Convergence and constraint in eukaryotic release factor 1 (eRF1) domain 1: the evolution of stop codon specificity

Yuji Inagaki*, Christian Blouin, W. Ford Doolittle and Andrew J. Roger

Program in Evolutionary Biology, Canadian Institute for Advanced Research, Department of Biochemistry and Molecular Biology, Dalhousie University, Halifax, Nova Scotia B3H 4H7, Canada

Received August 27, 2001; Revised November 5, 2001; Accepted November 13, 2001 DDBJ/EMBL/GenBank accession nos +

ABSTRACT

Class 1 release factor in eukaryotes (eRF1) recognizes stop codons and promotes peptide release from the ribosome. The 'molecular mimicry' hypothesis suggests that domain 1 of eRF1 is analogous to the tRNA anticodon stem–loop. Recent studies strongly support this hypothesis and several models for specific interactions between stop codons and residues in domain 1 have been proposed. In this study we have sequenced and identified novel eRF1 sequences across a wide diversity of eukaryotes and re-evaluated the codon-binding site by bioinformatic analyses of a large eRF1 dataset. Analyses of the eRF1 structure combined with estimates of evolutionary rates at amino acid sites allow us to define the residues that are under structural (i.e. those involved in intramolecular interactions) versus non-structural selective constraints. Furthermore, we have re-assessed convergent substitutions in the ciliate variant code eRF1s using maximum likelihood-based phylogenetic approaches. Our results favor the model proposed by Bertram et al. that stop codons bind to three 'cavities' on the protein surface, although we suggest that the stop codon may bind in the opposite orientation to the original model. We assess the feasibility of this alternative binding orientation with a triplet stop codon and the eRF1 domain 1 structures using molecular modeling techniques.

INTRODUCTION

In the standard genetic code, 61 amino acid codons are recognized by tRNAs via the codon–anticodon interaction. The remaining three codons, UAA, UAG and UGA, are thought to be recognized by proteins called class 1 release factors (RFs) that terminate nascent protein synthesis (1–4). Bacteria use two class 1 RFs, RF1 and RF2, that likely resulted from an ancient gene duplication in the bacterial lineage. RF1 binds to UAA and UAG codons, while RF2 binds to UAA and UGA codons (1–4). Most eukaryotes and archaea each have single RF proteins, called eRF1 and aRF1, respectively, which recognize all three stop codons (2,5,6). aRF1 and eRF1 show obvious sequence similarity and thus the eukaryotic termination system is thought to have originated from an archaeal-like version in the common ancestor of eukaryotes and archaea (6). This view is strongly supported by the fact that aRF1 can catalyze the release of nascent peptides from eukaryotic ribosomes in response to UAA, UAG and UGA stop codons in vitro (5). Interestingly, sequence similarity between bacterial RFs and the eRF1/aRF1 family is difficult to detect; it is unclear whether the two RF families evolved from a single ancestral protein (2,6).

Extensive mutational studies using Escherichia coli have clearly indicated that three consecutive residues in RF1 and RF2, Pro-Ala-Thr and Ser-Pro-Phe, respectively, govern their codon specificities (7). Molecular studies of the eukaryotic and archaeal termination systems are less advanced and it is still unclear how eRF1 and aRF1 distinguish stop codons from codons specifying amino acids. Presumably, the stop codon-binding sites (and the catalytic centre promoting peptidyl-tRNA hydrolysis) are highly conserved, since accurate eRF1 activities are obviously essential for cell viability. An attractive hypothesis has been formulated in which class 1 RFs mimic tRNAs in terms of their functions and tertiary structures (1,3,8). Indeed, the recently determined human eRF1 structure resembles that of tRNA to a striking degree (9). This structural resemblance implies that eRF1 domains 1 and 2, which are highly conserved among standard code eRF1s (SC-eRF1s), correspond to the anticodon stem–loop and acceptor stem of tRNA, respectively (9).

The analogy between the anticodon stem–loop of tRNA and eRF1 domain 1 is further supported by genetic studies using yeast (10). Mutations in yeast eRF1 that abolish recognition of only one of the three stop codons (so called ‘unipotent’ suppressors) are located in domain 1. Molecular modeling by Bertram et al. of the binding of the trinucleotide UAA to human eRF1 has shown that the unipotent suppressor sites were located near surface cavities that could potentially bind the stop codon (10). Others propose, however, that it is the...
conserved residues in the α2-helix–loop–α3-helix region (henceforth referred to as α2–loop–α3) that form the stop codon-binding site (11–13). Either hypothesis is consistent with the structural similarity shared between eRF1 and tRNA and the fact that these conserved residues occupy a position on eRF1 that is analogous to the tRNA anticodon (11–13). Putative models for the interaction between eRF1 and stop codons are summarized in Table 1.

Domain 2 of eRF1 is believed to promote peptidyl-tRNA hydrolysis at the peptidyltransferase center in ribosomes. At the tip of domain 2 there is a strictly conserved motif, Gly-Gly-Gln (GGQ) (positions 183–185 in human eRF1). Genetic studies in yeast eRF1 showed that only Gln185 is essential for cell viability (9). On the other hand, human proteins with substitutions at positions 183 and 184 could not catalyze peptide release from ribosomes in vitro (14), while the release activity was not lost completely by substitutions at position 185 (15). Although these results from yeast and human eRF1s are incongruent in detail, the GGQ motif clearly plays a critical role in promoting peptidyl-tRNA hydrolysis. Interestingly, the mutations in human eRF1 affected neither ribosome binding nor its interaction with the GTPase subunit eRF3 (14). Domain 3, the least conserved of the three eRF1 domains, is thought to bind to eRF3, since deletions of this domain abolished the eRF1–eRF3 interaction (16–20).

