried out by multiple sequence alignments across the entire sequences of the YRS and WRS enzymes. The alignments were performed with the program PILEUP, from the University of Wisconsin Genetics Computer Group package (12). These alignments confirmed that the clusters of sequence similarities for the eubacterial and eukaryote YRS and WRS do not follow the amino acid specificity groups, which would be expected to include all WRSs together, for instance. Instead, dendograms built from the pairwise comparisons used for the alignments resulted in clustering of the enzyme sequences according to their eubacterial or eukaryote origin. (Consistent with the endosymbiotic origin of mitochondrial proteins, mitochondrial YRS and WRS group together with their respective eubacterial counterparts.) This peculiar relationship is exemplified by the alignment of one of the highly conserved sequence motifs characteristic of class I tRNA synthetases (Table 1). Five of 17 positions in this "KMSKS" consensus region are more conserved among synthetases for the two amino acids within the same kingdom than they are for synthetases for the same amino acid across kingdoms. There is no example of the opposite case.

Multiple sequence alignments of all WRS and YRS sequences were built with a variety of gap opening and extending penalties, and all those that correctly aligned the conserved class-defining HIGH and KMSKS motifs (7) were kept for further analysis. Preliminary phylogeny trees were constructed with all of these alignments by maximum parsimony methods as implemented in the program PROTPARS [all phylogeny programs used are included in the package PHYLIP (13)]. In every case the resulting phylogeny suggested a clustering of the enzymes on the basis of their eukaryotic or eubacterial nature, and not on the basis of their amino acid specificities.

For a more stringent analysis, a single alignment [in better agreement with previously published structure-based align-

Abbreviations: YRS, tyrosyl-tRNA synthetase; WRS, tryptophanyl-tRNA synthetase; other aminoacyl-tRNA synthetases are abbreviated similarly, using the single-letter amino acid symbols; B-, eubacterial; E-, eukaryotic.

Table 1. Sequences of YRSs and WRSs around the KMSKS region

Species	Synthetase	Sequence
N. crassa	Y mit.	V P L L T D S S G A K F G K S A - G N
P. anserina	Y mit.	V P L L T D S A G V K F G K S A - G N
B. caldotenax	Y	I P L V T K A D G T K F G K T E S G T
B. stearothe.	Y	I P L V T K A D G T K F G K T E S G T
B. subtilis	Y	I P L V T K A D G T K F G K T E G G A
E. coli	Y	V P L I T K A D - T K F G K T E G G A
B. stearothe.	W	M S L V D P T K - - K M S K S D P N P
B. subtilis	W	M S L N D P L K - - K M S K S D P N Q
E. coli	W	M S L L E P T K - - K M S K S D D N R
S. cerevisiae	W mit.	L S L S T P E K - - K M S K S D P N H
B. taurus	W	P A L Q G A Q T - - K M S A S D P N S
H. sapiens	W	P A L Q G A Q T - - K M S A S D P N S
O. cuniculus	W	P A L Q G A Q T - - K M S A S D P N S
M. musculus	W	P A L Q G A Q T - - K M S A S D P N S
S. cerevisiae	Y	P G - L A Q G G - - K M S A S D P N S
H. sapiens	Y	P G T L T G G G - - K M S S S D P N S
		* * * * * *

Alignment of the region around the KMSKS motif of all YRS and WRS sequences used in this study. Full species names are *Neurospora crassa, Podospora anserina, Bacillus caldotenax, Bacillus stearothermophilus, Bacillus subtilis, Escherichia coli, Saccharomyces cerevisiae, Bos taurus, Homo sapiens, Oryctolagus cuniculus,* and *Mus musculus.* Synthetase abbreviations: Y, YRS; W, WRS; mit., mitochondrial enzyme. Boxes indicate positions conserved only in the eubacterial or eukaryotic sequences. Asterisks indicate generally conserved residues.

ments (10, 14)] was used. Maximum parsimony methods (PROTPARS) and distance methods (KITSCH) were applied (13). Both methods indicated a clear phylogenetical trend with a separate origin for the eukaryotic YRS and WRS from their eubacterial counterparts. Bootstrapping analysis with the program SEQBOOT was used to analyze the robustness of the tree obtained for the YRS and WRS alignments (15). Bootstrapping analysis of 1000 variant data sets generated from the original alignment indicated that these evolutionary relationships are strongly favored over other possible trees (12) (Fig. 1).

