PCP
PLANT & CELL PHYSIOLOGY

Rapid Paper

RESOPS: A Database for Analyzing the Correspondence of RNA Editing Sites to Protein Three-Dimensional Structures

Kei Yura[1,2,*], Sintawee Sulaiman[3], Yosuke Hatta[4], Masafumi Shionyu[4] and Mitiko Go[4,5,6]

[1]Computational Biology, Graduate School of Humanities and Sciences, Ochanomizu University, 2-1-1 Otsuka, Bunkyo, Tokyo, 112-8610 Japan
[2]Center for Informational Biology, Ochanomizu University, 2-1-1 Otsuka, Bunkyo, Tokyo, 112-8610 Japan
[3]Bioinformatics Program, King Mongkut's University of Technology Thonburi, 126 Pracha-utid Road, Bangmod, Toongkru, Thailand
[4]Department of Bioscience, Faculty of Bioscience, Nagahama Institute of Bio-Science and Technology, 1266 Tamura, Nagahama, Shiga, 526-0829 Japan
[5]Tokyo Medical and Dental University, 1-5-45 Yushima, Bunkyo, Tokyo, 113-8510 Japan
[6]Research Organization of Information and Systems, 4-3-13, Toranomon, Minato, Tokyo, 105-0001 Japan

Transcripts from mitochondrial and chloroplast DNA of land plants often undergo cytidine to uridine conversion-type RNA editing events. RESOPS is a newly built database that specializes in displaying RNA editing sites of land plant organelles on protein three-dimensional (3D) structures to help elucidate the mechanisms of RNA editing for gene expression regulation. RESOPS contains the following information: unedited and edited cDNA sequences with notes for the target nucleotides of RNA editing, conceptual translation from the edited cDNA sequence in pseudo-UniProt format, a list of proteins under the influence of RNA editing, multiple amino acid sequence alignments of edited proteins, the location of amino acid residues coded by codons under the influence of RNA editing in protein 3D structures and the statistics of biased distributions of the edited residues with respect to protein structures. Most of the data processing procedures are automated; hence, it is easy to keep abreast of updated genome and protein 3D structural data. In the RESOPS database, we clarified that the locations of residues switched by RNA editing are significantly biased to protein structural cores. The integration of different types of data in the database also help advance the understanding of RNA editing mechanisms. RESOPS is accessible at http://cib.cf.ocha.ac.jp/RNAEDITING/.

Keywords: Chloroplast • Mitochondrion • Molecular evolution • Organelle genome • Protein 3D structure • RNA editing.

Abbreviations: 3D, three-dimensional; PDB, Protein Data Bank.

Introduction

RNA editing is a process that inserts, deletes and converts nucleotides in RNA after transcription, distinct from RNA splicing (Gray and Covello 1993, Gott and Emeson 2000, Keegan et al. 2001). The conversion type of RNA editing was first discovered in mammalian mRNA for apolipoprotein B (*apoB*) (Chen et al. 1987, Powell et al. 1987), but most of the known cytidine to uridine conversion-type RNA editing events are mainly found on mRNAs transcribed from mitochondrial and chloroplast DNA of land plants (Covello and Gray 1989, Hoch et al. 1991, Hiesel et al. 1994, Wakasugi et al. 1996,Yoshinaga et al. 1996, Freyer et al. 1997, Giege and Brennicke 1999, Kugita et al. 2003). In hornwort chloroplasts, uridine to cytidine conversion was also found (Kugita et al. 2003). RNA editing is not a rare event. The *Anthoceros formosae* chloroplast genome has at least 942 RNA editing sites (Kugita et al. 2003), and the *Arabidopsis thaliana*

Editor-in-Chief's choice

*Corresponding author: E-mail, yura.kei@ocha.ac.jp; Fax: +81-3-5978-5514.

Plant Cell Physiol. 50(11): 1865–1873 (2009) doi:10.1093/pcp/pcp132, available FREE online at www.pcp.oxfordjournals.org

mitochondrial genome has at least 441 RNA editing sites (Giege and Brennicke 1999). Most of these conversions occur in protein-coding regions, suggesting that RNA editing should impact protein structure and function. The top three patterns of amino acid residue conversions in RNA editing are serine to leucine, proline to leucine and serine to phenylalanine (Bock 2000), all of which are conversions from hydrophilic to hydrophobic residues. This conversion pattern further supports the notion that RNA editing has a substantial impact on protein structure and function. Many experiments have been carried out to demonstrate that the conversion of amino acid residues via RNA editing is crucial for protein function (Covello and Gray 1990, Bock et al. 1994, Bonnard and Grienenberger 1995, Phreaner et al. 1996, Zito et al. 1997, Kozaki et al. 2001, Sasaki et al. 2001); however, it was seldom the case that a converted residue was included in a protein active site (Yura and Go 2008). Hence, the molecular mechanism for function regulation via RNA editing has not been clarified.

Genome sequencing and structural genomics projects have produced massive quantities of data, including RNA editing sites, organelle genome sequences and protein three-dimensional (3D) structures. Based on these data, we reported previously that amino acid residues that are converted by RNA editing (hereafter called edited residues) tend to be located in protein structural cores (Yura and Go 2008). Combinations of genome and protein structure data enabled us to determine that the locations of edited residues were significantly biased toward the structurally important sites of proteins. RNA editing, therefore, seems to regulate protein function through protein folding, because in general when a protein has a hydrophilic mutation in the protein structural core, the protein becomes unstable at best and does not fold at worst (Vos et al. 2001, Loladze et al. 2002).

