Discovery of new genes and deletion editing in *Physarum* mitochondria enabled by a novel algorithm for finding edited mRNAs

Jonatha M. Gott*, Neeta Parimi and Ralf Bundschuh

Center for RNA Molecular Biology, Case Western Reserve University, Cleveland, OH 44106, USA and 1Department of Physics, The Ohio State University, Columbus, OH 43210, USA

Received June 13, 2005; Revised and Accepted August 19, 2005

**ABSTRACT**

Gene finding is complicated in organisms that exhibit insertional RNA editing. Here, we demonstrate how our new algorithm Predictor of Insertional Editing (PIE) can be used to locate genes whose mRNAs are subjected to multiple frameshifting events, and extend the algorithm to include probabilistic predictions for sites of nucleotide insertion; this feature is particularly useful when designing primers for sequencing edited RNAs. Applying this algorithm, we successfully identified the *nad*2, *nad*4L, *nad*6 and *atp*8 genes within the mitochondrial genome of *Physarum polycephalum*, which had gone undetected by existing programs. Characterization of their mRNA products led to the unanticipated discovery of nucleotide deletion editing in *Physarum*. The deletion event, which results in the removal of three adjacent A residues, was confirmed by primer extension sequencing of total RNA. This finding is remarkable in that it comprises the first known instance of nucleotide deletion in this organelle, to be contrasted with nearly 500 sites of single and dinucleotide addition in characterized mitochondrial RNAs. Statistical analysis of this larger pool of editing sites indicates that there are significant biases in the 2 nt immediately upstream of editing sites, including a reduced incidence of nucleotide repeats, in addition to the previously identified purine-U bias.

**INTRODUCTION**

Predicting the sequence of messenger RNAs (mRNAs) from genomic sequence data is complicated by processes such as splicing, alternative splicing and RNA editing. RNA editing can take many forms, including base changes, deletion of encoded nucleotides, insertion of non-encoded nucleotides and replacement of nucleotides. These site-specific alterations occur in a wide variety of organisms, extending from viruses and single cell protists to man and, depending on the mRNA and organism, can involve changes ranging from a single nucleotide to more than 50% of the residues in a mature transcript (1–4). Such editing events have a significant impact on gene expression, frequently changing the coding capacity of mRNAs and thus contributing to the diversity of the proteome.

Although there are a number of programs available for prediction of standard genes within sequenced genomes, the localization of genes whose open reading frames (ORFs) are created by the insertion of non-encoded nucleotides is problematic. A case in point is provided by the mitochondrial genome of *Physarum polycephalum*, whose sequence was reported by Takano et al. in 2001 (5). Nearly all characterized *Physarum* mitochondrial RNAs contain multiple nucleotides that are not encoded in the genome, which are present as either single or dinucleotide insertions (5–9); the vast majority of these are single C insertions. RNAs that are either known or predicted to be edited include the mRNAs encoding components of NADH dehydrogenase (*nad*1, *nad*3, *nad*4, *nad*5 and *nad*7), apocytochrome b (*cyrb*), cytochrome oxidase (*cox*1, *cox*2 and *cox*3), and ATP synthase (*atp*1 and *atp*9), both the large and small ribosomal RNAs (rRNAs: *lsu* and *ssu*), and 4 of the 5 mitochondrially-encoded tRNAs (*met*1, *lys*, *pro* and *glu*). Notably, four genes that are commonly found in mitochondrial genomes, *nad*2, *nad*4L, *nad*6 and *atp*8, are absent in the published physical map of the *Physarum* mitochondrial genome (5). In addition, although the known genes are generally closely spaced, there are a number of large gaps in the current mitochondrial map, some of which are transcribed (J.M.Gott and Y.W.Cheng, unpublished data). Thus it seemed likely that the *Physarum* mitochondrial
The genome contains additional genes that are invisible to the conventional bioinformatics programs.

Here we describe the first application of a new algorithm Predictor of Insertional Editing (PIE) designed to find genes whose ORFs are created by insertional editing, using previously characterized mRNAs from Physarum mitochondria as the initial training set (10). The utility of this algorithm is enhanced by inclusion of probabilistic predictions of editing sites, as illustrated here for the previously uncharacterized cox2 mRNA. Using this algorithm, we successfully identified the nad2, nad4L, nad6 and atp8 genes, which had gone undetected by existing programs. Characterization of their mRNA products led to the first known instance of overlapping genes in this organelle, as well as the unexpected discovery of nucleotide deletion editing in Physarum mitochondria, which was confirmed by sequencing bulk mitochondrial RNA.

MATERIALS AND METHODS

PIE algorithm

The PIE algorithm is described in detail in (10). Its general concept is reviewed below. The parameters used to build the position specific scoring matrices (PSSMs) which summarizes the multiple alignment of all the known protein sequences of the protein in question were the BLAST default parameters, namely the BLOSUM62 (11) scoring matrix using an affine gap cost 11 + k for every gap of k amino acids. Also within PIE we used a gap cost 11 + k for amino acid insertions and deletions. In addition to the amino acid gap cost, the application of PIE to the P. polycephalum mitochondrial genome depends on two parameters, namely the cost for an editing site after a purine–pyrimidine dinucleotide and the cost for an editing site that does not follow a purine–pyrimidine. Differentiating between these two cases takes into account the known context bias of C insertion sites in Physarum mitochondria, which preferentially occur after a purine–pyrimidine dinucleotide. As described in (10), these two parameters were optimized by evaluating the predictive power of PIE on the mRNAs derived from six genes (nad7, cox1, cox3, cyb, atp1 and atp9) for which the location of the editing sites was accessible in GenBank before this study. The resulting values are a cost of six for an editing site after a purine–pyrimidine dinucleotide and a cost of 12 for an editing site that does not follow a purine–pyrimidine. The source code of PIE ready to be compiled on Linux systems and instructions for use are freely available for non-commercial use on request from the authors (contact Ralf Bundschuh at Bundschuh@mps.ohio-state.edu).

