

Computational Prediction of RNA Editing Sites: Successes and Challenges Ahead

Shuba Gopal*

Assistant Professor, Bioinformatics, Department of Biological Sciences, Rochester Institute of Technology, 85 Lomb Memorial Drive, Rochester, NY 14623, USA

Abstract: RNA editing is an unusual phenomenon in which RNA transcripts are modified so that they no longer match their original, genomic template. The process occurs in a wide variety of eukaryotes from plant mitochondria to humans. Recent advances in computational and experimental methods have made it possible to survey entire genomes for candidate edit sites. Such efforts have yielded many promising results while raising tantalizing new questions. Key features of extant methods and the open questions and issues that have been raised by these efforts are highlighted in this review.

Keywords: RNA editing, A-to-I editing, C-to-U editing, computational methods.

1. INTRODUCTION

RNA editing is an unusual phenomenon in which nucleotides in the mRNA transcript are modified post-transcriptionally so that the resulting transcript no longer directly correlates with its genomic template. There are many forms of RNA editing, and they are observed in a wide spectrum of eukaryotes. This review focuses on substitution RNA editing, in which one nucleotide is replaced by another, typically as a result of a deamination reaction. Two forms are common in higher eukaryotes including humans: adenine (A) to inosine (I) (A-to-I) and cytidine (C) to uridine (U) (C-to-U). A-to-I editing in essence functions as an adenine to guanosine (G) conversion as the inosine nucleotide is read by cellular machinery as a guanosine (G). Both forms of editing involve specialized enzymes that are part of a complex that identifies sites where editing should occur and then carries out the specific deamination reaction required. The complexity of the process and the signals that drive edit site selection have been the focus of intense study both experimentally and computationally in recent years [1-3].

RNA editing was discovered two decades ago in a class of unusual eukaryotic parasites, the trypanosomatids, which extensively edit transcripts in their mitochondrial organelle, the kinetoplast [4]. Since then, a wide variety of eukaryotes have been shown to edit transcripts. Transcripts are edited in both the nucleus and in some organelles. Plant mitochondria and chloroplasts extensively edit their transcripts [2], while mammals edit transcripts in the brain and in the small intestine [1]. In humans numerous transcripts in the brain undergo A-to-I editing, and at least one transcript is regularly C-to-U edited in the small intestine [5, 6]. Additional instances of editing have been associated with disease states ranging from neurological disorders to tumor formation [7-9].

Aside from the relative recency of the discovery of this process, there are three especially striking aspects of the

RNA editing phenomenon. The first is the apparent modification of the information content of the mRNA to dramatically alter the encoded protein sequence. For example, in trypanosomatids, as many as 50% of the nucleotides are altered so that the resulting transcript has no direct correlation with its genomic counterpart [4]. Even in substitution editing, wherein a few nucleotides are altered, the resulting transcript can be dramatically different from its genomic counterpart. RNA editing can introduce splice donor and acceptor sites [10], alter the reading frame by creating a start codon [2] or prematurely truncate a protein by generating a stop codon [11]. In this sense, substitution editing may play a similar role to RNA splicing, in which sections of the pre-mRNA are removed to create a functional open reading frame from the genomic version of the coding region. However, the actual consequences of RNA editing can be dramatic and long-lasting, even more so than splicing in some cases.

A computational analogy can serve to illustrate this point. If we think of the genome as a computer program that is meant to run the cell, then RNA editing is an example of self-editing code. Self-editing code can yield unpredictable results and long-term instability in a system. Consequently, it is anathema to most programmers. In many ways, the cell's need for predictable results and stable systems is even more pressing than most computer platforms. Therefore, we might reasonably expect that modifications to cellular programs would be strictly curtailed. Indeed, we might predict that such modifications to mRNA run the risk of cellular catastrophe and should be selected against over time. So why might RNA editing have evolved at all? This question has yet to be definitively answered, but some hypotheses about the possible roles of RNA editing within the cell are discussed in Section 2.