A possible explanation for these results would be that the sequence relationships found in contemporary enzymes are the result of divergent coevolutionary forces induced by the interaction of the enzymes with the eubacterial and eukaryotic protein synthesis machineries. If that were the case, however, a similar distribution of sequence relationships should be expected among the rest of synthetases. When the sequences studied here were analyzed in the context of other class I tRNA synthetases, the YRS and WRS cluster was the only one not to show a clustering defined by amino acid specificity. An example is given in Fig. 2, where we present the phylogenetic tree obtained for the YRS and WRS cluster, together with glutaminyl- (Q), methionyl- (M), and arginyl- (R) tRNA synthetases. A similar analysis performed for all class I and II tRNA synthetases confirmed that all available sequences in either class of enzymes, except for WRS and YRS, cluster according to their amino acid specificities (data not shown).

It could be argued that the phylogenetic relationships found in this analysis are due to strong posterior evolution of the

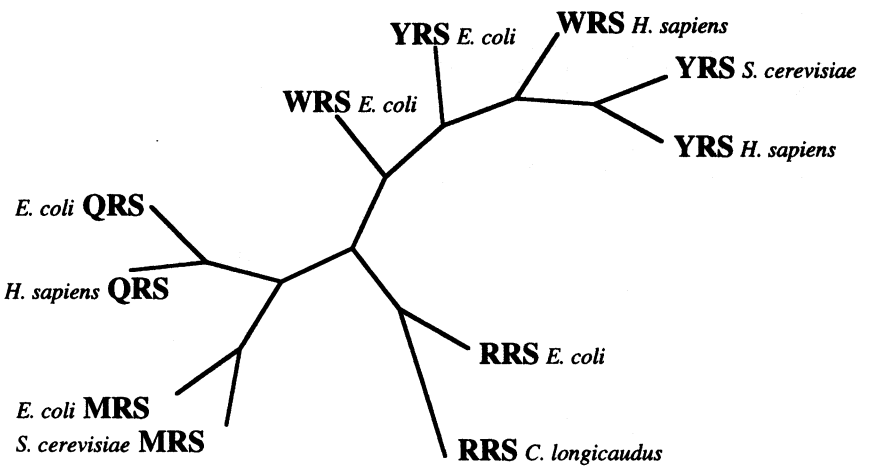

FIG. 1. Most parsimonious unrooted tree found for an alignment of several class I tRNA synthetase sequences between the HIGH and the KMSKS motifs (a region which encompasses around 190–230 residues, depending on the sequence considered). *C. longicaudus, Cricetulus longicaudus.*