Assignment of probabilities to predicted editing sites

In order to predict sites of C insertion, PIE assigns a score S(C) to every way C of inserting Cs into the genomic sequence and finds the way Cmax of inserting Cs that maximizes the score. In order to calculate a probability for a C insertion to be predicted at position i within the gene, we in addition define the weight of a way C to insert Cs into the genomic sequence as w(C) = exp[λS(C)] where λ is a scale factor that was used in creating the scoring system in the first place. This weight is maximal for the optimal way Cmax of inserting Cs and gets smaller as the score decreases from the optimum. Then, we define the probability p_

\[ p_i = \frac{\sum [C \text{ with an inserted } C \text{ at position } i]}{\sum [C \text{ with all } C]} \]

Both of these sums can be efficiently calculated in O(N^2) computer time where N is the length of the sequence using a dynamic programming scheme. If all high scoring ways of inserting Cs have a C insertion at position i the sum in the numerator is nearly as big as the sum in the denominator and the probability p_i is close to one. If none of the ways of inserting Cs with scores close to the optimal score have a C insertion at position i the numerator will be much smaller than the denominator and the probability p_i is close to zero. In the most interesting case that some of the ways to insert Cs with relatively high scores contain a C insertion at position i and some do not, the probability p_i calculated according to the formula above measures what fraction of the ways to insert Cs with high scores contains a C insertion at position i and how high their scores are relative to the optimal score.

Identification of significant sequence motifs

In order to identify sequence positions that contain significant amounts of information about the editing sites, we counted the number of times n_i(X) the base X is observed at position i, where the position i is counted relative to the position of the inserted C, i.e. i = −1 describes the position immediately upstream of the inserted C. Given the total number N_i = n_i(A) + n_i(U) + n_i(G) + n_i(C) of observed editing sites of a given type we obtain the observed frequencies f_i(X) = n_i(X) / N_i. These have to be compared with the corresponding background frequencies p_i(X). In order to obtain these background frequencies we collected all codons of the 11 known protein coding genes for the Physarum mitochondrion and eliminated all codons that contain an editing site. The background frequency p_i(X) of base X at position i is then given by the ratio of the number of times the base X is observed in these codons at the codon position corresponding to the position i relative to the editing site and the total number of codons collected. We quantify the difference between the observed f_i(X) and the expected p_i(X) in terms of the relative entropy H_i = \sum f_i(X) \ln(f_i(X) / p_i(X)). In order to assign a P-value, we generated many independent sets of N_i bases randomly chosen according to the background frequencies p_i(X) and calculate the relative entropy for these randomly generated sets of bases. Then, the P-value is the fraction of times the relative entropy of the randomly generated sets is larger than the relative entropy of the originally observed set. For positions immediately before and after the editing site the same procedure is used but restricted to the three bases A, U and G since a C cannot flank an unambiguous editing site. In order to detect significant correlations between two positions, the same approach is used but instead of X denoting the identity of a single base at a position, it denotes the identity of a pair of bases at a pair of positions. The background frequencies to be compared to are in the latter case the products of the observed frequencies of the bases in the two positions under study.

Oligonucleotide primers

2atp8: CAGCTTCTAATAAAAAAGCTAAC; 1nd4L: GGTTATTGATCTAAAGGACCATAAAA; 2atp8: TTGTCTAAAGGATTATATAATTC; 2nd6: TCTAGTAAATTATTAGAAACCA;
cDNA cloning
Mitochondrial DNA and RNA were isolated and PCR and RT–PCR were carried out as described previously (7) using the primers described above. Automated DNA sequencing was carried out by the Case Western Reserve Core Facility on a fee for service basis. The sequences are available from GenBank under the Accession nos DQ092488 (atp8), DQ092489 (cox2), DQ092490 (nad2), DQ092491 (nad4L) and DQ092492 (nad6).

Primer extension sequencing
Total Physarum RNA or plasmid DNA was heated to 65°C (RNA) or 95°C (DNA) for 3 min in 50 mM Tris–HCl (pH 8.3 at RT)/60 mM NaCl/10 mM DTT with end-labeled primer (11nd2 in Figure 5) and quickly cooled. Mg acetate was added to 6 mM and the primer was extended in the same buffer using AMV reverse transcriptase (Life Sciences) in the presence of 200 μM dNTPs and the appropriate ddNTP at 42°C for 30 min. Samples were mixed with an equal volume of 7 M urea gel dye and run on a denaturing 8% polyacrylamide gel before phosphorimager exposure.

RESULTS
PIE, a novel algorithm for prediction of genes whose transcripts are processed by RNA editing
The underlying basis of PIE is to compare a series of conceptual translations of the genome with aligned protein sequences from other organisms in order to localize potential genes and identify likely editing sites. The technical details of PIE and how its parameters were optimized using the set of edited Physarum mitochondrial mRNAs known before this study are described elsewhere (10). Here, we will give an informal overview of the mechanics of PIE, summarized in Figure 1, using the uncharacterized Physarum cox2 gene as an example.