The second intriguing aspect of substitution RNA editing is the apparent specificity with which editing occurs. Given a region of the transcript, both A-to-I and C-to-U editing will selectively edit one nucleotide while leaving neighboring candidate adenines or cytidines untouched. The editing itself is catalyzed by specialized enzymes that operate in complexes with auxiliary factors. These enzymes and associated

*Address correspondence to this author at the Assistant Professor, Bioinformatics, Department of Biological Sciences, Rochester Institute of Technology, 85 Lomb Memorial Drive, Rochester, NY 14623, USA; Tel: (585) 475-4498; Fax: (585) 475-2533; E-mail: sxgsbi@rit.edu

factors recognize edit sites, carry out the deamination reaction and may interact closely with other post-transcriptional processes such as splicing [1]. Yet it is not clear how the various components identified to date participate in edit site selection, and the exact mechanisms that underlie A-to-I and C-to-U editing are only just beginning to be elucidated. Section 5 explores our current state of understanding of RNA editing.

The third striking aspect of the RNA editing phenomenon is that so much of the progress has come from combined computational and experimental efforts. In many other areas of biological study experimental efforts lead the investigation and guide algorithmic design and implementation. In contrast, computational efforts have often been at the forefront of progress in the study of RNA editing. As described in Section 6, numerous computational methods have been developed that facilitate prediction of putative edit sites and have contributed to our understanding of signals associated with edit site selection. In this sense, RNA editing is the rare marriage of equals, and is thus particularly interesting for those with an interest in computational approaches to the study of biological phenomena.

2. THE RNA EDITOR: A JOB DESCRIPTION

There are many speculations about the role of RNA editing (see [1, 12] for some recent reviews). As explained in the ensuing sections, nearly every cellular process associated with gene regulation and expression has been impacted in one way or another by the discovery of RNA editing. In one sense, RNA editing may be a partner in every post-transcriptional processing step identified to date.

2.1. Generating Transcript Diversity

The most obvious impact of RNA editing, at least in mammalian cells, seems to be in the generation of multiple isoforms from a single gene region. In this sense, RNA editing is a sister process of alternative splicing, since both processes can generate complexity in the face of limited gene complements [9]. Processes such as editing and splicing can generate many isoforms from a single gene region, thereby doubling or tripling the number of possible proteins in a species.

This is an important consideration since the continuing paradox of the human genome project has been the deceptively low number of unique genes in the human genome; most recent estimates put this number near 25,000 genes [13]. When compared to invertebrates, such as nematodes (with 19,000 genes) and fruit flies (13,000 genes), it is difficult to reconcile the obvious complexity of the human being with the apparent penury of the human genome. The only plausible solution to this paradox is to postulate, as some have, that every gene region can generate multiple isoforms, thereby creating as many as 100,000 different protein products [14]. As some in the bioinformatics community will recall, early estimates of the number of genes in the human genome were placed at about 100,000; if every current gene region generated four isoforms, it might yet be possible to justify what has proven to be an overly optimistic gene number estimate.

Tongue-in-cheek speculations about the genuine complexity of humans aside, the problem presents a legitimate

puzzle. How do we account for complexity when gene number is relatively small? The discovery of processes such as RNA editing and splicing, which both appear to be extensive and prevalent in many higher eukaryotes, supports the notion that the complexity of higher mammals requires transcriptomic diversity. Two examples provide support for the hypothesis that RNA editing is yet another means of generating transcript diversity. The best documented instance of C-to-U editing in humans is the editing of the transcript of apolipoprotein B (apoB). This protein is expressed in the small intestine and in the liver. The transcript in the human small intestine is C-to-U edited to create a stop codon approximately half way through the transcript. The full transcript is expressed, unedited, in the liver. Each isoform has different lipid transport functions, altering the ways in which an organism regulates lipid formation, transport and deposition. In humans, this has profound consequences for the development of diseases such as atherosclerosis [11]. Thus, a single gene region can produce two distinct protein forms through RNA editing, and thereby accommodate complex lipid transport functions that would otherwise require two separate genes.