Although translation must be accurately terminated at stop codons, it is well known that stop codons in the standard code have been re-assigned to code for several different amino acids in some bacterial and eukaryotic lineages (reviewed in 21). The ciliates (Ciliophora) are a unicellular eukaryotic group that includes several lineages with variant genetic codes (22,23). Furthermore, an extensive survey of ciliate gene sequences revealed that the classes Colpodea (e.g. Colpoda) and Heterotrichea (e.g. Blepharitricha) utilize the UGA codon for Trp, while UAR codons are retained as termination signals (UAR = *) (23). In ciliate groups with variant codes, the stop codon-binding site on eRF1 in domain 1 must have been altered to accommodate the variant stop codons. We have previously shown that the domain 1 sequences in variant code eRF1s (VC-eRF1s) from ciliates are indeed significantly diverged from those of SC-eRF1s (24). On the other hand, little sequence divergence was detected in domains 2 and 3 (24). Based on similar sequence comparisons between SC- and VC-eRF1s, conserved residues in the α2–loop–α3 region of domain 1 were independently nominated for the stop codon-binding site by two groups. The Thr-Ala-Ser tripeptide (TAS, positions 58–60 in human eRF1) was proposed by Nakamura and collaborators to be the discriminator region based on analogy to the 'peptide anticodon' in bacterial RFs (11). Muramatsu and collaborators, in contrast, hypothesized that a set of residues, Glu55, Gly57, Thr58, Ser60 and Asn61, discriminates stop codons from amino acid codons (13). While the two candidates for the codon-binding site overlap, the stop codon–eRF1 interactions proposed were different in detail. Furthermore, Lozupone et al. found that substitutions at still a third site in ciliate eRF1 domain 1 are shared by variant codes to the exclusion of other SC-eRF1 homologs (23). Interestingly some of these sites were identical to residues forming the putative stop codon-binding cavities proposed by Bertram et al. (10), but completely different from those proposed by Muramatsu et al. or Nakamura et al. (11,13).

Evolutionary comparisons like these should provide important clues to the molecular mechanism of eRF1 codon recognition, by identifying sites that are highly conserved in SC-eRF1s but different in VC-eRF1s. However, ‘highly conserved’ sites have so far been identified by considering a SC-eRF1 dataset that is phylogenetically restricted to relatively few eukaryotic groups (e.g. animals, fungi and plants). Sequences conserved amongst these organisms may in fact be variable when a more diverse set of eukaryotes is considered. Therefore, to determine which sites are absolutely conserved among SC-eRF1s, it
is crucial that sampling of the gene be extended to encompass the phylogenetically diverse protists. Data from a variety of protistan SC-eRF1s should permit a more precise definition of functional residues in eRF1, including the stop codon-binding site. Furthermore, it is important to compare VC-eRF1s that have independently evolved similar recognition characteristics. Amino acid assignments that are specifically and universally involved in altered codon ‘converged unvaried sites’ and will be key in identifying amino acid substitutions specifically involved in altered codon recognition.

We sequenced or identified the genes encoding SC-eRF1 from three amitochondrial protists [Trichomonas vaginalis (Parabasalia), Entamoeba histolytica (Entamoebidae) and Encephalitozoon cuniculi (Microsporidia), three fungi (Candida albicans, Neurospora crassa and Aspergillus nidulans), a green alga (Chlamydomonas reinhardtii) and two plants (Oryza sativa and Glycine max)]. Using an eRF1 dataset that included diverse protistan eRF1s, we re-assessed sites that substituted convergently in the eRF1s from ciliates with the UGA = * code (Oxytricha, Stylonychia and Tetrahymena) and the UAR = * code (Euplotids and Blepharisma). Combining evolutionary computational analyses of our eRF1 dataset and detailed analyses of the tertiary structure of human eRF1, we examined candidates proposed for the stop codon-binding site in eRF1 domain 1. Although our observations favor the location of the stop codon-binding site in eRF1 suggested by Bertram et al. (10), we suggest that the stop codon triplet associates with the codon-binding ‘cavities’ in the reverse orientation to their model.

MATERIALS AND METHODS

Novel eRF1 genes

The Neurospora and Apergillus eRF1 cDNA clones (a4b11 and y4b01, respectively) were provided by the Fungal Genetic Stock Center (University of Kansas Medical Center, Kansas). A cDNA clone (TV-0925) of Trichomonas vaginalis eRF1 was provided by T. M. Embley (Natural History Museum, London, UK). Chlamydomonas cDNA clones CM016g10_r, CM021h05_r and HCL030f11_r were provided by the Kazusa DNA Research Institute (Kisarazu, Chiba, Japan) (25). All clones were sequenced completely on both strands.

The entire Oryza eRF1 gene in clone P0504E02 (GenBank accession no. AP003269) and the Encephalitozoon eRF1 gene on chromosome V (GenBank accession no. AL590445) were identified with TBLASTN (26) against the GenBank non-redundant (nr) database. We also found five genome survey sequences (GSS) from Entamoeba that show strong similarity to eRF1s with TBLASTN in the Genome Survey Database at GenBank. The GSSs were assembled to form a contig covering almost the entire coding region of eRF1.

The putative amino acid sequence of the Glycine eRF1 was obtained from a contig assembled from more than 20 expressed sequence tag (EST) sequences. The eRF1 gene of Candida was identified with TBLASTN (26) and the DNA sequence was obtained from the Stanford DNA Sequencing and Technology Center website (http://www-sequencing.stanford.edu/group/candida). Sequencing of Calbicicans was accomplished with the support of the NIDR and the Burroughs Wellcome Fund.