A portion of the cox2 gene was previously localized to a region spanning 29105–29598 nt on the complementary strand of the Physarum mitochondrial genome (5), but the exact boundaries of the gene and the sequence of the mRNA have not been reported. To predict the editing sites in the Physarum cox2 mRNA, Cox2 protein sequences from other organisms were first aligned (Figure 1). This can be conveniently achieved by choosing one Cox2 sequence from GenBank and searching the non-redundant database for homologs of this sequence with the program PSI-BLAST (12). PSI-BLAST not only finds the homologs but also constructs a multiple alignment. For Cox2 we used amino acids #230–506 (the part corresponding to Cox2) of the combined Cox1/Cox2 protein sequence of Dictyostelium discoideum (Accession no. 2655920) as the query sequence. This choice was motivated by the relatively close relationship between P. polycephalum and D. discoideum; however, the precise choice is not critical to the outcome (R. Bundschuh, unpublished data). Only a small portion of the resulting multiple alignment is shown in Figure 1. The actual alignment consisted of 7598 sequences averaging 200 amino acids in length after three iterations of PSI-BLAST. The information in this multiple alignment was then summarized by a PSSM. Amino acids which are frequently observed at a given position are assigned a high score, while a low score is assigned to residues that are rarely observed at a given position.

The second phase of PIE involves iterative manipulations of the genomic sequence under study. Because ~90% of the editing events in Physarum mitochondrial RNAs are insertions of single C residues, only C insertions are considered in this algorithm. A subsequence of the genome that includes the putative coding region for the Physarum cox2 gene was selected and a collection of related sequences were considered which are identical to the original subsequence except for the addition of individual C’s at multiple positions. In order to obtain the score of a given pattern of C insertions into the genomic subsequence, each of the resulting DNA sequences is translated into a protein sequence and compared to the PSSM, generating an alignment score minus a cost for every editing site.
Among all the $2^N$ possible ways of inserting Cs into a fragment of the genomic sequence of length $N$ (a C could potentially be inserted after each base), the one with the highest score comprises the predicted positions of the editing sites in the cox2 mRNA. The technical effort described in (10) is necessary in order to reorganize the computation in such a way that the highest scoring of the $2^N$ ways of inserting Cs can be chosen with an effort of computer time that behaves only like $N^2$. As a result of these efforts of purely computational nature, the approach outlined above can be executed within a few minutes on a regular workstation.

In order to predict the positions of start and stop codons, we allowed the alignment of a DNA sequence (modified by inserted Cs) and the PSSM to start and end at any start and stop codon, respectively, and chose the start and stop codon positions that generated the alignment with the highest score. However, the lengths of the N- and C-terminal regions typically vary considerably between different members of a protein family, and the sequence conservation itself is much weaker towards the terminal regions than in the protein core. Both of these effects make the determination of start and stop codons difficult. This problem is shared with all comparative gene finding programs as far as the position of the start codon is concerned; the position of the stop codon is, of course, trivial to detect in an organism without RNA editing, but is as challenging as finding the position of the start codon in the presence of editing. This problem is shared with all comparative gene finding programs as far as the position of the start codon is concerned; the position of the stop codon is, of course, trivial to detect in an organism without RNA editing, but is as challenging as finding the position of the start codon in the presence of editing. It is not surprising, therefore, that previous attempts were also unable to precisely locate the relevant start and stop codons (5).

With the parameter settings described in the Materials and Methods section, PIE predicts the exact location of 70% of the editing sites in the six protein coding genes (nad7, cox1, cox3, cyb, atp1 and atp9) for which the location of the editing sites was accessible in GenBank prior to this study correctly, with another 17% of predicted editing sites falling within 3 nt of the actual editing site (10). Predictions for the insertion sites within the cox2 mRNA were not as accurate as with the reference set, largely due to the variability of known Cox2 proteins. Characterization of the Physarum cox2 mRNA via RT–PCR and cDNA cloning indicated that the coding region of this mRNA (positions 29666–28983 on the complementary strand of the genome) contains 33 C residues not present in the mitochondrial genome (shown above the sequence in Figure 2).

Figure 2. Characterization of the cox2 mRNA. The region of the Physarum polycephalum mitochondrial genome that contains the cox2 gene is shown, with the predicted sites of C insertion shown below. Note that only one C is expected to be added at any given cluster; relative probabilities of insertion at any given site are indicated (see scale at the bottom). The experimentally determined editing sites are shown above the nucleotide in the genomic sequence that lies immediately 5' to the inserted C. Note that when an inserted C lies above an encoded C, the exact site of C insertion is ambiguous. Numbers refer to genomic coordinates from ref. (5).

Among all the $2^N$ possible ways of inserting Cs into a fragment of the genomic sequence of length $N$ (a C could potentially be inserted after each base), the one with the highest score comprises the predicted positions of the editing sites in the cox2 mRNA. The technical effort described in (10) is necessary in order to reorganize the computation in such a way that the highest scoring of the $2^N$ ways of inserting Cs can be chosen with an effort of computer time that behaves only like $N^2$. As a result of these efforts of purely computational nature, the approach outlined above can be executed within a few minutes on a regular workstation.

In order to predict the positions of start and stop codons, we allowed the alignment of a DNA sequence (modified by inserted Cs) and the PSSM to start and end at any start and stop codon, respectively, and chose the start and stop codon positions that generated the alignment with the highest score. However, the lengths of the N- and C-terminal regions typically vary considerably between different members of a protein family, and the sequence conservation itself is much weaker towards the terminal regions than in the protein core. Both of these effects make the determination of start and stop codons difficult. This problem is shared with all comparative gene finding programs as far as the position of the start codon is concerned; the position of the stop codon is, of course, trivial to detect in an organism without RNA editing, but is as challenging as finding the position of the start codon in the presence of editing. It is not surprising, therefore, that previous attempts were also unable to precisely locate the relevant start and stop codons (5).