A second, more striking example of RNA editing contributing to organismal complexity has to do with A-to-I editing of neuronal transcripts. In many eukaryotes including mammals and *Drosophila*, the majority of A-to-I edit sites are observed in the central nervous system (CNS) [1]. The best documented of these in mammals is the editing of ionotropic glutamate receptors (GluR). As many as eight different A-to-I edit sites exist in the GluR transcript, generating isoforms with extremely nuanced results. Editing alters splice sites, the type of ion that can be conducted through the channel and the sensitivity of the receptor to activation or inactivation [15]. An even more dramatic example is the editing of serotonin receptors, which can produce as many as 24 variants [16]. Thus, a single gene locus can generate dozens of variants, each perhaps fine-tuned for a specific set of circumstances or conditions.

The prevalence of A-to-I editing in the CNS is especially evident in primate brains, where A-to-I editing is surprisingly widespread [17]. There is even some evidence that it is particularly dominant in human brains compared with other primates [18]. The reason for this seems to be the prevalence of A-to-I editing in Alu repeat regions. Alu repeats are found in the primate lineage, but over-represented in the human genome compared with other primate species [19]. This finding has fueled speculation that the complexity of the human brain may in part be a consequence of the extensive editing of neuronal transcripts [17].

2.2. Regulating Gene Expression

Generating transcriptomic diversity is certainly a compelling role for RNA editing, but if this were its sole purpose then we would predict that the majority of edit sites would be in the coding regions of transcripts. Unfortunately for this hypothesis, the best estimates suggest that two or three times as many edit sites exist in non-coding regions, such as introns and regulatory regions upstream of genes, than in the protein-coding regions. This seems to be particularly common for A-to-I edit sites [18, 20]. This would suggest that RNA editing must play other roles in the cell. The most ob-

vious explanation for the presence of edit sites in introns and upstream regulatory regions would be to postulate that these sites modulate the expression of the transcript in some fashion. Circumstantial evidence from a variety of systems supports this hypothesis. Transgenic mice lacking the enzymes required for A-to-I editing die *in utero* or shortly after birth. It appears that A-to-I editing is required for fetal development [21]. The disruption of A-to-I editing has also been associated with several neurological disorders in humans [7]. Perhaps the most striking finding to date is the apparent correlation between aberrant editing of serotonin receptors and major mental illness, especially suicidal depression (reviewed in [8]). These findings would argue that A-to-I RNA editing plays a key regulatory role.

The same may also be true for C-to-U editing. Overexpression of the enzyme complex associated with mammalian C-to-U editing causes tumor formation in many mammals [22, 23]. The tumors show aberrant editing of transcripts that do not appear to be edited in normal cells [23]. A more compelling instance was the discovery that aberrant C-to-U editing of the neurofibromin 1 (nf1) transcript is associated with neurofibromatosis type 1, a disease characterized by repeated tumor formation in a variety of tissues [24, 25]. More recently, evidence of aberrant A-to-I editing has also been documented in several human cancers [26]. These findings bolster the argument in favor of RNA editing as a means of gene regulation.

It is possible that RNA editing exerts a regulatory role as part of a larger system of monitoring and regulating mRNAs within the cell. Recent evidence suggests that there is crosstalk between the RNA editing system and RNA interference (RNAi). As discussed in Section 5, a key aspect of both A-to-I and C-to-U editing is the formation of double-stranded RNA (dsRNA) structures around the edit site. The presence of dsRNA is known to activate RNAi [27], so it is perhaps not surprising that the two processes interact. What is more interesting is the possibility that RNA editing might introduce or modify signals associated with RNAi. For example, a recent paper proposes a mechanism in which A-to-I RNA editing blocks the target recognition of micro RNAs (miRNAs) [28]. Other researchers have found that A-to-I editing may inhibit gene silencing via small interfering RNAs (siRNAs) (reviewed in [12]). Recent evidence suggests at least one homolog of the C-to-U editing enzyme in humans may play a similar role [29].