Phylogenetic analyses

The putative amino acid sequences of the eRF1s that were newly determined or identified in this study were manually added to our previous eRF1 alignment (24). All sequences available as of June 2001 were used for the analyses. Ambiguously aligned sites and sites with a gap(s) were excluded from the initial alignment. The final dataset included 39 eRF1s with 339 aligned amino acid sites [henceforth referred to as the full-length (FL) dataset]. A maximum likelihood (ML) distance matrix was calculated from the FL dataset using TREE-PUZZLE v.4.0.2 (27) employing the JTT amino acid substitution model incorporating among-site rate variation [modeled by a gamma (Γ) distribution approximated with 16 discrete rate categories]. A phylogenetic tree based on this FL dataset was reconstructed from the JTT + Γ distance matrix using FITCH with global rearrangements implemented in PHYLIP v.3.573 (28). Branch lengths for optimal distance trees were estimated by the ML method using the JTT + Γ model implemented in TREE-PUZZLE v.4.0.2 (16 discrete Γ distribution categories). Each node in the distance tree was examined by bootstrap analyses. 500 resampled datasets were generated from the FL dataset using SEQBOOT in PHYLIP v.3.573 and subsequently JTT + Γ ML distance matrices were calculated using TREE-PUZZLE v.4.0.2 and PUZZLEBOOT v.1.02 (distributed by A. J. Roger and M. E. Holder; http://members.tripod.de/korbi/puzzle/) with parameter settings as above. Bootstrap proportions (BPs) were obtained using SEQBOOT, FITCH and CONSENSE in PHYLIP v.3.573. We also examined the tree topology by percent occurrence in 10 000 quartet puzzling trees (QP score), calculated using TREE-PUZZLE v.4.0.2. For all of the analyses above, a shape parameter α = 0.95 was used, obtained by ML estimation on a neighbor-joining tree using TREE-PUZZLE v.4.0.2.

We generated a data subset that included 103 sites in eRF1 domain 1 from the FL dataset. Since four vertebrate, Podospora and Neurospora and two Stylonychia, eRF1 domain 1 sequences are identical, redundant sequences were removed to save computational time. The partial eRF1 sequence of Blepharisma japonicum (GenBank accession no. AJ291710) was included, yielding a final dataset of 35 eRF1s [the domain-1 (D1) dataset]. The FITCH (Γ-FITCH) tree based on a JTT + Γ ML distance matrix for the D1 dataset was estimated with branch lengths optimized as described above. A shape parameter of α = 1.01 was re-estimated from the D1 dataset for this series of analyses.

A SC-eRF1 domain 1 dataset was generated from the D1 dataset by exclusion of 11 ciliate eRF1s. From this SC-eRF1 dataset, we used the conditional mode relative substitution rates of each site (site rates) derived using TREE-PUZZLE v.4.0.2 (27) with a discrete Γ distribution approximated with 16 categories (with shape parameter α = 0.89 estimated from the data). The calculated relative site rates were categorized into five classes and mapped on the tertiary structure of human eRF1 domain 1 (PDB accession no. 1DT9), with a color gradient from blue to red; the slowest evolving (blue; site rate < 0.2000), the second slowest evolving (light blue; 0.2000 < site rate < 0.5000), moderately evolving (white; 0.5000 < site rate < 0.7000), the second fastest evolving (dark blue; 0.7000 < site rate < 1.0000), and the fastest evolving (red; site rate ≥ 1.0000).
that seven sites were converged unvaried sites in those that have undergone convergent evolution in the ciliate lineages with the same codon specificity (UGA = *). Here we define a subset of such sites we call ‘converged unvaried’ sites to refer to those sites in which an identical residue is shared among the evolving residues that most likely stabilize the backbone fold of eRF1 domain 1. Slowest evolving residues failing to meet this criterion were deemed non-structural and are potentially responsible for RNA binding and/or codon specificity. All tertiary structures were visualized using VMD v.1.6.1 (29).

Identification of converged unvaried sites using site log-likelihood comparison Since the tree topologies inferred from the FL and D1 datasets (FL and D1 topologies) were incongruent (see Results and Discussion), log-likelihood (lnL) scores for individual amino acid alignment positions (site lnLs) were calculated from the D1 dataset under the FL and D1 topologies using the JTT + Γ model with eight discrete Γ rate categories (shape parameter = 1.01) using CODEML implemented in PAML v.3.0 (30). The FL topology was modified by adding B. japonicum to form a clade with Blepharisma americamum and excluding redundant sequences (three vertebrates, Podospora and Solychnia mytilus). By subtracting the site lnL scores under the FL topology from those under the D1 topology, a number of sites that favored the D1 topology were identified (i.e. those sites with positive ΔlnL scores).

We postulate that some fraction of these had undergone convergent evolution in the ciliate lineages with the same codon specificity (UGA = * or UAR = *). Here we define a subset of such sites we call ‘converged unvaried’ sites to refer to those sites in which an identical residue is shared among the eRF1s with the same codon specificity (i.e. UGA = * or UAR = *) but a different residue is present in their common ancestral sequence. Ancestral sequences of each node in the FL topology were reconstructed using the AAML option of CODEML implemented in PAML v.3.0 (30). The details of the analysis are described in the previous section. Our analyses indicated that seven sites were converged unvaried sites in Oxycircha + Stylochnia and Tetrahymena (UGA = *) or in Euplotes and Blepharisma (UAR = *) (see Results and Discussion). To test whether the presence of these seven converged unvaried sites in the alignment are sufficient to yield the anomalous relationship among ciliates in the D1 tree, FITCH trees were reconstructed from the D1 dataset: (i) excluding the seven converged unvaried sites; or (ii) excluding the seven sites with negative ΔlnL scores. Phylogenetic analyses were carried out as described above. For these calculations, shape parameters (α) were re-estimated for each new dataset.

Molecular modeling
Stop codon-binding was simulated using residues 28–132 of the human eRF1 structure and a trinucleotide 5′-UAA-3′. The unipotent suppressor data (10) and the converged sites data (above and 23) suggested that the trinucleotide in the domain 1 structure can be modeled in the opposite orientation to that proposed by Bertram et al. (10) (see detailed explanation below). Therefore, the first, second and third bases of the trinucleotide were roughly positioned within cavities 3, 2 and 1, respectively [Table 1; cavity numbering as per Bertram et al. (10)]. The binding conformational search was performed using AFFINITY (MSI, San Diego, CA) using a random docking complex search on a flexible ligand. The best complex was refined with a 50 step (0.1 ps/step) simulated annealing (500–300 K), using the force field cf91. Finally, 1000 iterations of minimization were performed on the bound complex.