With the parameter settings described in the Materials and Methods section, PIE predicts the exact location of 70% of the editing sites in the six protein coding genes (nad7, cox1, cox3, cyb, atp1 and atp9) for which the location of the editing sites was accessible in GenBank prior to this study correctly, with another 17% of predicted editing sites falling within 3 nt of the actual editing site (10). Predictions for the insertion sites within the cox2 mRNA were not as accurate as with the reference set, largely due to the variability of known Cox2 proteins. Characterization of the Physarum cox2 mRNA via RT–PCR and cDNA cloning indicated that the coding region of this mRNA (positions 29666–28983 on the complementary strand of the genome) contains 33 C residues not present in the mitochondrial genome (shown above the sequence in Figure 2). Of the 33 C insertion sites present in the cox2 mRNA, 18 (54%) were predicted exactly, with another 8 (25%) predicted editing sites falling within 3 nt of the experimentally determined C insertion sites. Although this level of accuracy is more than an adequate for mRNA characterization, practical considerations led us to incorporate a probability function into the PIE algorithm, as described below.

Providing probabilistic predictions for sites of nucleotide insertion

Although PIE has strong predictive power, the accuracy of the predictions varies between editing sites depending on the degree of conservation observed within the protein sequences at the position of the editing site. This can be a significant problem when characterizing edited RNAs in that primers based on the sequence of the (unedited) genome may not anneal well to the edited RNA if they overlap an editing site that was incorrectly predicted by PIE. Given that editing sites in some regions of the gene can be predicted with less certainty than in others, it is useful to know which predicted editing sites are more likely to be reliable. Thus, we devised a method to assign a probability of correct prediction to each editing site.

This method, described in detail in the Materials and Methods section, is based on the extent of conservation between the aligned proteins used to generate the original scoring matrix.
and takes into account both the optimal and suboptimal ways of inserting Cs. If the editing site is in a very strongly conserved region of the multiple alignment, often only one position at all is plausible. This is reflected in the fact that this position is the same in the optimal as well as in all relevant suboptimal ways of inserting Cs. Such an editing site will be assigned a high probability. If, on the other hand, a specific editing site is in a region of the protein with relatively little conservation within the multiple sequence alignment, several amino acid sequences are similarly plausible and the position of the editing site between the optimal and other high scoring patterns of C insertions may differ. Such an editing site is assigned a low probability.

After calculating all probabilities, we graphically depict the degree of certainty associated with each predicted editing event, as shown in Figure 2 for the cox2 example. The predictions of PIE are shown as Cs immediately below the genomic sequence, shaded according to their assigned certainty. For some editing sites, e.g. the one at position 29647, the prediction is very certain while in other regions of the gene, e.g. in the stretch from position 29280 to position 29350, many different editing site positions are possible. The latter region would clearly not be a good place to choose as a primer binding site, whereas the region from position 29260 to position 29280, which is highly unlikely to contain an editing site based on protein alignments, would be an ideal location for primer annealing.

Examining the experimentally determined editing sites shown immediately above the genomic cox2 sequence in Figure 2, we find that the predictions of PIE are correct for every editing site that is assigned a certainty of at least 0.5. Just as significantly, only a single editing site (the one at position 29213) occurs at a position where PIE judged the probability for an insertion site to be <0.05 indicated by the absence of a C below the genomic sequence (note, that the apparent misses at editing sites at 29388 and 29069 are due to the fact that these Cs are added next to an encoded C). This implies that choosing primers in regions where PIE predicts the absence of editing sites will be a reliable strategy for primer design.

**Application of PIE to identify the location of new genes**

For the six characterized genes used as the training set (10) and the cox2 gene described above the approximate genomic location had been previously identified by analyzing the mitochondrial genome with the standard tools BLASTX (13) and BEAUTY (14). However, these bioinformatics programs failed to identify genes encoding a number of expected mitochondrial proteins, including atp8, nad2, nad4L and nad6 (5). To determine whether these latter four genes are even present in the mitochondrial genome of *P. polycephalum*, we first generated protein alignments and PSSMs for each. We then divided the full mitochondrial genome into overlapping fragments 1120 bases in length, i.e. the first piece contains bases 1–1120, the second contains bases 561–1680, etc. We then applied PIE to each of these fragments (and their reverse complements) as described above for the cox2 mRNA, generating a series of scores for the atp8, nad2, nad4L, and nad6 PSSMs. In addition to predicting the optimal way of inserting extra Cs, PIE also measures how similar the protein product of any putative mRNA is to the protein sequence of the query gene. Once scores for all fragments of the genome and their reverse complements were generated, we found that for a given gene, the score for a single region was significantly larger than all other scores. Thus, the data generated by PIE provided strong evidence that these four genes were indeed encoded in the mitochondrial genome of *P. polycephalum*, identified their probable locations, and predicted the sites of C insertions present in each mRNA. These predictions were then tested by characterizing cDNA clones generated by RT-PCR, using oligonucleotide primers based on statistical predictions of editing sites.