The molecular mechanism of the regulation suggested above is based on current advances in data production from omics analysis, and a suggested mechanism should be continuously tested as data are augmented by new results. In addition, combining data related to RNA editing will advance our understanding of the mechanisms and origin of RNA editing in land plant organelles, allowing, for example, the development of RNA editing site prediction methods (Cummings and Myers 2004, Mower 2005, Thompson and Gopal 2006, Du et al. 2007, Yura et al. 2008, Du et al. 2009). So far, there are no databases providing information about the relationship between RNA editing sites and protein 3D structures, multiple sequence alignments of homologous proteins or statistics on RNA editing sites. We therefore launched RESOPS, a database of RNA editing sites of land plant organelles that contains up-to-date RNA editing site raw data, multiple amino acid sequence alignments with editing site information in detail and edited residues in protein 3D structures. The database is freely accessible at http://cib.cf.ocha.ac.jp/RNAEDITING/.

Results

Collection of RNA editing sites from the GenBank and PDB database

In the August 2009 version of RESOPS, based mainly on the GenBank database release 172, there are 710 entries that contain at least one edited residue in an amino acid sequence from plant mitochondria and chloroplasts. A single flat file with 710 entries in pseudo-UniProt format, containing amino acid and cDNA sequences marked with RNA editing sites, can be obtained from the download page. The download page describes the details of the format and the history of manual corrections. A comparison between homologous sequences in the data set is performed via the construction of multiple sequence alignments.

The current data contain 5,754 RNA editing sites, of which 2,059 (35.8%) sites are located on the first letter of a codon, 3,165 (55.0%) are on the second letter and 530 (9.2%) are on the third letter. These figures are dynamically calculated by summing over the alignment data. The distribution of the RNA editing sites on codons is similar to a distribution calculated previously (Bock 2000).

RNA editing events frequently convert coded amino acid residues, because >90% of RNA editing sites are located on either the first or the second letter of codons. The conversion pattern of amino acid residues is automatically tabulated from the flat file as shown in **Fig. 1**. The most frequent conversion in amino acid residues is from serine to leucine, followed by proline to leucine and serine to phenylalanine. The trend of altering from hydrophilic to hydrophobic residues, mentioned before (Gray and Covello 1993, Bock 2000, Yura and Go 2008), still holds.

RESOPS stores data for the location of edited residues in both the primary and tertiary structures (**Fig. 2**). Edited residues are shown in color in the multiple amino acid sequence alignment. If the first letter of the codon is edited then the residue is colored in red, if the second letter then in green and if the third letter then in blue. If more than one letter is edited, then the residue is in the mixed color. In a multiple sequence alignment, the conservation patterns of edited residues amongst species show that RNA editing improves sequence identity among homologous proteins. This evidence further supports the notion that RNA editing is a process of 'transcript repair' (Bock 2000). When a group of homologous proteins includes one protein for which the 3D structure has been determined and stored in the Protein Data Bank (PDB) (Berman et al. 2003), the amino acid sequence of the structurally determined protein is shown at the top of the alignment. This alignment forms the basis for mapping edited residues onto protein 3D structures. The 3D structure of a protein is shown as a ribbon model, in which each chain is in a different color, and residues corresponding to the edited residues are marked by space-filling