RESULTS AND DISCUSSION
Characteristics of novel SC-eRF1s
Neurospora cDNA clone a4b11 is 2002 bp long and contains the entire eRF1 coding region. Aspergillus clade a4b01 (1426 bp) lacks a part of eRF1 domain 1, but, by assembly with EST sequence AA965980, we succeeded in recovering almost the entire coding region of the eRF1. Trichomonas cDNA clone TV-0925 appeared to contain almost the entire coding region of eRF1 (1291 bp long, 429 residues) with a 3′-untranslated region and poly(A) tail. Comparison of the putative amino acid sequence to other complete eRF1 sequences suggested that this cDNA probably lacks a very small portion of the N-terminus (data not shown). We sequenced three Chlamydodonas cDNA clones, but only HCL030f11_r covered the entire eRF1 coding region. The HCL030f11_r clone is 2568 bp long and includes a 1317 bp open reading frame (ORF). In the 3′-untranslated regions, four polymorphic sites were detected among the three clones (data not shown). The Entamoeba eRF1 gene occupies nearly two-thirds of the contig comprising five GSSs (AZ357755, AZ354941, AZ533867, AZ534936 and AZ540149; total 1762 bp). The identified coding region is 1241 bp long (413 residues) and the region corresponding to the C-terminal end of eRF1 (corresponding to the β15 sheet and the acidic amino acid tail in human eRF1) was truncated (data not shown). No obvious introns were identified in the Entamoeba eRF1 gene. The consensus DNA sequence from the Glycine eRF1 EST contig has an ORF of 1314 bp (437 residues). The Oriya eRF1 gene was identified in a BAC clone of chromosome I (GenBank accession no.: AP003269; positions 132930–134231). The gene is 1311 bp long (436 residues) and no introns were found. The Encephalitozon eRF1 gene (1161 bp long; 386 residues), without any introns, was identified between positions 49544 and 50704 in the chromosome V sequence (GenBank accession no. AL590445). The Candida eRF1 gene was found in contig 6-2330 (positions 12392–13639). The gene is 1311 bp long and encodes 346 residues without any introns.

eRF1 phylogeny
Phylogenetic relationships among eukaryotic groups were poorly resolved by the eRF1 tree based on the FL dataset. Only seven major nodes were supported by a BP >70% (Fig. 1A): (i) an animal clade (BP = 94%); (ii) a plant + fungi + Protist clade (BP = 96%); (iii) a kinetoplastid (Trypanosoma + Leishmania) clade (BP = 100%); (iv) an Aspergillus + Neurospora + Podospora clade (BP = 100%); (v) a Saccharomyces + Candida clade (BP = 97%); (vi) a Spirotrichea clade (Oxytricha, Stylochnia and Euplotes) (BP = 75%); (vii) a Spirotrichea + Tetrahymena clade (BP = 71%).
Although the ciliates as a whole were recovered as monophyletic in the optimal Γ-FITCH tree, this clade received low statistical support (BP = 40%; Fig. 1A). In contrast, a single clade for the Spirotrichea was not recovered in the tree based on the D1 dataset. Instead, the ciliate lineages with UGA = * codes (Oxytricha + Stylonychia and Tetrahymena) were clustered together (BP = 46%; Fig. 1B). In general, the resolution of this tree was worse than the FL dataset tree. We suspected that the aberrant relationships among ciliates in the D1 tree could result from sites in the D1 dataset that had independently converged on the same amino acid residues, conferring the same variant codon specificities of eRF1 on the independent ciliate lineages, such as Oxytricha + Stylonychia and Tetrahymena (UGA = *) or Euplotes and Blepharisma (UAR = *). To test this, we compared the lnL scores of domain 1 sites under the FL and D1 topologies by subtracting the former from the latter (Fig. 2A). This comparison indicates that 55 of 103 sites in domain 1 had higher lnL scores under the D1 topology than the FL topology.

It is important to note that the FL phylogeny for ciliates, although poorly supported, is in agreement with those based on sequences of small subunit and large subunit rRNA and features of cell ultrastructure (reviewed in 31), as well as macronuclear genome organization (32,33). Thus, the conclusion that, in ciliates, VC-eRF1s have arisen four separate times from SC-eRF1s (UGA = * code twice independently in Oxytricha + Stylonychia and Tetrahymena and UAR = * code twice independently in Euplotes and Blepharisma) rests on much more than the apparent phylogeny of the eRF1s themselves (22,23). The fact that the D1 dataset supports a topology that unites the two UAR = * code groups therefore likely reflects convergence in the eRF1 primary sequences, constrained by similarly altered codon recognition requirements. Amino acid position assignments that are identical within either the ciliate groups with the UAR = * or UGA = * code, and different from inferred ancestral assignments for ciliates (based on our expanded SC-eRF1 dataset), are therefore good candidates for involvement in codon recognition. We call these ‘variant code-specific converged and then unvaried sites (converged unvaried sites)’. Variant code-specific converged unvaried sites in ciliate eRF1

Since the FL and D1 trees differed in the placement of SC-eRF1s as well the branching order of ciliate eRF1s (Fig. 2A and B), only a subset of the 55 ‘D1 topology-favoring’ sites correspond to ciliate variant code-specific converged unvaried sites. To identify the latter sites, we filtered out 48 positions where identical residues were not conserved within the eRF1s corresponding to the UGA = * or UAR = * code. This left seven sites, four of which were identical among the UGA = * code ciliates (Oxytricha + Stylonychia and Tetrahymena) and three that were shared among UAR = * code ciliates (Euplotes and Blepharisma) (Fig. 3). We used ML reconstruction of ancestral sequences of the ciliate eRF1s at nodes 1–3 in the FL topology (Fig. 1A) to confirm that these seven sites were indeed lineage-specific convergent substitutions within the UGA = * or UAR = * code groups (Fig. 3). Five of the sites we identify as convergent were previously identified by Lozupone et al. (23) (boxed and marked with closed circles, Fig. 3). However, our method identifies two novel converged unvaried sites in the UAR = * eRF1s (boxed, Fig. 3). In no cases were the converged unvaried sites in the two variant code groups located at the same alignment position (Fig. 3).