**Characterization of mRNAs from candidate genes identified by PIE**

**atp8 and nad4L.** Our algorithm predicted that the atp8 and nad4L genes were adjacent to one another, localizing to the region of the mitochondrial DNA between 35567 and 36062. Genomic and cDNA clones from this region were generated using primers 2atp8 and 2nd4L, which anneal to the regions indicated in Figure 3. An alignment of the unedited (mtDNA) and edited (cDNA) sequences is presented in Figure 3, with the conceptual translation for both the atp8 and nad4L mRNAs shown below. The atp8 start codon and nad4L start and stop codons were predicted correctly, while the actual stop codon occurring 11 nt downstream of the predicted termination codon (indicated by a dotted underline). The creation of both the atp8 and nad4L ORFs involves the insertion of single C residues relative to the mitochondrial genome. The atp8 mRNA contains 9 C insertions, while the nad4L mRNA has 13 added C residues (Figure 3).

The coding sequences for atp8 and nad4L are present on the same transcript. RT-PCR using total mitochondrial RNA as template gave a product of the expected length using primers that anneal 5′ of atp8 and 3′ of nad4L (primers 2atp8 and 2nd4L in Figure 3). In addition, we found no evidence for a separate nad4L transcript when carrying out primer extension sequencing of total mitochondrial RNA with a primer annealing near the 5′ end of the nad4L ORF (primer 1nd4L in Figure 3, data not shown). Thus, the vast majority of the mRNAs encoding atp8 and nad4L are polycistrionic. Because the atp8 stop codon overlaps the start codon of the nad4L ORF, we speculate that translation of the two ORFs is likely to be coupled.

**nad6.** The nad6 gene is found on the opposite DNA strand within the region encompassing positions 56758–56280. A comparison of the sequence of the nad6 gene and mRNA is shown in Figure 4. The nad6 mRNA is processed by the insertion of a single U and 18 single C residues, resulting in the creation of an ORF encoding 165 amino acids. We predicted the stop codon correctly, but the actual start codon is 38 nt downstream of the predicted start codon. Similar to what was observed with the atp8 and nad4L transcript described above, the nad6 gene is cotranscribed with the downstream cox3 gene, which is only separated from the nad6 ORF by a single nucleotide (Figure 4); oligonucleotide primers annealing within the cox3 coding region were used to generate many of the nad6 cDNA clones. Primer extension sequencing using a primer annealing near the 5′ end of the nad6 coding region indicated that the 5′ end of the nad6
transcript falls \(\sim 110\) nt upstream of the \textit{nad6} start codon (data not shown).

\textit{nad2}. Initially, it appeared that the \textit{nad2} gene could potentially be encoded by three different loci in the mitochondrial genome. However, upon further examination, two of these positions mapped to the known \textit{nad4} and \textit{nad5} genes, which show sufficient sequence similarity to \textit{nad2} to give relatively high scores with the \textit{nad2} protein alignments, with the most likely \textit{nad2} candidate localizing to positions \(30699\)–\(32105\) on the \textit{Physarum} mitochondrial genome.

Characterization of overlapping cDNA clones generated by RT–PCR indicated that the \textit{nad2} mRNA contains 55 additional C residues, four extra U residues and, surprisingly, a deletion of three adjacent A residues relative to the genomic DNA. The triple A deletion was present in all cDNA clones generated with two different RT–PCR primer sets, indicating that the vast majority of the \textit{nad2} mRNA was likely to contain this deletion. However, because deletion editing has not been reported previously in \textit{Physarum} mitochondria, we confirmed this observation by primer extension sequencing of total RNA from \textit{Physarum} mitochondria (Figure 5). An end-labelled oligonucleotide primer was annealed to both \textit{nad2} DNA (left) and bulk mitochondrial RNA (right) and extended by reverse transcriptase in the presence of dideoxynucleotides. The data clearly show that the three A residues are present in the \textit{nad2} DNA (indicated by arrowheads in Figure 5), but are absent in the mRNA (thick line to the right of the RNA sequence). As expected, C residues added by editing are present in the bulk RNA sample (indicated by asterisks) but not in the DNA, confirming that the template for the primer extension reaction was indeed \textit{nad2} mRNA. Thus, virtually all of the \textit{nad2} mRNA present in \textit{Physarum} mitochondria lack these three encoded A residues, confirming the existence of deletion editing in this organism.

### Statistical analysis of editing site positions

Prior to our study, the editing sites of six protein coding genes (\textit{nad7}, \textit{cox1}, \textit{cox3}, \textit{cyt b}, \textit{atp1} and \textit{atp9}) were known. As shown in the first column of Table 1, these contained 250 editing sites, of which 222 are C insertions. Of these C insertions, the exact insertion site is known for 140 C residues; the precise site of insertion of the remaining 82 cases of C insertion is ambiguous, i.e. they are flanked by an encoded C, which makes it impossible to determine which C in the mRNA sequence is added from sequence data alone. The number of editing sites within stable RNAs (the large and small rRNAs and four
The current size of the database after inclusion of the editing sites in the five genes (\textit{cox}2, \textit{nad}2, \textit{nad}4L, \textit{nad}6 and \textit{atp}8) characterized in this study are also shown in Table 1. Because our additions amount to a significant expansion of the database, we reexamine here the statistics of these editing site positions and the sequences surrounding the editing sites.

The only known general feature of the sequence environment of \textit{Physarum} editing sites is a significant bias towards having a purine-U sequence directly upstream of unambiguous C insertion sites (8). 140 of the previously known 206 unambiguous C insertion sites (or 68%) show this sequence feature. This ratio remains unchanged in our new data set, where 201 of the 293 unambiguous C insertion sites (69%) follow a purine-U.

The position of C insertion in coding sequences is also non-random in that roughly two thirds of the unambiguous C insertions appear at the third codon position, while the second codon position is significantly underrepresented (8). This general trend remains unchanged as we take into account the editing sites from our five new genes. Of the

Figure 4. Editing sites within the \textit{nad}6 mRNA. Notations are as in the legend to Figure 3 except that the \textit{cox}3 start codon is indicated by a double underline.