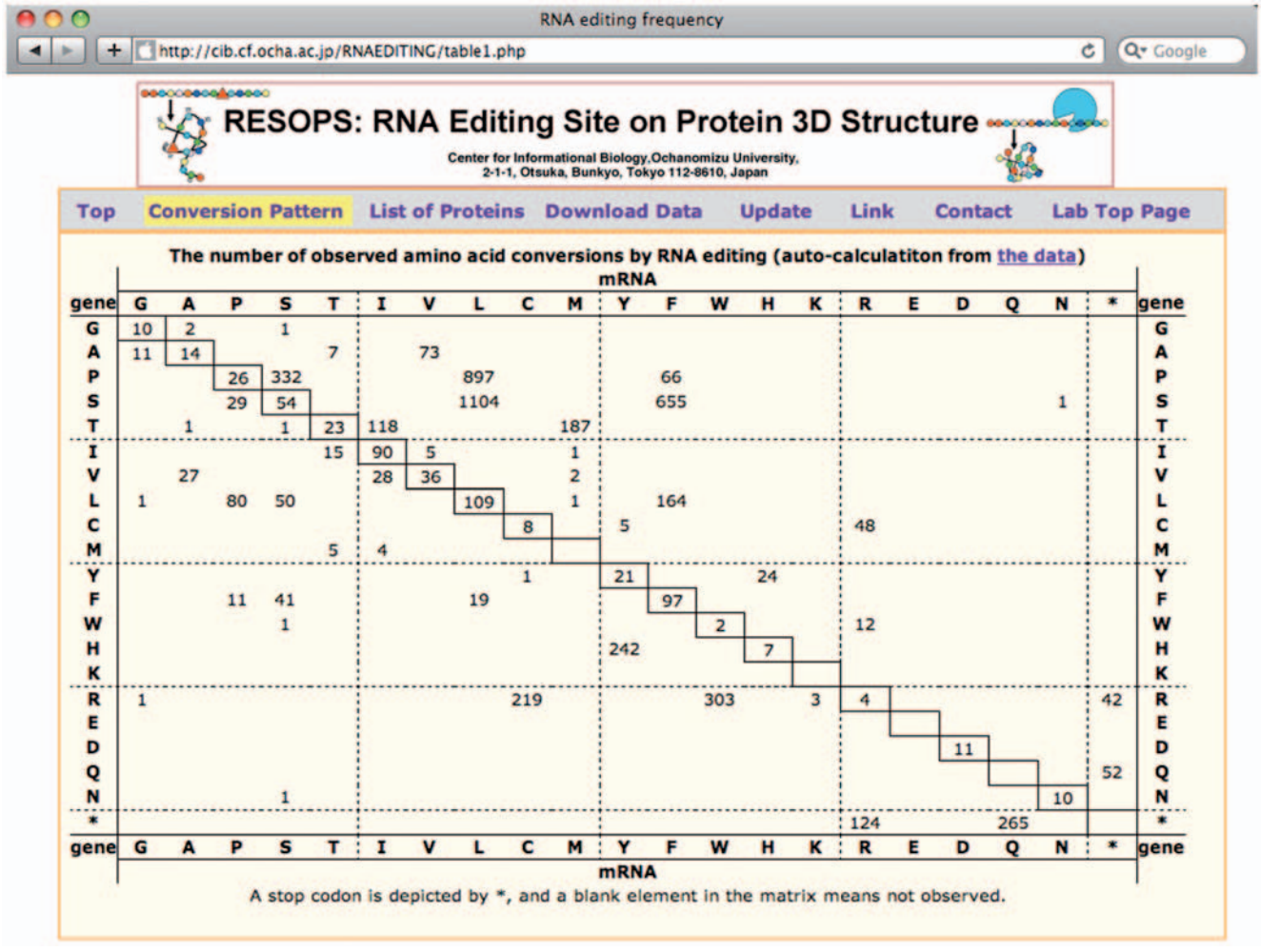

Fig. 1 Conversion patterns of amino acid residues by RNA editing. The vertical axis is the type of amino acid coded in the genome, and the horizontal axis is the type of amino acid in mature mRNAs. The table is dynamically generated from the flat file data each time a user accesses the website. The trend in the pattern is very similar to the trend previously reported.

representations using the molecular graphic software, Jmol (http://www.jmol.org/). The edited residues that reside in the protein structural core are shown in purple, and the others are in blue.

Bias of RNA editing sites toward a protein structural core

It was shown that the location of edited residues was significantly biased in favor of the protein structural core (Yura and Go 2008). In this database, the statistical test for this biased distribution can be automatically performed. In the August 2009 version of RESOPS, 3D structures of 48 groups of proteins were assigned. In these 48 proteins, 1,985 residues resided in the structural cores (41 residues per protein) and 14,290 residues were categorized as non-core residues. Therefore, about 12% of residues were categorized as residues in the structural cores and 88% were non-core residues. Multiple sequence alignments in RESOPS were able to map edited residues onto a protein 3D structure. It was found that 251 out of 1,277 edited residues resided in protein structural cores, and 1,026 were non-core residues. The expected number of residues in structural cores, based on a random distribution model, is ~153 ($=1,277\times0.12$), whereas the number of expected non-core residues is ~1,124 ($=1,277\times0.88$). A χ^2 test with one degree of freedom yields 66.3 ($P<3.8\times10^{-16}$). This result indicates that the distribution of edited residues is biased toward protein structural cores in the current data set. The biased distribution of edited residues to the protein structural core might be derived from the fact that edited residues tend to be hydrophobic residues and that hydrophobic residues tend to be buried inside the protein. RESOPS has a function to test automatically the distribution of hydrophobic residues only (phenylalanine and leucine), which eliminates the inherent