A Γ-FITCH tree based on the D1 dataset with the seven converged unvaried sites removed no longer grouped the

---

**Figure 1.** (A) An unrooted eRF1 tree based on the FL dataset using the Γ-FITCH protein distance method. Abbreviations of the sequence names are listed in Table 2. BP are listed for major nodes. Percent occurrence in 10 000 quartet puzzling trees are given in brackets. Dashes indicate that given nodes were not supported by bootstrap or quartet puzzling analyses. The novel sequences determined or identified in this study are underlined. The redundant sequences removed from the analyses using the D1 dataset are marked by arrows. The ciliate lineages that independently converted the standard code to variant codes [UGA = * (UAR = Gln) and UAR = * (UGA = Cys or Trp)] are highlighted by thick black and gray lines, respectively. (B) An unrooted eRF1 tree based on the D1 dataset using the Γ-FITCH protein distance method.
UGA = * code ciliates together, instead recovering a monophyletic Spirotrichea clade (Fig. 2B). In contrast, a tree from the D1 dataset with the seven negative $\Delta$lnL scores removed (marked by open circles in Fig. 2A) once again recovered the UGA = * code ciliates as a group (Fig. 2C). These results indicate that the seven sites identified by our site lnL comparisons and ancestral sequence reconstructions as convergently substituted among the UGA = * or UAR = * code lineages are sufficient to cause the phylogenetically anomalous relationships among ciliates in the D1 tree.

Nine of the ‘code-specific’ substitutions detected by Lozupone et al. (23) (positions 60, 73, 86, 99, 106–107, 126 and 132) are not identified as converged unvaried sites by our method (marked with open circles in Fig. 3). For instance, positions 60 and 86 had negative $\Delta$lnL scores, indicating that they favor the FL topology and disqualifying them from consideration by our criteria. At position 60 the Tetrahymena sequence has a Thr, whereas the Oxytricha and Stylonychia sequences have a Gln. Position 86 is not conserved between UGA = * code ciliate sequences and is unlikely to be necessary for UGA = * code specificity, since the Gln in the Tetrahymena sequence is chemically more similar to the ancestral Asn, found in the SC-eRF1s, than the Lys and Arg residues with long and basic side chains found in the Oxytricha and Stylonychia sequences.

Table 2. eRF1 sequences used in this study

<table>
<thead>
<tr>
<th>Species</th>
<th>Abbreviations</th>
<th>Stop codon(s)</th>
<th>DDDBJ/EMBL/GenBank accession</th>
</tr>
</thead>
<tbody>
<tr>
<td>Homo sapiens</td>
<td>Homsap</td>
<td>Univ†</td>
<td>U90176</td>
</tr>
<tr>
<td>Xenopus laevis</td>
<td>Xenlae</td>
<td>Univ†</td>
<td>Z14253</td>
</tr>
<tr>
<td>Oxytrilugus ciculiculus</td>
<td>Orycin</td>
<td>Univ†</td>
<td>AB29089</td>
</tr>
<tr>
<td>Mesocruciscus auratus</td>
<td>Meaur</td>
<td>Univ†</td>
<td>X81626</td>
</tr>
<tr>
<td>Caenorhabditis elegans</td>
<td>Caeele</td>
<td>Univ†</td>
<td>A016452</td>
</tr>
<tr>
<td>Drosophila melanogaster</td>
<td>Dromel</td>
<td>Univ†</td>
<td>AB03391</td>
</tr>
<tr>
<td>Polysandricarpus misakiensis</td>
<td>Polmis</td>
<td>Univ†</td>
<td>AB053102</td>
</tr>
<tr>
<td>Saccharomyces cerevisiae</td>
<td>Sacche</td>
<td>Univ†</td>
<td>Z3612</td>
</tr>
<tr>
<td>Schizosaccharomyces pombe</td>
<td>Schpoon</td>
<td>Univ†</td>
<td>D63883</td>
</tr>
<tr>
<td>Candida albicans</td>
<td>Canalb</td>
<td>Univ†</td>
<td></td>
</tr>
<tr>
<td>Podospora anserina</td>
<td>Podans</td>
<td>Univ†</td>
<td></td>
</tr>
<tr>
<td>Aspergillus nidulans</td>
<td>Asnpn</td>
<td>Univ†</td>
<td>AY050664</td>
</tr>
<tr>
<td>Neurospora crassa</td>
<td>Neucra</td>
<td>Univ†</td>
<td>AY050665</td>
</tr>
<tr>
<td>Pneumocystis carinii</td>
<td>Pneear</td>
<td>Univ†</td>
<td>AB052893</td>
</tr>
<tr>
<td>Eucaphaloizoon cuniculi</td>
<td>Exoan</td>
<td>Univ†</td>
<td>AL500445</td>
</tr>
<tr>
<td>Arabidopsis thaliana</td>
<td>Arath1/2/3</td>
<td>Univ†</td>
<td>X69375/1U40217/AC012187</td>
</tr>
<tr>
<td>Oryza sativa</td>
<td>Orysat</td>
<td>Univ†</td>
<td>A003269</td>
</tr>
<tr>
<td>Glycine max</td>
<td>Glymax</td>
<td>Univ†</td>
<td>EST contig</td>
</tr>
<tr>
<td>Chlamydomonas reinhardtii</td>
<td>Chlre</td>
<td>Univ†</td>
<td>AY050666</td>
</tr>
<tr>
<td>Plasmodium falciparum</td>
<td>Plafal</td>
<td>Univ†</td>
<td>A001402</td>
</tr>
<tr>
<td>Giardia lamblia</td>
<td>Gialam</td>
<td>Univ†</td>
<td>AF98107</td>
</tr>
<tr>
<td>Dictyostelium discoideum</td>
<td>Dicdis</td>
<td>Univ†</td>
<td>AF298834</td>
</tr>
<tr>
<td>Leishmania major</td>
<td>Lemail</td>
<td>Univ†</td>
<td>AL161416</td>
</tr>
<tr>
<td>Trypanosoma brucei</td>
<td>Tybru</td>
<td>Univ†</td>
<td>AF273872</td>
</tr>
<tr>
<td>Trichomonas vaginalis</td>
<td>Trivag</td>
<td>Univ†</td>
<td>AY050667</td>
</tr>
<tr>
<td>Entamoeba histolytica</td>
<td>Enthis</td>
<td>Univ†</td>
<td>GSS contig</td>
</tr>
<tr>
<td>Nucleomorph in Guillardia theta</td>
<td>Nm._Guilthe</td>
<td>Univ†</td>
<td>AF165518</td>
</tr>
</tbody>
</table>