![DNA and RNA sequence alignment](image)

Figure 5. Primer extension sequencing of \textit{nad}2 DNA and mRNA showing the region encompassing the triple A deletion. Arrowheads indicate the A residues present in the DNA, but missing from the bulk RNA. The missing As are indicated by a thick line at the right; inserted Cs are marked with asterisks.

| Table 1. Total observed editing events |
|--------------------------------------|
| | Previous coding | Total coding | Stable RNA | Previous total | Total |
|-----------------|-----------------|--------------|-------------|----------------|-------|
| Editing sites   | 250             | 390          | 107         | 357            | 497   |
| C insertion     | 222             | 353          | 107         | 319            | 450   |
| Unambiguous     | 140             | 227          | 66          | 206            | 293   |

This table gives an overview of the total number of editing events in the mRNAs characterized before our study (\textit{nad}7, \textit{cox}1, \textit{cox}3, \textit{cytb}, \textit{apt}1 and \textit{apt}9), all mRNAs including the ones studied here (\textit{cox}2, \textit{nad}2, \textit{nad}4L, \textit{nad}6 and \textit{atp}8), and the stable RNAs, as well as the total characterized before our study and including the results of our study.

the tRNAs) and the current size of the database after inclusion of the editing sites in the five genes (\textit{cox}2, \textit{nad}2, \textit{nad}4L, \textit{nad}6 and \textit{atp}8) characterized in this study are also shown in Table 1. Because our additions amount to a significant expansion of the database, we reexamine here the statistics of these editing site positions and the sequences surrounding the editing sites.

The only known general feature of the sequence environment of \textit{Physarum} editing sites is a significant bias towards having a purine-U sequence directly upstream of unambiguous C insertion sites (8). 140 of the previously known 206 unambiguous C insertion sites (or 68%) show this sequence feature. This ratio remains unchanged in our new data set, where 201 of the 293 unambiguous C insertion sites (69%) follow a purine-U.

The position of C insertion in coding sequences is also non-random in that roughly two thirds of the unambiguous C insertions appear at the third codon position, while the second codon position is significantly underrepresented (8). This general trend remains unchanged as we take into account the editing sites from our five new genes. Of the
total unambiguous C insertion sites within mRNAs, 58 (26%) occur at the first position, 24 (11%) at the second position and 145 (64%) at the third position within the codon. However, with our larger data set it becomes clear that the suppression of C insertions at the second codon position is not as severe as concluded from the previous data set, in which only 6% of all insertions occurred at the second position.

The distribution of codons created by C insertions is also highly skewed, as detailed in Table 2. Comparing the data from previously characterized mRNAs (in parenthesis) with the data from all 11 protein coding genes, we find that the four dominant codons AUG, ACC, GUC and GCC remain the same, accounting for roughly half of the codons created by C insertion. This is predictable, at least for the codons in which the inserted C is unambiguous, based on the purine-U and third position biases noted above. The most notable outcome of our enlarged data set is that the number of codons created by C insertion has been expanded to include the codons CGG and CGC. This suggests that there is no fundamental reason why certain codons cannot be created by editing, but that the codons which have not been observed simply are so rare that they do not occur in the limited data set. For instance, the fact that the codons ACA, ACU, ACG and UCG are not present in our current data set (Table 2) is likely due to the rarity of C insertions at the second position, given that Cs are added in these contexts at other codon positions and in stable RNAs. This is further supported by the fact that ACU and UCG have been observed as edited codons in the nad1 mRNA (8) which we could not include in our study since its edited sequence is not published.

Table 2. Codons created by C insertions

| Codon | A | U | G | C |
|-------|---|---|---|---|
| AXC   | 5 (4) | 76 (49) | 10 (5) | 40 (29) |
| AXA   | 0 (0) | 6 (0) | 0 (0) | 6 (3%) |
| UXG   | 4 (2) | 11 (7) | 1 (1) | 12 (8) |
| UXG   | 6 (1) | 11 (3) | 0 (0) | 12 (8) |
| GXC   | 3 (2) | 33 (23) | 0 (0) | 29 (17) |
| GAX   | 1 (1) | 5 (4) | 1 (0) | 12 (8) |
| CAX   | 8 (5) | 3 (3) | 1 (1) | 2 (1) |
| CUX   | 17 (12) | 12 (5) | 4 (2) | 2 (1) |
| CXG   | 4 (3) | 6 (6) | 0 (0) | 1 (0) |
| CCX   | 14 (9) | 12 (7) | 2 (2) | 2 (1) |

The numbers shown comprise all 11 characterized mRNAs; numbers in parenthesis include data from the six previously known mRNAs.

Table 3. Correlation between the two positions immediately preceding the editing site

| A        | U        | G        | Total |
|----------|----------|----------|-------|
| −2−1 A   | 6 (14)   | 122 (115)| 13 (11)| 141 (62%) |
| 1 U      | 9 (3)    | 16 (6)   | 6 (2) | 31 (14%)  |
| 1 G      | 6 (5)    | 43 (39)  | 0 (4) | 49 (21%)  |
| C        | 2 (1)    | 4 (6)    | 0 (1) | 6 (3%)    |
| Total    | 23 (10%) | 185 (82%)| 19 (8%)| 227 (100%)|

The main entries are the actual numbers of observations of each of the pairs of bases while the numbers in parenthesis are the number of observations expected from the percentages of the totals in the individual positions.

in order to eliminate effects of the strong codon bias seen in Table 2. The deviation between the observed frequencies and the background frequencies was quantified in terms of a p-value as described in the Materials and Methods section. After taking into account the codon bias, only the positions −1 and −2 immediately upstream of the editing site showed a truly significant deviation (p-value of 0.002 or smaller) from background. Thus, we conclude that at the current level of 227 unambiguous editing sites in coding regions, no significant sequence pattern beyond the known purine-U bias directly upstream of the editing site could be found.