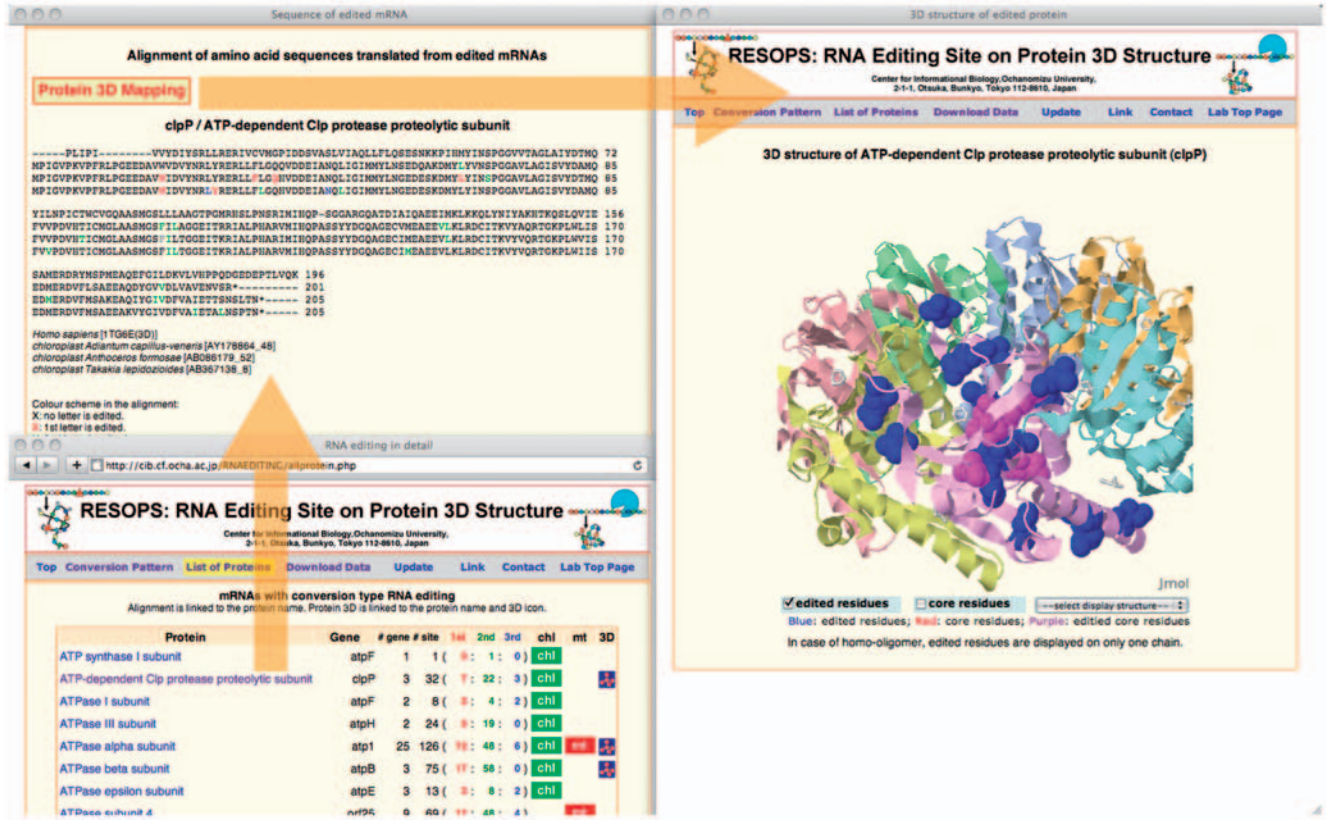

Fig. 2 A typical analysis in RESOPS. A user can find a protein/gene name in the List of Proteins if the protein expression is influenced by RNA editing. The List of Proteins also indicates the number of genes, locations of the edited sites in a codon, the origin (chloroplast and mitochondrion) of the sequence and whether a 3D structure of the protein is known. The multiple sequence alignment of homologous edited proteins is linked to the name of the protein. If the 3D structure of the protein is known, then all edited residues are mapped to the 3D structure. The 3D structure is also directly accessible by an icon link in the 3D column of the List of Proteins.

biased distribution of hydrophobic residues in protein 3D structures from the test. A χ^2 test on the adjusted data set still yields 10.3 ($P < 1.3 \times 10^{-3}$), and the distribution of the edited hydrophobic residues is found to be significantly biased toward protein structural cores.

Discussion

Impact of RNA editing on protein structure and function

The role of RNA editing in plant organelles was suggested to be a means of regulating organellar protein expression. A number of experiments were performed to test the function of unedited proteins, most of which turned out to be less functional than the edited proteins (Covello and Gray 1990, Bock et al. 1994, Bonnard and Grienenberger 1995, Phreaner et al. 1996, Zito et al. 1997, Kozaki et al. 2001, Sasaki et al. 2001). However, the molecular mechanism for regulation has yet to be uncovered, because only a few of the edited sites comprise the active sites of proteins (Yura and Go 2008). Based on the biased distribution of edited residues toward protein structural cores, we suggest that the expression of function should be regulated via protein folding, because unedited proteins tend to contain more hydrophilic residues in the parts that are supposed to be protein structural cores. Mutation to hydrophilic residues in the protein structural core destabilizes the protein, because the hydrophobic core is required to build a functional protein 3D structure (Vos et al. 2001, Loladze et al. 2002). An unedited protein has, on average, two to three hydrophilic mutations in its protein structural core, and a single hydrophilic mutation in a protein structural core destabilizes proteins by ~5 kcal mol^{-1}, which is comparable in magnitude with the reduction of free energy in protein folding, ~10–15 kcal mol^{-1} (Creighton 1990).

Kotera et al. (2005) identified a nuclear protein CRR4 involved in RNA editing in *A. thaliana*. It was shown that CRR4 protein was specifically involved in RNA editing on the initiation codon of *ndhD*, because mutation of CRR4 changed the extent of the RNA editing. Following this study, many

Table 1 Correspondence between functional/structural effects by suppressing RNA editing and the suppressed RNA editing sites on protein 3D structures

Gene	Position[a]	Effect	3D[b]	Reference
rpoA	C200	May prevent PEP assembly	Core	Chateigner-Boutin et al. (2008)
rpoB	C338	Partial loss of activity	Non-core	Zhou et al. (2009)
clpP	C559	Unclear	Surface	Chateigner-Boutin et al. (2008)
cox2	C167	Malfunction in electron transport chain	Surface	Kim et al. (2009)
cox3	C572	Malfunction in electron transport chain	Core	Kim et al. (2009)
rps4	C954	No effect	ND	Zehrmann et al. (2009)
accD	C794	Low solubility	Core	Sasaki et al. (2001)
		Albino phenotype		Yu et al. (2009)
		No effect		Robbins et al. (2009)

[a]The position number is taken from the original paper and it is the nucleotide number of the edited nucleotide in mRNA.
[b]ND indicates that the corresponding residue is not included in the 3D structure determined. Each 3D structure is given in **Supplementary data 1–6**.

nuclear proteins involved in RNA editing of chloroplasts and mitochondria were identified (Chateigner-Boutin et al. 2008, Cai et al. 2009, Kim et al. 2009, Robbins et al. 2009, Yu et al. 2009, Zehrmann et al. 2009, Zhou et al. 2009). These studies identified the target sites of the nuclear proteins for RNA editing, and the functional effect of mutation on the nuclear proteins, mainly the impact of suppressing RNA editing of the target sites. We found that many effects in these cases could be qualitatively explained based on protein 3D structures in RESOPS. The result is summarized in **Table 1** and the details are described below.