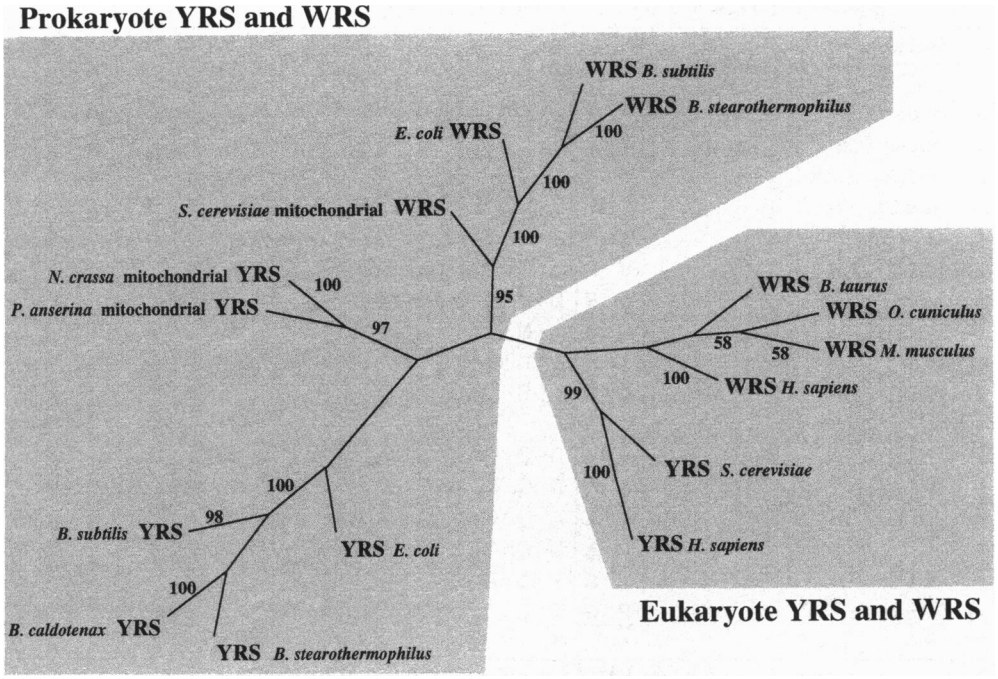

FIG. 2. Most parsimonious unrooted tree built from a bootstrap analysis of the WRS–YRS alignment (1000 data sets). Numbers at the branches correspond to percentage bootstrap frequencies for each particular branch. The alignment included 16 different sequences over their entire lengths. All phylogenetic analyses were carried out by using the package PHYLIP (13). All sequences used were available in the GenBank and SwissProt data bases except for the human YRS, which was determined by sequencing cDNA clones identified from a human cDNA library by C.L.Q. (GenBank accession no. U40714).

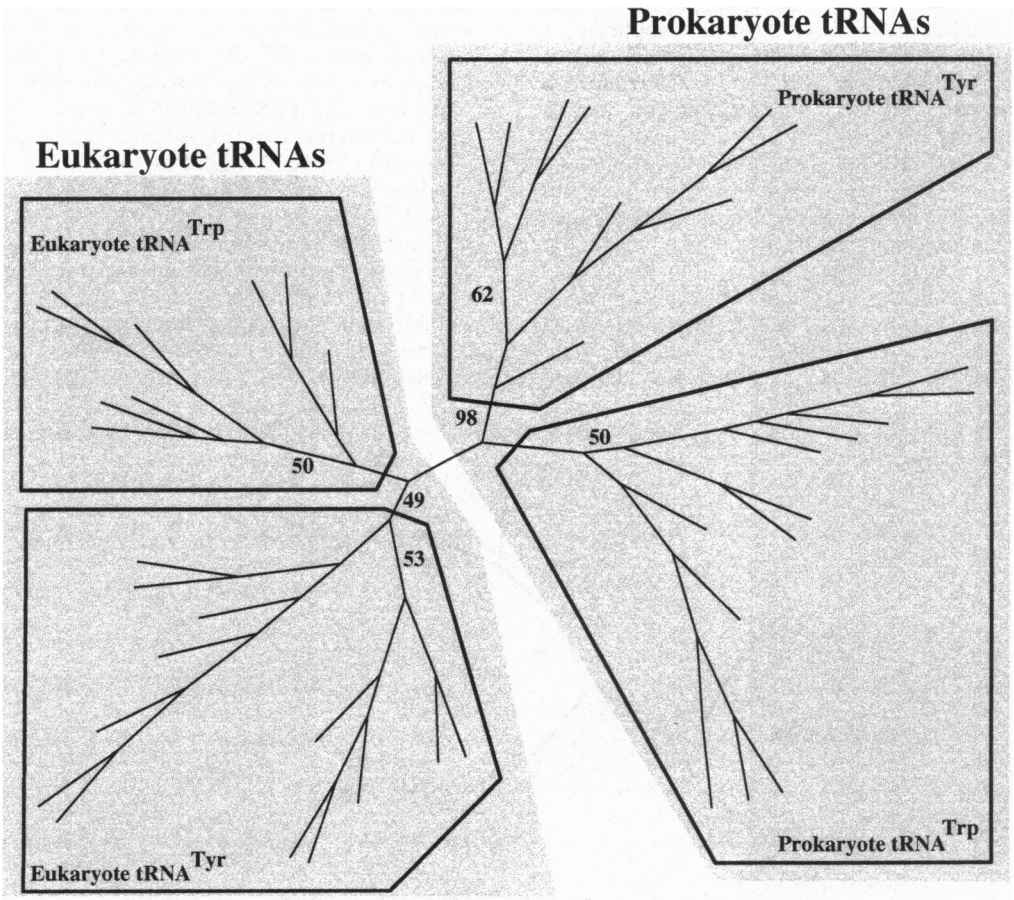

FIG. 3. Most parsimonious unrooted tree built from a bootstrap analysis (100 data sets) of an alignment of tryptophan and tyrosine tRNAs. Numbers at the initial branches of the tree correspond to bootstrap frequencies. The alignment included 43 different sequences over their entire lengths (all sequences are available in the GenBank data base). The correctness of these alignments was evaluated as a function of the correct positioning of the variable loop residues of eubacterial tyrosyl-tRNAs.