We also investigated whether there are noteworthy correlations between pairs of positions within the window of 15 positions up- and downstream of editing sites (see Materials and Methods). A pair correlation of two positions within the window means that a given pair of bases appears significantly more or less often at the two positions than what is expected from looking at the individual frequencies of occurrence of the bases at the corresponding positions. If, for example RNA secondary structures in the vicinity of insertion sites play an important role in editing, one would expect pairs that form Watson–Crick base pairs to occur more often than non-Watson–Crick combinations within the relevant regions. It is also possible that more than one sequence pattern directs editing that on average does not look different from the background. In this case, pairs of bases that belong to the same motif should be enhanced over other pairs. Thus, correlations between pairs of positions can be used to uncover subtle patterns that might not be otherwise observed.

When editing sites within coding regions were used for this analysis, we observed a correlation between positions −2 and −1 as well as between positions 10 and 11, both with an individual P-value of 0.0006. These P-values are not much smaller than the P-values from an equivalent 435 pairs of completely random sequences (P-value = 1/435 or 0.0023), and thus both correlations are only marginally significant. Indeed, if we expand our data set to include the unambiguous editing sites within the stable RNAs in this analysis, the correlation between positions 10 and 11 loses its statistical significance. Thus, we will not discuss the 10/11 pair correlation beyond noting that it reflects a preference for identical bases in these two positions. In contrast, the persistence of the two positions −2 and −1 (immediately upstream of the editing site) in the larger data set warrants a closer look at this specific correlation. Table 3 shows the number of each pair of bases at these two positions relative to unambiguous C insertion sites. The purine-U bias discussed above is clearly reflected in the
DISCUSSION

A major goal of any genome sequencing project is the accurate identification of all encoded gene products. This is particularly challenging in organisms whose transcripts are subject to the addition of nucleotides that are not encoded in the gene itself. Indeed, such ‘cryptogenes’ are often invisible to standard gene-finding algorithms, as evidenced by the initial characterization of the mitochondrial genomes of Trypanosoma brucei and Leishmania tarentolae (16), as well as that of P. polycephalum (5).

In Physarum mitochondria, ∼1 out of every 25 nt in edited mRNAs and 1 of every 40 nt in stable RNAs are added during transcription, most as single nucleotide insertions. ORFs are created by repeated frameshifts, which can occur as often as every fourth codon. It is therefore not surprising that standard bioinformatics tools failed to identify all genes present in the Physarum mitochondrial genome. Although the genes for aps8, nad2, nad6 and nad4L were expected to be encoded in Physarum mitochondria, these four genes were found only upon application of PIE, an algorithm expressly developed to look for genes which require insertional editing for their expression. Additional potential genes have recently been mapped to other regions of the Physarum mitochondrial genome using PIE (C. Ainsley, H. Lee, and R. Bundschuh, unpublished data), demonstrating further the efficacy of this algorithm.

Use of the PIE algorithm is not restricted to identification of sites of C insertion in Physarum mRNAs. Mitochondrial mRNAs in a number of other myxomycetes are edited by the insertion of single U residues. We have tested the general utility of the PIE algorithm by applying it to the cox1 genes of Clastoderma debaryanum, Arcyria cinerea, Stemotitis flavogenita, and Didymium nigripes (10), whose mRNAs were previously characterized by Horton and Landweber (17). In principle, the same general strategy should be applicable to all types of insertional editing, with only minor changes to reflect the characteristics of editing in that organism. This includes the possible search for novel cryptogenes in kinetoplastids, although in this case the algorithm would have to be modified to accommodate deletions and insertions of longer stretches of consecutive uridines. However, given that kinetoplastid editing sites are known to be specified by guide RNAs (18), an approach involving a search for cryptogenes and their cognate guide RNAs (19) would be more direct.

The genes for aps8, nad2, nad6 and nad4L are not as well conserved as the previously localized Physarum mitochondrial genes, and this may account for the failure of BLASTX and BEAUTY to find these genes. Not surprisingly, the accuracy of editing site predictions for those four genes were not as high as those of the cox2 mRNA, which was anticipated based on the insertion site probabilities generated for each gene. However, although the exact boundaries were not correctly identified for every gene, the predictions of PIE allowed facile characterization of their respective cDNAs. In doing so, two new features of Physarum mitochondrial gene expression were identified: overlapping genes and deletion editing.

Nucleotide deletions have been observed previously in kinetoplastid mRNAs (18), but the discovery of deletion editing in Physarum mitochondria was surprising, given that no deletions have been reported in the previously characterized mitochondrial RNAs, which include seven mRNAs (nad7, cox1, cox3, cyt6, atp1, atp9, and the unpublished nad1), the large and small rRNAs, and four tRNAs (5–7). Although nucleotide deletions are much less frequent than insertions in T. brucei (322 deletions versus 3030 insertions), L. tarentolae (161 deletions versus 1436 insertions), and other kinetoplastids (18), they still make up a substantial proportion of the total number of editing events in these organelles. In contrast, the three deleted A residues described in this work constitute <1% of the known editing sites in Physarum mitochondria.