Chateigner-Boutin et al. (2008) speculated that abolishing RNA editing on amino acid residue 67 of RpoA in Arabidopsis chloroplast mutants may prevent assembly of plastid-encoded RNA polymerase (PEP). The speculation implies that RpoA becomes unstable. In RESOPS, we find that the residue forms a protein structural core of RpoA, and alteration of the residue to a small hydrophilic amino acid probably destabilizes the protein, and hence affects interactions with other subunits of the polymerase (**Supplementary data 1**). Zhou et al, (2009) showed that *ys1* mutants had a defect in RNA editing of *rpoB* in Arabidopsis chloroplast and that the defect possibly caused a partial loss of RpoB activity. In RESOPS, we find that the residue is buried, but not in a structural core (**Supplementary data 2**) and hence the alteration of the residue probably has a partial impact on protein stability. Chateigner-Boutin et al. (2008) found that their *clb19* mutants abolished one of the RNA editing events on *clpP*. The impact of abolishing the RNA editing event on *clpP* was not clear in their work. In RESOPS, the edited residue is found on the surface of the protein, even though it is a hydrophobic residue (**Supplementary data 3**). We speculate that ClpP is stable and functional in the mutant. Kim et al (2009) demonstrated that rice *ogr1* mutants had defects in RNA editing of *cox2* and *cox3* and speculated that the defects caused malfunction in the mitochondrial electron transport chain. The edited residue in Cox2 is found on the surface of a transmembrane helix, which suggests that the residue is in contact with membrane lipids (**Supplementary data 4**). The edited residue in Cox3 is found in the protein structural core, in the internal interfaces of the helix bundle (**Supplementary data 5**). The structural data, therefore, suggest that the mutation on Cox3 should have a more significant impact on protein function than that on Cox2. Robbins et al. (2009) showed that *rare1* mutants abolished RNA editing at C794 of *accD*. The mutants were unexpectedly robust and they suggested that RNA editing at C794 of *accD* was not essential for acetyl-CoA carboxylase activity, or that other carboxylases should compensate for the loss of *accD* function. In RESOPS, we find that the edited residue is included in a protein structural core, and mutation of the residue evidently has an impact on protein stability (**Supplementary data 6**). Our analysis is consistent with previous works by Sasaki et al. (2001) and Yu et al. (2009), and we suggest that the second suggestion by Robbins et al. (2009) is much more likely than the first one.

Possible origin of RNA editing in land plant organelles

Multiple sequence alignment of the edited proteins suggests a multiple origin of RNA editing in organelles. Most of the sites with RNA editing are not unanimously edited in homologous proteins. When the type of amino acid is compared at each site, amino acids of non-edited sequences are almost always the same as the residues of the edited sequence, but not the unedited sequence. This suggests that RNA editing was not introduced into the non-edited sequences at

the site. If RNA editing had been introduced in the common ancestor of the genes, and if the current non-edited sequence had lost its RNA editing mechanism, then the type of amino acid residue should be the same as the type of unedited amino acid. Hence, this observation suggests that RNA editing should be introduced at a site in the most recent common ancestor of the genes that share RNA editing sites at the same position, which also suggests that the introduction of an RNA editing site should have occurred many times in many genes. It is well known that RNA editing in plant organelles has only been found in land plants (Gray and Covello 1993, Bock 2000). This suggests that RNA editing was introduced at the time land plants came into being (Yoshinaga et al. 1996). Because RNA editing is introduced later than the time that plants acquired two organelles, it should be rare to find RNA editing events in homologous sites of proteins in mitochondria and chloroplasts. By checking through multiple sequence alignments in RESOPS, however, we found 12 such events in five genes, as shown in **Table 2**. The amino acid sequence alignment of *ndhC*/*nad3* products is shown in **Fig. 3**. Other alignments are given in **Supplementary data 7**. These correspondences could reflect preferred sites for introduction of RNA editing events.

Table 2 RNA editing events in homologous sites in genes from chloroplasts and mitochondria

Gene name	Position in alignment[a]	Chloroplast	Mitochondrion
rps7	195 (2)	S→L	S→L
rps19	16 (2)	S→L	S→L
ndhB/nad2	186 (2)	T→M	S→L
	242 (2)	S→F	S→F
	285 (2)	S→L	S→F
ndhC/nad3	130 (2)	S→F	S→F
	144 (2)	S→L	S→F
	153 (2)	S→L	S→L
petB/cob	64 (2)	T→M	S→L
	68 (1)	H→Y	H→Y
	116 (2)	S→L	S→F/F→S
	209 (2)	T→M	W→S

[a]Detail of the amino acid sequence alignment is given in **Supplementary data 7**. The number in parentheses is the position of the edited nucleotide in the codon.

```
chloroplast Adiantum capillus-veneris  1 MFLSHQYDSFWIFLLVCISIPLLAFSITRFAAPPREG--PEKSTSYESGIE
chloroplast Anthoceros formosae        1 MFLVSKYNYFWIFLLIASLIPTIAFSISRVIAPISKG--PEKFTSYECGIE
chloroplast Takakia lepidozioides      1 MFLLPKNDSLWIFLLITSLIPITAFSISKIIAPVSEG--PEKFTSYESGIE
mitochondrion Arabidopsis thaliana     1 --MMSEFAPISIYLVISLLVSLILLGVPFLFA-SNSSTYPEKLSAYECGFD
mitochondrion Beta vulgaris            1 ---MLEFAPICIYLVISLLVSLILLGVPFLFS-SNTSTYPEKLSAYECGFD
mitochondrion Brassica napus           1 ---MLEFAPIFIYLVISLLVSLILLGVPFLFA-SNSSTYPEKLSAYECGFD
mitochondrion Carthamus tinctorius     1 ---MLEFAPIFIYLVISLLVSLILLGVPFLFA-SNSSTYPEKLSAYECGFD
mitochondrion Megaceros aenigmaticus   1 ----MEFVPICVYLVISSLLPLILIGVSLLFAFPSSSAYPEKLSAHECGSD
mitochondrion Oryza sativa            35 SAFLSEFAPICIYLVISLLVSLILLGVPFLFA-SNSSTYPEKLSAYECGFD
mitochondrion Triticum aestivum        1 ---MLEFAPICIYLVISLLVSLILLGVPFLFA-SNSSTYPEKLSAYECGFD
mitochondrion Zea mays                 1 ---MLEFAPICIYLVISLLVSLILLGVPFLFA-SNSSTYPEKLSAYECGFD
mitochondrion Sorghum bicolor          1 ---MSEFAPICIYLVISLLVSLILLGVPFLFA-SNSSTYPEKLSAYECGFD

                                                  130         144       153
50 PKGDTWIRFQIRYYMFALVFTVFDVETVFLYPWATSFEELGLFAFVEVIVFIFILIVGLVYAWRKGALEWC* [AY178864_29]
50 PMGDAWIQFQIRYYMFALVFVIFDVETVFLYPWAMSFKQLGIPAFIEVFIFVFILIIGLIYAWRKGALEWS* [AB086179_33]
50 PMGDAWIQFHIRYYMFALVLVIFDVETVFLYPWAMSFNELGISAFIEALIFVSILIIGLVYAWRKGALEWS* [AB299142_4]
49 PFGDARSRFDIRFYLVSILFLIFDLEVTFFFPWAVSLNKIDLFGFWSMMAFLFILTIGFLYEWKRGALDWE* [Y08501_82]
48 PFGDARSRFDIRFYLVSILFIIFDLEVTFFFPWAVSLNKIDLFGFWSMMAFLLILTIGFLYEWKRGALDWE* [BA000009_100]
48 PFGDARSRFDIRFYLVSILFLIFDLEVTFFFPWAVSLNKIDLFGFWSMMAFLFILTIGFLYEWKRGALDWE* [AP006444_50]
48 PFGDARSRFDIRFYLVSILFIIFDLEVTFFFPWAVSLNKIDLFGFWSMMAFLLILTIGFLYEWKRGALDWE* [DQ534204_1]
48 PSDDARSRFDIRLYLVLTSSIISDSEVTSSFPWAVPPNKIGLFGSWSMMVFSLISTIGFVYEWKKGASDWE* [EU660574_18]
85 PFGDARSRFDIRFYLVSILFIIFDLEVTFFFPWAVSLNKIDLFGFWSMMAFLLILTIGFLYEWKRGALDWE* [BA000029_20]
48 PFGDARSRFDIRFYLVSILFIIFDLEVTFFFPWAVSLNKIDLFGFWSMMAFLLILTIGFLYEWKRGALDWE* [X59153_1]
48 PFGDARSRFDIRFYLVSILFIIFDLEVTFFFPWAVSLNKIDLFGFWSMMAFLLILMIGSLYEWKRGALDWE* [AY506529_90]
48 PFGDARSRFDIRFYPVSILFIIFDLEVTFFFPWAVSLNKIDLFGFWSMMAFLLILTIGFLYEWKRGALDWE* [Z85978_1]
```