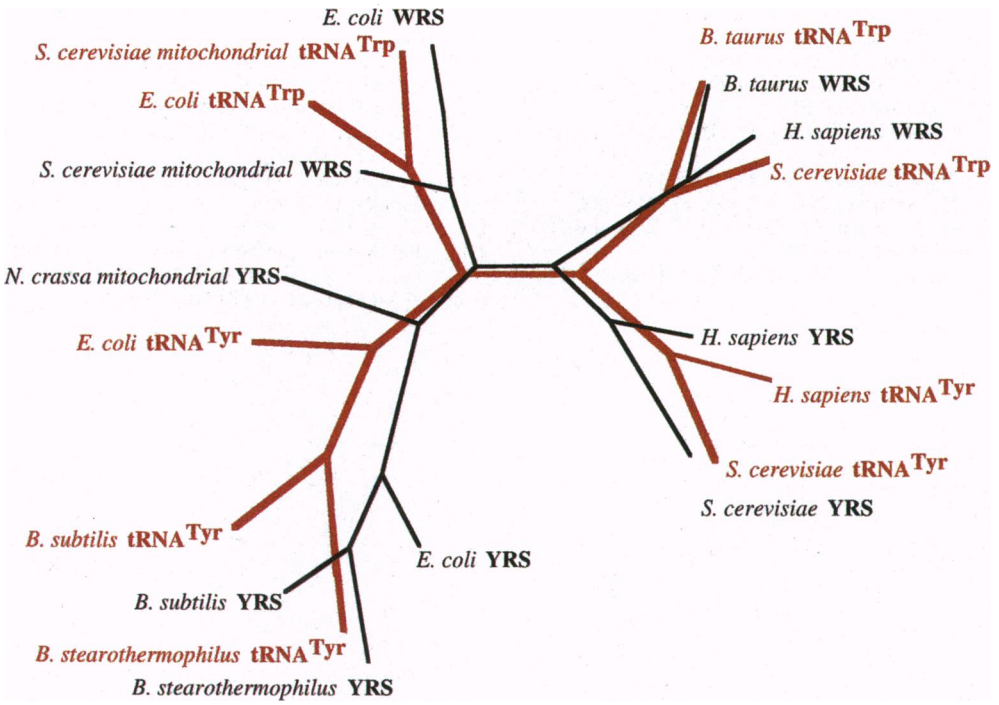

FIG. 4. Superimposition of a maximum parsimony unrooted tree of 10 YRSs and WRSs with the analogous tree obtained for their corresponding tRNA sequences.

anticodon recognition domains (13), which tend to be less conserved than the active-site domains of the enzymes and whose sequences could be biasing the evolutionary analysis. However, when the same analysis was repeated with the sequences between the HIGH and KMSKS motifs (which mark the conserved active-site domains of these enzymes) an identical phylogenetical distribution was found (data not shown). This result suggests that the complete sequences of eukaryotic YRS and WRS evolved separately and independently from their eubacterial equivalents.

Analysis of Sequences for $tRNA^{Trp}$ and $tRNA^{Tyr}$. If a strong coevolution exists between a tRNA synthetase and its cognate tRNAs, then the sequences of $tRNA^{Trp}$ and $tRNA^{Tyr}$ may be expected to show a similar sequence distribution. Indeed, the comparison of 43 sequences of eukaryotic and prokaryotic tyrosine and tryptophan tRNAs shows a clustering of sequence identities that groups sequence families around their evolutionary phyla, and not according to their cognate amino acids (Fig. 3). These analyses were carried out with the maximum sequence parsimony program DNAPARS and the distance methods program FITCH (13). The bootstrapping analysis for the tRNA trees obtained indicates that the data for these sequences are not as robust as those obtained for the proteins. This difference may be due to the smaller amount of evolutionary information provided by short nucleic acid sequences.

However, a clear coevolutionary process must exist between the two molecules in order to maintain molecular recognition despite genetic drift. Thus, the trend observed for the tRNA sequences may reflect a real phylogenetic situation. The superimposition of the most parsimonious trees for a number of YRSs and WRSs and a sequence of their cognate tRNAs is in agreement with the existence of such a coevolutionary process (Fig. 4).