2.3. Cellular Defense

Another intriguing possibility is that RNA editing in some way protects genomes from retroviruses. A homolog of the enzyme responsible for C-to-U editing in humans appears to inhibit HIV (human immunodeficiency virus) replication by blocking a key step in the process of reverse transcription [30]. There are nearly a dozen homologs of the C-to-U editing enzyme, so it is possible that these enzymes are actually part of a novel defense mechanism against retroviral infection (reviewed in [31]). Related to this finding, researchers have speculated that A-to-I editing may have helped slow the spread of retrotransposons within the genome. The proliferation of retrotransposons can have disastrous consequences for a genome, but it is difficult for cells to combat these elements because they are embedded within

the genome. One possible solution may be to edit the transcripts that contain portions of these repeat regions. Transcripts that are overly edited are retained in the nucleus and targeted for degradation, so editing transcripts may be a means of preventing retrotransposon proliferation [3]. Indeed, in a recent study of repeat regions and A-to-I editing in mouse, the level and extent of editing seems to be correlated with the size, length and variety of repeat regions present in a genome [32].

It may also be the case that the editing activity occurs at the DNA level as well as at the RNA level. One homolog of the C-to-U editing enzyme in humans has the ability to edit DNA; it is employed in the differentiation of immune system cells in response to specific antigens [33, 34]. Thus, it is possible that some or all RNA editing enzymes have dual activity, editing either RNA or DNA. If this is the case, then the hypothesis that RNA editing serves a protective function against rogue elements in the genome is all the more plausible.

2.4. Error Correction

When RNA editing was first discovered, its role was postulated to be much simpler than the many nuanced activities that have since been discovered. Initially, it appeared that RNA editing was a form of error correction, selectively adjusting codon bias and open reading frames without making dramatic changes to the genomic version of the transcript. This still appears to be at least one key function of RNA editing, especially in instances where editing occurs in mitochondria or other organelles [35].

In the case of substitution RNA editing, the best examples of the error correcting activity come from studies of plant organellar RNA editing. Since the mitochondria and chloroplasts have their own genomes and are derived from prokaryotic ancestors, one might expect that the codon usage and bias of these genomes will be distinct from their hosts' preferences. We know that these organelles, with their prokaryotic-derived genomes, can sometimes have distinct codon bias and preferences from their host organisms. We might therefore predict that RNA editing would serve a role in adjusting codon usage to more closely match that of the host. If this is the case, then we would anticipate relatively few edit sites in non-coding regions, as these have little impact on translation. Indeed, nearly all the C-to-U edit sites in several plant mitochondrial genomes studied to date are in coding regions [36, 37, 38]. Furthermore, at least one study that compared genes edited in some plants with their unedited counterparts showed that over time, the genomically encoded C is replaced by genomic T (which would then generate the correct U in the resulting mRNA transcript) [39]. In other words, it appears that a selective pressure does exist to correct the genomic template, and that RNA editing is perhaps the interim solution to long-term encoding of the correct sequence within the genome.

In the best illustration of "fixing" an error in the genome via RNA editing, a start codon is generated for a chloroplast transcript that otherwise cannot be translated. Specifically an ACG codon is edited by C-to-U editing to create the requisite AUG [40]. More subtly, RNA editing in plant mitochondrial genomes seems to selectively alter codon usage and sometimes alters the amino acid encoded by the transcript

[36]. At least one computational analysis of edit sites from three plant mitochondrial genomes found evidence that edited codons tend to yield amino acids that may improve the resulting protein's stability [41]. In the few instances where an edit site occurs in a non-coding region, such as an intron, the C-to-U edit may yield a more stable secondary structure or otherwise facilitate downstream processing of the mRNA transcript [42]. Thus, at least in mitochondrial genomes, the role of RNA editing may be to assist in ensuring functional proteins are generated from transcripts that might otherwise be poorly expressed in the host environment.

3. WHY STUDY RNA EDITING COMPUTATIONALLY?

Given the myriad roles of RNA editing in cells, identifying the sites of editing and dissecting the mechanisms involved in editing will be critical in better defining cellular activities in both normal and disease states. Yet, until recently, most edit sites were discovered by serendipity. A case in point was the recent discovery that C-to-U editing of the interleukin-12 (IL-12) transcript is associated with a form of atopy, a disease characterized by hypersensitive allergic responses [43]. Initially, the observed C-U mismatch between expressed IL-12 from patients with the condition and the reference human genome were postulated to be the result of a single nucleotide polymorphism (SNP). Yet, when the region was sequenced from patients, the C in question proved to be genomically encoded. Further analysis then revealed that this was in fact a case of C-to-U editing. Such approaches to identifying edit sites, while critical for the study of RNA editing, are frustratingly slow because they depend entirely on chance.