The existence of nucleotide deletions extends the list of editing types that occur in Physarum mitochondria, which already includes single C insertions, U insertions, dinucleotide insertions (GU, CU, UA, GC, AA and UU), and C to U changes (5–7). It remains to be determined whether these nucleotide deletions occur co-transcriptionally, as observed for the nucleotide insertions (20), or post-transcriptionally, as is the case for C to U changes (21). Although all forms of editing in Physarum mitochondria are virtually 100% efficient in vivo (21), RNAs made in vitro contain a mixture of unedited, edited and mis-edited sites (22). Interestingly, one form of misediting that is observed during run-on transcription in partially purified mitochondrial transcription elongation complexes is the deletion of three encoded nucleotides immediately downstream of an insertion site. This finding intimates that nucleotide deletions may also occur co-transcriptionally, although this remains to be tested.

ACKNOWLEDGEMENTS

We would like to acknowledge useful contributions to this work by Linda Visomirski-Robic, Joey Hunter, Christopher Webb, Angela Stout, and Marianne Lee and thank Amy Rhee for critical reading of the manuscript. This work was supported by NIH grant GMS4663 to J.M.G. and National Science Foundation grant DMR0404615 to R.B. This work was initiated at the Rustbelt RNA meeting supported by the National Science Foundation under Grant MCB0121758.
Funding to pay the Open Access publication charges for this article was provided by NIH grant GM54663 to J.M.G.

Conflict of interest statement. None declared.

REFERENCES

1. Gott, J.M. (2003) Expanding genome capacity via RNA editing. C. R. Biol., 326, 901–908.
2. Smith, H.C., Gott, J.M. and Hanson, M.R. (1997) A guide to RNA editing. RNA, 3, 1105–1123.
3. Brennicke, A., Marchfelder, A. and Binder, S. (1999) RNA editing. FEMS Microbiol. Rev., 23, 297–316.
4. Gray, M.W. (2003) Diversity and evolution of mitochondrial RNA editing systems. J. Mol. Evol., 55, 227–233.
5. Takano, H., Abc, T., Sakurai, R., Moriyama, Y., Nozaki, H., Kawanoue, S., Sasaki, N. and Kuroiwa, T. (2001) The complete DNA sequence of the mitochondrial genome of Physarum polycephalum. Mol. Gen. Genet., 264, 539–545.
6. Mahendran, R., Spottswood, M.R. and Miller, D.L. (1991) RNA editing by cytidine insertion in mitochondria of Physarum polycephalum. Nature, 349, 434–438.
7. Gott, J.M., Visomirski, L.M. and Hunter, J.L. (1993) Substitutional and insertional RNA editing of the cytochrome c oxidase subunit 1 mRNA of Physarum polycephalum. J. Biol. Chem., 268, 25483–25486.
8. Miller, D., Mahendran, R., Spottswood, M., Costandy, H., Wand, S., Ling, M.L. and Yang, N. (1993) Insertional editing in mitochondria of Physarum. Semin. Cell Biol., 4, 261–266.
9. Wang, S.S., Mahendran, R. and Miller, D.L. (1999) Editing of cytochrome b mRNA in Physarum mitochondria. J. Biol. Chem., 274, 2725–2731.
10. Bundschuh, R. (2004) Computational prediction of RNA editing sites. Bioinformatics, 20, 3214–3220.
11. Henikoff, S. and Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks. Proc. Natl Acad. Sci. USA, 89, 10915–10919.
12. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res., 25, 3389–3402.
13. Gish, W. and States, D.J. (1993) Identification of protein coding regions by database similarity search. Nat. Genet., 3, 266–272.
14. Worley, K.C., Wiese, B.A. and Smith, R.F. (1995) BEAUTY: An enhanced BLAST-based search tool that integrates multiple biological information resources into sequence similarity search results. Genome Res., 5, 173–184.
15. Byrne, E.M. and Gott, J.M. (2004) Cotranscriptional editing of Physarum mitochondrial RNA requires local features of the native template. RNA, 8, 1174–1185.
16. Simpson, L., Neckelmann, N., de la Cruz, V.F., Simpson, A., Feagin, J., Jasmer, D.P. and Stuart, K. (1987) Comparison of the maxicircle (mitochondrial) genomes of Leishmania tarentolae and Trypanosoma brucei at the level of nucleotide sequence. J. Biol. Chem., 262, 6182–6196.
17. Horton, T.L. and Landweber, L.F. (2000) Evolution of four types of RNA editing in myxomycetes. RNA, 6, 1339–1346.
18. Alfonzo, J.D., Thiemann, O. and Simpson, L. (1997) The mechanism of U insertion/deletion RNA editing in kinetoplastid mitochondria. Nucleic Acids Res., 25, 3751–3759.
19. von Haesler, A., Blum, B., Simpson, L., Sturm, N. and Waterman, M.S. (1992) Computer methods for locating kinetoplastid cryptogenes. Nucleic Acids Res., 20, 2717–2724.
20. Cheng, Y.W., Visomirski-Robic, L.M. and Gott, J.M. (2001) Nontemplated addition of nucleotides to the 3′ end of nascent RNA during RNA editing in Physarum. EMBO J., 20, 1405–1414.
21. Gott, J.M. and Visomirski-Robic, L.M. (1998) RNA editing in Physarum mitochondria. In Grosjean, H. and Benne, R. (eds), Modification and Editing of RNA. ASM Press, Washington, DC, pp. 395–411.
22. Byrne, E.M., Stout, A. and Gott, J.M. (2002) Editing site recognition and nucleotide insertion are separable processes in Physarum mitochondria. EMBO J., 21, 6154–6161.