†The standard genetic code, where UAA, UAG and UGA serve as termination signals.
‡R for A or G and UGA for Cys.
§R for A or G and UGA for Trp.
¶Refer to the Stanford Genome Technology Center website (http://www-sequence.stanford.edu/group/candida). The DNA sequences of the E.histolytica GSS and G.max EST contigs and the putative amino acid sequences are available on request from Y.I. The novel eRF1 sequences determined/identified in this study are shown in bold.

Residues at positions 73, 99, 126 and 132 have positive $\Delta$lnL scores (Fig. 2A) but can be excluded from consideration for
other reasons. At position 99, the residue found in the *Tetrahymena* eRF1 (Lys) is chemically and structurally dissimilar to the amino acid found in the *Oxytricha* and *Stylonychia* sequences (Val) (Fig. 3). At position 73, the Thr with a polar side chain in the *Oxytricha* and *Stylonychia* sequences is not similar to the negatively charged Asp in the *Tetrahymena* eRF1 (Fig. 3). Position 126 in *Tetrahymena* and the *Oxytricha* + *Stylonychia* group appears to represent a convergent substitution in these two lineages, as Lozupone et al. argued (23). However, the *Oxytricha nova* sequence has an Asp at this position, which is chemically and structurally dissimilar to the Phe in the *Tetrahymena*, *Oxytricha trifallax* and *Stylonychia* sequences, questioning its absolute necessity for conferring the UGA = * code specificity (Fig. 3). At position 132, where a His is strictly conserved among SC-eRF1s, the *Oxytricha*, *Stylonychia* and *Euplotes* sequences (Spirotrichea) have various residues. However, the *Tetrahymena* and *Blepharisma* sequences retain the His residue from the ancestral sequences (Fig. 3). Thus the substitutions found at this position in the Spirotrichea eRF1s are unlikely to contribute to their variant code specificities.

Finally, the insertion of a single amino acid (Thr or Cys) between positions 106 and 107 is found only in the *Oxytricha*, *Stylonychia* and *Blepharisma* eRF1s, whereas neither the *Tetrahymena* nor *Euplotes* proteins share this insertion (Fig. 3). Clearly, such insertions cannot be required for variant code specificity in the UGA = * or UAR = * eRF1s (a ΔlnL score is not available, as the missing data mandated its exclusion from the site lnL analyses).

**Mapping the converged unvaried sites on the human eRF1 domain 1 structure**

We postulated that some or all of the convergent sites we identified could account for the alteration of codon specificities in the ciliate eRF1s, responding to the UGA = * or UAR = * code. However, mapping these sites onto the tertiary structure reveals that they are widely spread across eRF1 domain 1 (Fig. 4A). There appears to be no general correlation between the convergent substitutions and the estimated site rates at a given site, although positions 35, 51, 57, 70 and 126 are located within or near the cluster of slowest evolving residues (Fig. 4B and C). Of most interest are positions 51 (Met → Leu in the UGA = * eRF1s) and 126 (Met → Ile in the UAR = * eRF1s); these sites map to cavities 1 and 2 on the protein surface, respectively, identified by Bertram et al. as potential stop codon-binding ‘cavities’ (10). Postulated functions of the slowest evolving and converged unvaried sites are summarized in Table 3.

Another interesting convergent substitution is Ser → Ala at position 70 in the UAR = * eRF1s (Figs 3 and 4A). In the human structure, the two slowest evolving residues, Ser70 and Ser33, are predicted to form a potential hydrogen bond (Table 3). The ciliate VC-eRF1s with this substitution retain a Ser at position 33 (Fig. 3) and so would not be able to form a hydrogen bond. This substitution may not have a direct mechanical impact on codon binding, but it could alter the codon specificity of eRF1 by indirectly affecting the flexibility of the stop codon-binding site.

Indeed, the structural rationale behind many of the converged unvaried sites is likely to extend beyond the postulated function of codon recognition. As discussed above, only a few residues are directly involved in the cavity-binding model proposed by Bertram et al. (10). However, many of the convergent changes (including that at position 70) are more likely to reflect group-specific constraints on what residues are involved in stabilizing the structure. This is consistent with the observation that many of the converged unvaried changes are occasionally observed in SC-eRF1 homologs from other organisms (see for example positions 35, 70 and 126 in Fig. 3).