More widespread experimental surveys of candidate RNA editing sites have emerged in recent years. For A-to-I editing, the presence of the unusual I in transcripts is a clear hallmark of that form of RNA editing. It is possible to experimentally screen transcripts for the presence of inosine and estimate its prevalence from a variety of tissues in *Caenorhabditis elegans* [44]. This strategy is more difficult to implement in the case of C-to-U editing since one cannot easily distinguish the edited U in a transcript from one that was genomically encoded. Nevertheless, a recent report details a method for surveying the instances of C-to-U editing in the organelles of plants [45].

These approaches are critical for creating a body of experimentally verified sites of A-to-I or C-to-U RNA editing. However, such approaches can be limited by a number of factors, not least the time and expense involved in such endeavors. In contrast, the wealth of genomic sequence data and associated data on expressed sequences from a variety of species is a resource that has been under-exploited until recently. A computational survey of a genome and transcripts has the potential to dramatically increase the number of candidate sites, and these can then be evaluated experimentally to verify the findings. In the past few years, at least three genome-wide surveys of the human genome and transcriptome have yielded dramatic results: many thousands of candidate A-to-I edit sites [18, 20, 46]. These results are described in more detail in Section 4.1.

Aside from increasing the number and variety of RNA editing sites, computational approaches offer a second pow-

erful advantage. With a suitable set of candidate edit sites, computational methods can help discern the signals that allow for the selection of a specific nucleotide to be edited. As described in Section 5, experimental analyses have provided critical insights into some of these signals. But the power of computational approaches lies in their ability to discern subtle patterns in sequences that correlate well with known edit sites. In this way, computational methods can further our understanding of the mechanisms that underlie the RNA editing process. As described in Section 6, a number of such approaches have been developed for plant organellar RNA editing. These approaches have both enhanced our ability to discover new sites of editing well as provided some insights into the signals and processes that regulate this phenomenon. When used in conjunction with experimental assays, we have a far better chance of discerning the underlying processes that drive RNA editing than with either a solely experimental or computational approach. Recent advances in experimental techniques and assays to identify RNA editing sites and the explosion of genomic data could not come at a better time for those interested in solving the mysteries of RNA editing. The time is now ripe for computational approaches to tackle the key questions of where editing occurs and what signals might help direct the editing machinery to sites of editing.

4. WHAT DO WE NEED FOR A COMPUTATIONAL APPROACH?

If computational approaches to identifying edit sites are desirable, then it is reasonable to ask what is needed to implement such approaches. The key requirement, clearly, is suitable data for training and testing methods. As with many areas of bioinformatics analysis, the need for data is generally far greater than the data available at hand. This is certainly the case for the RNA editing realm, although this has altered dramatically in the past five years.

4.1. Data

There are three forms of data on RNA editing that have accrued since the phenomenon was first discovered in higher eukaryotes two decades ago. The first is data garnered through experimental discovery and analysis. In the case of A-to-I editing, one means of identifying edit sites experimentally is to identify mRNA transcripts containing inosine, a nucleotide that is unlikely to occur in any other context. The results of such experimental surveys suggest that A-to-I editing is widespread in a variety of eukaryotes [44]. No such obvious mechanism exists for identifying C-to-U edited transcripts. Rather, each instance of C-to-U editing has been discovered by serendipity, and this may in part explain the very few such instances documented to date in humans [47].