Fig. 3 Amino acid sequence alignment of *ndhC*/*nad3* products. An edited residue is differently colored based on the edited letter in the codon. When the first letter is edited then the residue is colored in red, the second letter in green and the third letter in blue. When the first and the second letters are edited, then the residue is colored in orange. The top three sequences are derived from chloroplast genomes and the others are from mitochondrion genomes. Three positions with a triangle have edited residues in sequences from both chloroplasts and mitochondria. The number on the triangle is the alignment position given in the RESOPS database.

Future development of the database

RESOPS will be updated regularly following the major update of GenBank every 2 months. The procedure for updating is automatic, except for the initial process of adding data from the literature and of checking the consistency of the GenBank database (see Materials and Methods). We hope that the inconsistencies in the public database may be resolved by the original depositors in the near future. In the last 2 years, we have seen some corrections introduced into the annotations of RNA editing events listed in the GenBank database. To promote corrections, we continue to contact the original depositors when we find ambiguous annotations. RESOPS will also be upgraded as a tool for mapping RNA editing sites on protein 3D structures in the future.

Materials and Methods

Collection of nucleotide sequences with conversion-type RNA editing in plant organelles

GenBank/EMBL/DDBJ (Cochrane et al. 2008, Benson et al. 2009, Sugawara et al. 2009) stores the nucleotide position numbers for RNA editing sites without a standardized description and, therefore, interpretation of the collection of RNA editing sites in nucleotide and amino acid sequences is not straightforward. Manual inspection, with the aid of in-house C programs, was performed to decipher the GenBank/EMBL/DDBJ database descriptions, specifically for plant organelle conversion-type RNA editing site descriptions. A whole character search was performed to find a string of characters that matched 'RNA' and 'editing' in the '/note' field of 'misc_feature' lines for plant entries in the GenBank database release 172. The C program then extracted protein-coding regions with RNA editing sites, generated both edited and unedited cDNA sequences, and translated edited mRNAs into amino acid sequences.

Some of the entries contained an error in the nucleotide position number of the RNA editing site, or a discrepancy between the types of nucleotide described in the misc_feature line and the corresponding nucleotide in the deposited nucleotide sequence. In these cases, manual correction was done based either on the literature or by communication with the depositors. In the GenBank database release 172, corrections were needed to AJ006146, BA000029, DQ645537, DQ984517, X07566, X69720, X80170, X92735, X96536, Y14434, Y14435 and Y17812. We could not correct all errors encountered, because we could not make contact with all depositors. The entries with errors were discarded. We started the error correction procedure about 3 years ago, and the depositors of AB254134, AY521591 and AY820131 have evidently made contact with GenBank to rectify the annotations; the annotations of these three entries are corrected in the latest version of GenBank.

Manual checks also included the curation of RNA editing information from the literature and a check for duplicated data. For RNA editing on *rbcL* transcripts from a number of different species, we copied the RNA editing site described in the table from the literature (Freyer et al. 1997) into the following entries: D14882, D43696, L11055, L11056 and L13485. Occasionally, the same gene and cDNA were sequenced by different groups and independently deposited with different IDs. These entries were stored as they were, because the editing sites may differ, even if the sequences were the same.

Multiple amino acid sequence alignments and protein 3D structures

Amino acid sequences were clustered based on sequence identity. When a cluster contained more than one sequence, a sequence in the cluster had at least one different sequence with identity no less than 25%. Representative sequences of each cluster were then used as a query to find homologous proteins in the PDB (Berman et al. 2003) with BLAST (Altschul et al. 1997). When the amino acid sequences with identity no less than 30% were found, we selected the largest structure in the PDB with the highest sequence identity for assigning structural properties to the amino acid sequences in the cluster. Multiple sequence alignments, including amino acid sequences of proteins from the PDB, were then performed for each cluster, and edited residues were located both in the alignment and in the protein 3D structure.