**Revised cavity-binding model**

The pattern of site rates mapped on the human eRF1 structure (Fig. 4B) sensibly corresponds to the general area of binding of the stop codon in the model proposed by Bertram et al. (10). This large number of slowest evolving residues comprises a structural (and hydrophobic) core that is likely to position the side chains whose function it is to recognize the stop codons...
Four of the solvent-accessible (and slowest evolving) residues, Glu55, Tyr125, Cys127 and Asp128, have the added property that they are located near conserved cavities on the protein surface (Fig. 4C) suggested to specifically bind the nucleotides of the stop codons on mRNA in the cavity-binding model proposed by Bertram et al. (10). Assuming that eRF1 interacts with the ribosome in a similar fashion to a tRNA molecule, the distance between the codon and the elongation peptide chain should be similar to the preceding tRNA already in the ribosome. Thus we measured the distances between the anticodon and the CCA end based on the *Saccharomyces cerevisiae* phenylalanine tRNA (PDB accession no. 1I9V) and between Gln185 in eRF1 domain 2 (which is thought to be analogous to the 3′-end of tRNAs) and various slowest evolving residues in domain 1 (Table 4). Interestingly, the distance between Gln185 and Tyr125 in cavity 2 (domain 1) is 84 Å, roughly matching the distance between the anticodon and the CCA end of 75–80 Å in the tRNA (Table 4).

In the cavity-binding model of Bertram et al. (10), residue Met51 is part of cavity 1, where it is thought to bind the first nucleotide of stop codons (Table 1). However, it is hard to explain how the Met→Leu convergence found in the UGA = *eRF1s (Fig. 3), and Met→Ile in a UAG unipotent suppressor mutant (10), selectively forbid the binding of a G at the third position of the stop codon. Furthermore, cavity 3, which was proposed to originally bind with the third nucleotide (10), is lined with absolutely conserved residues among SC-eRF1s and ciliate VC-eRF1s (Glu55, Tyr125 and Cys127; Fig. 3). These results suggest to us that the three nucleotides of the stop codon and the three cavities on the eRF1 surface could interact in an opposite orientation to that originally proposed (10); cavities 1 and 3 could account for discrimination of the third and first nucleotides of stop codons, respectively [Table 1; cavity numbering as per Bertram et al. (10)].

To test this, we modeled binding of the UAA trinucleotide such that the first nucleotide is inserted in cavity 3, the second nucleotide still interacts with cavity 2, and the third nucleotide is placed in cavity 1. In this conformation, the mRNA backbone is extended along a surface groove in which the proposed binding site is located (Fig. 5). The first nucleotide, U, is held in place by absolutely conserved Glu55 and Tyr125 residues, while its sugar moiety interacts with Cys127. The second position is tucked into cavity 2, with matching Asp128 and His132. Finally, the third position is located in the least evolutionarily conserved region of the cavity.
in principle, bind to mRNA. These results also support the two interacting with other slowest evolving residues and so could, structural' by our criteria (Fig. 4C). These residues are probably not Arg65, are slowest evolving and are not classified as 'struc-

Although currently available data (this study; 10,23) fit stop codon recognition? the cavity-binding models, a set of residues in the

Our cavity-binding model shown in Figure 5 also rationalizes why a higher aliphatic branching at position 51 (i.e. the Met→Leu convergence in UGA = * ciliates and the Met→Ile change in the unipotent suppressor mutation) would prevent binding of a G by steric hindrance of the amino group at C2 of the purine ring. It provides a structural basis for binding of the first nucleotide (which apparently never changes amongst SC or VC organisms) in a patch of absolutely conserved residues and positions the three nucleotides within the surface cavities without imposing a strain on the mRNA backbone. However, it is currently impossible to definitively decide whether the model presented in Figure 5 or that proposed by Bertram et al. (10) has the correct codon-binding orientation. Based on what is known about mRNA–tRNA interaction in the ribosome (34), the binding groove is located in the face of domain 1 that is distal to the 5′-end of the mRNA. Thus, the mRNA has to wrap around the release factor in either cavity-binding model; only the direction of the wrap would differ.

Is the α2-helix-loop–α3-helix region directly involved in stop codon recognition?

Although currently available data (this study; 10,23) fit well with the cavity-binding models, a set of residues in the α2–loop–α3 region, Gln185, Thr58, Ser60, Lys63, Ser64 and Arg65, are slowest evolving and are not classified as ‘structural’ by our criteria (Fig. 4C). These residues are probably not interacting with other slowest evolving residues and so could, in principle, bind to mRNA. These results also support the two models proposed by Muramatsu et al. (13) and Nakamura et al. (11), which suggest that the side chains of these residues could interact with stop codons directly (Table 1). However, the distances between Gln185 in eRF1 domain 2 and the slowest evolving residues in the α2–loop–α3 region are ~100 Å, much larger than the estimated distance based on the tRNA structure (~80 Å) or the cavity-binding models (~84 Å) (Table 4). Further crystallographic studies are needed to examine whether the tertiary structure of eRF1 bound with the GTPase subunit, eRF3, or in the A site of the ribosome might be different from that determined by Song et al. (9).

Other lines of evidence also cast doubt on involvement of the α2–loop–α3 region in stop codon recognition. For instance, a recent in vitro study showed that the Methanococcus jannaschii aRF1 with Gln at position 58 (rather than the Thr conserved amongst eukaryotes) is capable of terminating protein synthesis on human ribosomes responding to UAA, UAG and UGA codons (5). Similarly, Ser60 has changed to Thr in the Entamoeba SC-eRF1 (Fig. 3). These results indicate that positions 58 and 60 are changeable (at least to similar amino acids) without generating a concomitant change in the codon specificity of SC-eRF1. No substitution at position 59 has been found in SC-eRF1s before this study and Ala59 is a part of the potential ‘peptide anticodon’ in the model proposed by Nakamura et al. (11). However, this residue is substituted by Val in the SC-eRF1 from Aspergillus (Fig. 3). Our examination of the human eRF1 structure indicates that the Ala59 side chain points into and forms part of the slowest evolving domain 1 core (Fig. 4C). Thus it is likely that the high degree of conservation of Ala at this position is not necessary for eRF1 codon recognition. Instead, the residue is probably constrained to maintain the conformation of domain 1.
The combination of site rate analysis and structural observations of eRF1 domain 1 defined Ile62 as one of the slowest evolving residues in the domain 1 core (Fig. 4B). Other slowest evolving residues, Asn61, Lys63 and Ser64, adjacent to Ile62, are all polar or positively charged and solvent exposed (Fig. 4C), suggestive of a potential role as a ‘peptide anticodon’. In fact, position 64 was once suggested to be the eRF1 discriminator residue, since the *Tetrahymena* sequence, the only VC-eRF1 available at the time, possessed Asp (35). Yet this hypothesis was rejected, because other ciliate VC-eRF1s have the regular Ser at this position (Fig. 3) (23). Additionally, our novel SC-eRF1 from *Trichomonas* has the substitution, Ser64→Asn (Fig. 3), suggesting that Ser at position 64 is unlikely to be involved in eRF1 codon specificity.