In other systems, however, the discovery of C-to-U edit sites has been much more rapid. In plant mitochondrial genomes, C-to-U edit sites are relatively easy to identify because the majority are in coding regions and editing is required to yield the correct reading frame or protein sequence [2]. Thus, a plethora of experimentally verified edit sites exist in these genomes. While experimental approaches to identification of edit sites are clearly desirable and necessary, the slow pace and reliance on chance in some species including humans is less than satisfactory. Therefore, a second approach to generating the desired data on edit sites has been

to utilize comparative approaches based on phylogenetic comparisons. For example, in one study, researchers compared known A-to-I edit sites in *Drosophila* species to identify common features of edit sites. They then scanned the genome and identified other candidate edit sites. By this approach, they nearly doubled the number of known A-to-I edit sites in *D. melanogaster* and experimentally verified one of these sites [48]. The development of computational methods to predict candidate edit sites through the use of such approaches is described in more detail in Section 6.2.1.

A third approach, utilizing the wealth of sequence data available today, has been by far the most successful in terms of expanding the set of candidate edit sites. Several large-scale surveys of the human genome have identified many thousands of candidate A-to-I edit sites [18-20, 46]. In one of these studies, by Furey and colleagues, a general survey of the reference human genome was conducted to evaluate its overall sequence quality. Mismatches between transcript and genomic sequences were assessed to evaluate the relative completeness and accuracy of the reference human genome. A key feature of this study was the use of multiple transcript sequences so that sequencing errors in one transcript sequence would not overly influence the estimate of sequencing errors in the reference genome. For example, if a mismatch occurred between a genomic region and an EST sequence, the team sought at least two other ESTs generated independently that supported the first EST sequence. In this fashion, the team were able to identify sequencing errors in the reference genome with high confidence. In the process, however, they also identified numerous instances where three or more ESTs and other transcript sources supported a mismatch that could not be explained as a sequencing error. In these instances, the mismatches were compared to a panel of individuals to ascertain if these were single nucleotide polymorphisms (SNPs) as well as to the SNP database (dbSNP, [49]). After elimination of such variable sites, the team still had a sizable number of mismatches that were well-supported by transcript data. Of these, over 9000 appear to represent candidate instances of RNA editing [46]. Similar approaches are described in more detail in Section 6.2.2.

4.2. Defining the Problem

The combined efforts of experimental and computational analyses have generated a large set of candidate edit sites for A-to-I editing in humans [18-20, 46] as well as a sets of known C-to-U edit sites in plant organelles (see [36-38] among others). With this in hand, we can try to define the specific aspects of RNA editing that are amenable to computational analysis. It is clear that trans-acting factors, namely proteins associated with the editing process, are required for RNA editing. Identifying trans-factors requires experimental analysis to ascertain the precise contribution of a given protein to the RNA editing process. However, computational methods can assist in narrowing the field of candidates by highlighting homologs of known enzymes and factors, conducting comparative or phylogenetic analyses of candidate factors and in modeling the structures of such candidates [2, 10, 47, 50]

Computational analysis can play a larger role in the identification of cis factors, or sequence-based signals that help

drive the selection of individual edit sites. Indeed, it is in the identification of these cis-factors and in the utilization of these data that computational approaches have been able to make novel predictions. For example, in C-to-U editing in plant mitochondria, both experimental and computational methods have highlighted the importance of the nucleotides that flank the edit site [36, 41, 51, 52]. The apoB C-to-U edit site in humans also shows strong nucleotide biases in the regions immediately upstream and downstream of the edit site [53, 54]. Such cis-signals are described in more detail in Section 5.

5. MECHANISMS OF EDITING

Before we can predict sites likely to be edited by either A-to-I or C-to-U editing, we must first have some sense of the biological mechanism and signals associated with selection of a given nucleotide for editing. Extensive experimental analyses have identified numerous trans-acting factors, including the enzymes that catalyze the actual deamination reaction. For a detailed review of the biological mechanism and trans-factors that participate in editing processes, see [1, 2, 10] among others. Here, I will focus on what is known about cis-signals associated with editing because from a computational perspective the real focus has been on identifying these signals and utilizing them in predicting new sites of editing.