Identification of protein structural cores

A structural core was determined by identifying clusters of buried residues and peripheral residues as described previously (Yura and Go 2008). The solvent-accessible surface area of each residue was calculated (Shrake and Rupley 1973) and solvent-inaccessible residues were identified first. When carbon atoms from two different solvent-inaccessible residues were in contact (≤4.0 Å), then the pair of residues was defined as a cluster. In the next step, every carbon atom in residues with accessibilities to solvent molecules (Go and Miyazawa 1980) between 0 and 0.05 was selected, and if the atom was in contact (≤4.0 Å) with one of the carbon atoms in the cluster residues, then the residue not in the cluster was defined as peripheral. Both cluster and peripheral residues were defined as structural core residues.

Supplementary data

Supplementary data are available at PCP online.

Funding

Japan Society for the Promotion of Science KAKENHI [Grant-in-Aid for Scientific Research (B) No. 18370061 to

M.G.]; University Education Internationalization Promotion Program of the Ministry of Education, Culture, Sports, Science, and Technology-Japan (to Ochanomizu University).

Acknowledgments

K.Y. and M.G. thank the late Professor Hans Kössel for introducing us to his RNA editing work in RuBisCO large subunit transcripts when he visited Nagoya University. K.Y. also thanks Mr. Kazuhiro Kobayashi, Ms. Kazuko Kaji and Ms. Atsuko Doi for gathering RNA editing data from the literature.

References

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., et al. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* 25: 3389–3402.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Sayers, E.W. (2009) GenBank. *Nucleic Acids Res.* 37: D26–D31.

Berman, H.M., Henrick, K. and Nakamura, H. (2003) Announcing the worldwide Protein Data Bank. *Nature Struct. Biol.* 10: 980.

Bock, R., Kössel, H. and Maliga, P. (1994) Introduction of a heterologous editing site into the tobacco plastid genome: the lack of RNA editing leads to a mutant phenotype. *EMBO J.* 13: 4623–4628.

Bock, R. (2000) Sense from nonsense: how the genetic information of chloroplasts is altered by RNA editing. *Biochimie* 82: 549–557.

Bonnard, G. and Grienenberger, J.M. (1995) A gene proposed to encode a transmembrane domain of an ABC transporter is expressed in wheat mitochondria. *Mol. Gen. Genet.* 246: 91–99.

Cai, W., Ji, D., Peng, L., Guo, J., Ma, J., Zou, M., et al. (2009) LPA66 is required for editing *psbF* chloroplast transcripts in Arabidopsis. *Plant Physiol.* 150: 1260–1271.

Chateigner-Boutin, A.-L., Ramos-Vega, M., Guevara-García, A., Andrés, C., Gutiérrez-Nava, M.L., Cantero, A., et al. (2008) CLB19, a pentatricopeptide repeat protein required for editing of *rpoA* and *clpP* chloroplast transcripts. *Plant J.* 56: 590–602.

Chen, S.H., Habib, G., Yang, C.Y., Gu, Z.W., Lee, B.R., Weng, S.A., et al. (1987) Apolipoprotein B-48 is the product of a messenger RNA with an organ-specific in-frame stop codon. *Science* 238: 363–366.

Cochrane, G., Akhtar, R., Aldebert, P., Althorpe, N., Baldwin, A., Bates, K., et al. (2008) Priorities for nucleotide trace, sequence and annotation data capture at the Ensembl Trace Archive and the EMBL Nucleotide Sequence Database. *Nucleic Acids Res.* 36: D5–D12.

Covello, P.S. and Gray, M.W. (1989) RNA editing in plant mitochondria. *Nature* 341: 662–666.

Covello, P.S. and Gray, M.W. (1990) RNA sequence and the nature of the CuA-binding site in cytochrome *c* oxidase. *FEBS Lett.* 268: 5–7.

Creighton, T.E. (1990) Protein folding. *Biochem. J.* 270: 1–16.

Cummings, M.P. and Myers, D.S. (2004) Simple statistical models predict C-to-U edited sites in plant mitochondrial RNA. *BMC Bioinformatics* 5: 132.

Du, P., He, T. and Li, Y. (2007) Prediction of C-to-U RNA editing sites in higher plant mitochondria using only nucleotide sequence features. *Biochem. Biophys. Res. Commun.* 358: 336–341.

Du, P., Jia, L. and Li, Y. (2009) CURE-Chloroplast: a chloroplast C-to-U RNA editing predictor for seed plants. *BMC Bioinformatics* 10: 135.

Freyer, R., Kiefer-Meyer, M.-C. and Kössel, H. (1997) Occurrence of plastid RNA editing in all major lineages of land plants. *Proc. Natl Acad. Sci. USA* 94: 6285–6290.

Giege, P. and Brennicke, A. (1999) RNA editing in Arabidopsis mitochondria effects 441C to U changes in ORFs. *Proc. Natl Acad. Sci. USA* 96: 15324–15329.

Go, M. and Miyazawa, S. (1980) Relationship between mutability, polarity and exteriority of amino acid residues in protein evolution. *Int. J. Pept. Protein Res.* 15: 211–224.

Gott, J.M. and Emeson, R.B. (2000) Functions and mechanisms of RNA editing. *Annu. Rev. Genet.* 34: 499–531.

Gray, M.W. and Covello, P.S. (1993) RNA editing in plant mitochondria and chloroplasts. *FASEB J.* 7: 64–71.

Hiesel, R., Combettes, B. and Brennicke, A. (1994) Evidence for RNA editing in mitochondria of all major groups of land plants except the Bryophyta. *Proc. Natl Acad. Sci. USA* 91: 629–633.