Given the novel eRF1s determined/identified in this study and the *in vitro* experiments with the *M.jannaschii* aRF1 (5), it appears that at least one of the residues at positions 58–60 and 64 is substitutable to UAA, UAG or UGA without loss of codon specificity. It is intriguing how divergent this suite of residues can be, since most aRF1s have more than one
substitution at these positions as compared with the sequence conserved in most SC-eRF1s (TASNIKS) (24). Particularly, the aRF1s from *Aeropyrum pernix* (GenBank accession no. AP000063, ITDNIKL; versus the standard TASNIKS motif), *Sulfolobus solfataricus* (GenBank accession no. AE006836, IAQNIKL) and *Methanobacterium thermautotrophicum* (GenBank accession no. AE000666, QSA NIKS) are extremely divergent in this region. Yet, the high degree of conservation of the α2–loop–α3 region amongst eukaryotic class 1 RFs suggests that this region has an important function. Interestingly, we have noted that yeast ribosomal protein L30 (RPL30; PDB accession no. 1CK8) appears to have an analogous tertiary structure to human eRF1 domain 1 (despite having a different connectivity of secondary structural elements) by direct comparison of three-dimensional protein structure with the VAST algorithm (36,37). In this case, residues in the helices of RPL30 that are analogous to the eRF1 α2 and α3 helices interact with, and bind to, the rRNA molecule (38). eRF1 works at the A site of ribosomes and it is likely to interact with rRNA and ribosomal proteins adjacent to the A site during the translation termination process. Extrapolating the rRNA–RPL30 interaction to human eRF1 domain 1, one may speculate that the side chains of slowest evolving residues in the α2–loop–α3 region may interact with a part of small subunit rRNA (Table 3). Further *in vitro* experiments combining various aRF1s and eukaryotic ribosomes as well as large-scale sequencing of eRF1 genes will help elucidate the function of the residues in the α2–loop–α3 region.

### Table 4. Dimensional comparisons between tRNA and eRF1

<table>
<thead>
<tr>
<th>Molecule (PDB acc no)</th>
<th>Distance* (Å)</th>
</tr>
</thead>
<tbody>
<tr>
<td>eRF1 (1DT9)</td>
<td></td>
</tr>
<tr>
<td>Gln184 to Thr185</td>
<td>93</td>
</tr>
<tr>
<td>Gln184 to Ser188</td>
<td>96</td>
</tr>
<tr>
<td>Gln184 to Asn189</td>
<td>98</td>
</tr>
<tr>
<td>Gln184 to Lys190</td>
<td>101</td>
</tr>
<tr>
<td>Gln184 to Ser193</td>
<td>101</td>
</tr>
<tr>
<td>Gln184 to Glu194</td>
<td>85</td>
</tr>
<tr>
<td>Gln184 to Ser197</td>
<td>73</td>
</tr>
<tr>
<td>Gln184 to Tyr200</td>
<td>84</td>
</tr>
<tr>
<td>Gln184 to Asp203</td>
<td>82</td>
</tr>
<tr>
<td>tRNA (19P)</td>
<td></td>
</tr>
<tr>
<td>CCA-5′ end to anticodon, 2′′ nucleotide</td>
<td>79–80</td>
</tr>
</tbody>
</table>

*Measured from the Cα of Gln185 to the tip of the target side chain.
†Part of the GGQ motif in eRF1 domain 2.
‡In the α2-helix–loop–α3-helix region in domain 1.
§In cavity 2 in domain 1. Cavity numbering as per Bertram *et al.* (10).

---

Figure 5. Stereoview of stop codon 5′-UAA-3′ (blue) bound to the human release factor eRF1 N-terminal domain (purple) in the reverse orientation to the previously proposed model (10). Only the backbone atoms of eRF1 and the side chain interacting with the trinucleotide are displayed. Green dotted lines indicate possible dipolar interactions conferring binding specificity to the stop codon.
CONCLUSION

Are there multiple ways to change the codon specificity of eRF1?

As discussed above, both cavity-binding models, regardless of the trinucleotide orientation, appear to explain the data better than the anticodon mimicry models between the tRNA anticodon loop and the eRF1 o2–loop–o3 region. If either of the cavity-binding models were true, variant codon specificity would be modulated by (i) overall shapes of the codon-binding cavities and/or (ii) interactions between stop codon nucleotides and residues that are adjacent to the cavities. It is known that diplomonads and dasyycladacean green algae have converted the standard code to the UGA = * code independently (39–41). Furthermore, independent code conversions (from the standard to UGA = * code) from those in either the Oligohymenophorea or Spirotichea have been reported in other ciliate lineages [i.e. the classes Nassophorea and Karyorelictida and Condylosoma (Heterotricha)] (23). Some VC-eRF1s from these lineages may alter their codon specificities via substitutions that have occurred in different positions to the VC cistines tested here (e.g. positions 51 and 126). Thus, sequencing of additional SC-eRF1s and VC-eRF1s and structural analyses are required to further elucidate the mechanism of the translation termination system at the molecular level.

ACKNOWLEDGEMENTS

We thank J. M. Archibald and J. O. Andersson (Dalhousie University, Halifax, Canada) for critical reading of this manuscript. We also thank T. M. Embley (Natural History Museum, University, Halifax, Canada) for critical reading of this manuscript.

REFERENCES