DISCUSSION

Unfortunately, the lack of sequences from archaebacterial organisms prevents the analysis of the sequence relationships at the early stages of the prokaryote-eukaryote separation. Other examples of late evolutionary events concerning tRNA synthetases have been proposed for QRS and ERS, where horizontal genetic transfer of the QRS gene from eukaryotes to eubacteria has been postulated (16). In the case of YRS and WRS, however, a parallel transfer of genetic information would fail to account for the similarity patterns found in this study, because such patterns do not cross the evolutionary tree.

Our results do not explain why WRS and YRS are distinguished from all other synthetases in the way shown here. Nor should they be taken to mean that early eukaryotes necessarily had fewer than 20 tRNA synthetases or did not have at least a way to incorporate both tryptophan and tyrosine into proteins.

A possible explanation would be that, after the appearance of the ancestor eukaryotic cell, either one of the genes encoding the primitive YRS or WRS was lost in the eukaryotic branch. This could have been achieved by the replacement of the lost gene by a duplicated allele of the other gene. This theory would require a further explanation on the process by which a functional and essential enzyme is replaced by the duplication of another, functionally distinct, enzyme.

Alternatively, a single ancestral enzyme of YRS and WRS may have been able to interact with both amino acids and attach them selectively to their respective cognate tRNAs. This ancestor could have remained functional and, after the separation of prokaryotes from eukaryotes, have duplicated independently in both branches. The caveat of this theory is that it requires an improbable double duplication and divergence event.

Both scenarios, however, suggest a late existence of a highly dynamic genetic expression machinery which, at the time of the eukaryote–prokaryote divergence, was still capable of undergoing changes in its essential components.

We thank Profs. Charles Carter, Robert Cedergren, Joe Felsenstein, and Linda Gilmore for critical suggestions and comments. We are also grateful to Drs. James Brown and Douglas Buechter, and to Arturo Morales and Eric Schmidt, for helpful discussions. This work was supported by Grant GM 23562 from the National Institutes of Health and by a grant from the National Foundation for Cancer Research.

1. de Duve, C. (1991) *Blueprint for a Cell: The Nature and Origin of Life* (Neil Patterson, Burlington, NC), pp. 173–198.
2. Schimmel, P. (1987) *Annu. Rev. Biochem.* **56,** 125–158.
3. LaPointe, J. & Giegé, R. (1991) in *Translation in Eukaryotes*, ed. Trachsel, H. (CRC, Boca Raton, FL), pp. 35–69.
4. Cusack, S., Härtlein, M. & Leberman, R. (1991) *Nucleic Acids Res.* **19,** 3489–3498.
5. Moras, D. (1992) *Trends Biochem. Sci.* **17,** 159–164.
6. Carter, C. W., Jr. (1993) *Annu. Rev. Biochem.* **62,** 715–748.
7. Eriani, G., Delarue, M., Poch, O., Gangloff, J. & Moras, D. (1990) *Nature (London)* **347,** 203–206.
8. Nagel, G. M. & Doolittle, R. F. (1991) *Proc. Natl. Acad. Sci. USA* **88,** 8121–8125.
9. Brown, J. R. & Doolittle, W. F. (1995) *Proc. Natl. Acad. Sci. USA* **92,** 2441–2445.
10. Doublié, S., Bricogne, G., Gilmore, C. & Carter, C. W., Jr. (1995) *Structure* **3,** 17–31.
11. Brick, P., Bhat, T. N. & Blow, D. M. (1988) *J. Mol. Biol.* **208,** 83–98.
12. Devereaux, J., Haeberli, P. & Smithies, O. (1984) *Nucleic Acids Res.* **12,** 387–395.
13. Felsenstein, J. (1989) *Cladistics* **5,** 164–166.
14. Landès, C., Perona, J. J., Brunie, S., Rould, M. A., Zelwer, C., Steitz, T. A. & Risler, J. L. (1995) *Biochemie* **77,** 194–203.
15. Felsenstein, J. (1985) *Evolution* **39,** 783–791.
16. Lamour, V., Quevillon, S., Diriong, S., N'guyen, V. C., Lipinski, M. & Mirande, M. (1994) *Proc. Natl. Acad. Sci. USA* **91,** 8670–8674.