Hoch, B., Maier, R.M., Appel, K., Igloi, G.L. and Kössel, H. (1991) Editing of a chloroplast mRNA by creation of an initiation codon. *Nature* 353: 178–180.

Keegan, L.P., Gallo, A. and O'Connell, M.A. (2001) The many roles of an RNA editor. *Nature Rev. Genet.* 2: 869–878.

Kim, S.-R., Yang, J.-I., Moon, S., Ryu, C.-H., An K., Kim, K.-M., et al. (2009) Rice *OGR1* encodes a pentatricopeptide repeat-DYW protein and is essential for RNA editing in mitochondria. *Plant J.* 59: 738–749.

Kotera, E., Tasaka, M. and Shikanai, T. (2005) A pentatricopeptide repeat protein is essential for RNA editing in chloroplasts. *Nature* 433: 326–330.

Kozaki, A., Mayumi, K. and Sasaki, Y. (2001) Thiol–disulfide exchange between nuclear-encoded and chloroplast-encoded subunits of pea acetyl-CoA carboxylase. *J. Biol. Chem.* 276: 39919–39925.

Kugita, M., Yamamoto, Y., Fujikawa, T., Matsumoto, T. and Yoshinaga, K. (2003) RNA editing in hornwort chloroplasts makes more than half the genes functional. *Nucleic Acids Res.* 31: 2417–2423.

Loladze, V.V., Ermolenko, D.N. and Makhatadze, G.I. (2002) Thermodynamic consequences of burial of polar and non-polar amino acid residues in the protein interior. *J. Mol. Biol.* 320: 343–357.

Mower, J.P. (2005) PREP-Mt: predictive RNA editor for plant mitochondrial genes. *BMC Bioinformatics* 6: 96.

Phreaner, C.G., Williams, M.A. and Mulligan, R.M. (1996) Incomplete editing of *rps12* transcripts results in the synthesis of polymorphic polypeptides in plant mitochondria. *Plant Cell* 8: 107–117.

Powell, L.M., Wallis, S.C., Pease, R.J., Edwards, Y.H., Knott, T.J. and Scott, J. (1987) A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. *Cell* 50: 831–840.

Robbins, J.C., Heller, W.P. and Hanson, M.R. (2009) A comparative genomics approach identifies a PPR-DYW protein that is essential for C-to-U editing of the *Arabidopsis* chloroplast *accD* transcript. *RNA* 15: 1142–1153.

Sasaki, Y., Kozaki, A., Ohmori, A., Iguchi, H. and Nagano, Y. (2001) Chloroplast RNA editing required for functional acetyl-CoA carboxylase in plants. *J. Biol. Chem.* 276: 3937–3940.

Shrake, A. and Rupley, J.A. (1973) Environment and exposure to solvent of protein atoms. Lysozyme and insulin. *J. Mol. Biol.* 79: 351–371.

Sugawara, H., Ikeo, K., Fukuchi, S., Gojobori, T. and Tateno, Y. (2009) DDBJ dealing with mass data produced by the second generation sequencer. *Nucleic Acids Res.* 37: D16–D18.

Thompson, J. and Gopal, S. (2006) Genetic algorithm learning as a robust approach to RNA editing site prediction. *BMC Bioinformatics* 7: 145.

Vos, S.D., Backmann, J., Prevost, M., Steyaert, J. and Loris, R. (2001) Hydrophobic core manipulations in ribonuclease T1. *Biochemistry* 40: 10140–10149.

Wakasugi, T., Hirose, T., Horihata, M., Tsudzuki, T., Kössel, H. and Sugiura, M. (1996) Creation of a novel protein-coding region at the RNA level in black pine chloroplasts: the pattern of RNA editing in the gymnosperm chloroplast is different from that in angiosperms. *Proc. Natl Acad. Sci. USA* 93: 8766–8770.

Yoshinaga, K., Iinuma, H., Masuzawa, T. and Ueda, K. (1996) Extensive RNA editing of U to C in addition to C to U substitution in the *rbcL* transcripts of hornwort chloroplasts and the origin of RNA editing in green plants. *Nucleic Acids Res.* 24: 1008–1014.

Yu, Q.-B., Jiang, Y., Chong, K. and Yang, Z.-N. (2009) AtECB2, a pentatricopeptide repeat protein, is required for chloroplast transcript accD RNA editing and early chloroplast biogenesis in *Arabidopsis thaliana*. *Plant J.* 59: 1011–1023.

Yura, K. and Go, M. (2008) Correlation between amino acid residues converted by RNA editing and functional residues in protein three-dimensional structures in plant organelles. *BMC Plant Biol.* 8: 79.

Yura, K., Miyata, Y., Arikawa, T., Higuchi, M. and Sugita, M. (2008) Characteristics and prediction of RNA editing sites in transcripts of the moss *Takakia lepidozioides* chloroplast. *DNA Res.* 15: 309–321.

Zehrmann, A., Verbitskiy, D., van der Merwe, J.A., Brennicke, A. and Takanaka, M. (2009) A DYW domain containing pentatricopeptide repeat protein is required for RNA editing at multiple sits in mitochondria of *Arabidopsis thaliana*. *Plant Cell* 21: 558–567.

Zhou, W., Cheng, Y., Yap, A., Chateigner-Boutin, A.-L., Delannoy, E., Hammani, K., et al. (2009) The Arabidopsis gene *YS1* encoding a DYW protein is required for editing of *rpoB* transcripts and the rapid development of chloroplasts during early growth. *Plant J.* 58: 82–96.

Zito, F., Kuras, R., Choquet, Y., Kössel, H. and Wollman, F.A. (1997) Mutations of cytochrome b_6 in *Chlamydomonas reinhardtii* disclose the functional significance for a proline to leucine conversion by *petB* editing in maize and tobacco. *Plant Mol. Biol.* 33: 79–86.

(Received September 15, 2009; Accepted September 24, 2009)