5.1. A-to-I Editing: Mechanism and Signals

A-to-I editing occurs as the result of deamination of an adenine to an inosine. This reaction is catalyzed by a class of enzymes known as the adenine deaminases that act on RNA (ADAR) [5]. In humans, there are three ADARs. ADAR1 and ADAR2 are expressed throughout the body, while ADAR3 appears to be exclusively expressed in the brain [1]. In order for editing to occur, the ADARs rely on the formation of a double-stranded RNA (dsRNA) structure formed by basepairing of the region flanking the edit site and a sequence downstream of the edit site known as the editing site complementary sequence (ECS). This is shown in Fig. (1). These complementary sequences can be quite distant, involving hundreds or even thousands of bases in the duplex RNA structure. A common motif seems to be the basepairing of an exonic region with an intron or a distal repeat unit such as an Alu region [18]. The dsRNA need not be composed of perfectly complementary sequences, so even loose complementarity that can yield the duplex structure is sufficient [20, 48]. The ADARs recognize the dsRNA structure and then selectively edit some adenines to inosines within the region. How individual adenines are selected for editing within this larger structure remains uncertain. Studies in *Drosophila* species using phylogenetic comparisons showed a strong conserved sequence region around edited adenines [48], but this conserved region does not appear to be part of a common consensus sequence that corresponds with every site of A-to-I editing.

To date, most computational approaches have therefore sought to use the possible presence of a dsRNA structure to anchor the search for A-to-I edit sites. In most cases, this is then supplemented with comparisons of sequence data for the region from expressed sequence tags (ESTs) and cDNAs

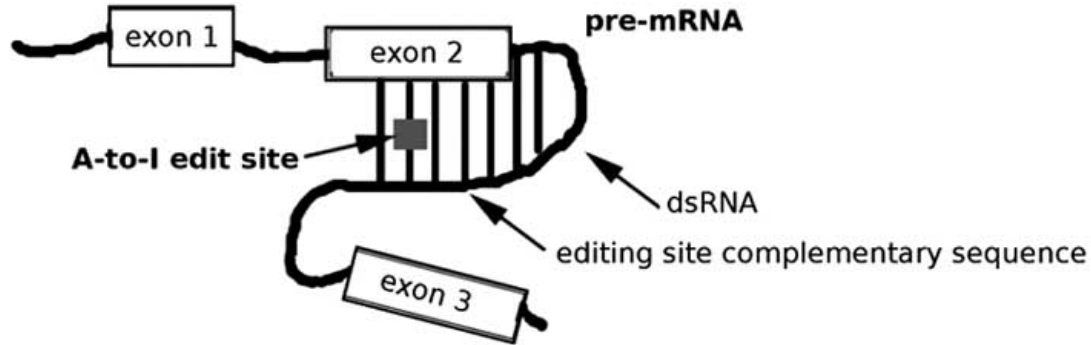


Fig. (1). A-to-I editing requires the formation of a dsRNA structure around the edit site.

against the genomic template. Mismatches between the two sources of sequence data are then candidates for further analysis and verification [18-20, 46]. These approaches are described in more detail in Section 6.2.2.

5.2. C-to-U Editing: Mechanism and Signals

C-to-U editing involves the deamination of cytidines to uridines, and occurs in both nuclear and organellar transcripts. In humans, a handful of C-to-U edit sites have been confirmed [6, 24, 43], and in plant organelles, a large collection of C-to-U edit sites have been identified in mitochondrial transcripts [36-38] and in chloroplasts [40, 55]. While all mammals seem to share a common mechanism for C-to-U editing, the mechanisms involved in plant organellar editing appear quite distinct. Therefore, in the ensuing discussion, I first focus on what is known about the mechanism of C-to-U editing in humans and other mammals. I then turn to the plant organellar editing process. The latter is quite interesting from a computational perspective as the abundance of known edit sites has spawned a variety of methods for predicting edit sites in these organellar genomes (see Section 6.1.2).

5.2.1. C-to-U Editing in Humans

The canonical example of C-to-U editing in humans is the editing of apolipoprotein B (apoB), and it remains the best studied instance of C-to-U editing in mammals. The editing of apoB was first described nearly two decades ago, when it was observed that an isoform of apoB in the small intestine was the result of a post-transcriptional modification to the transcript [6]. The editing of apoB involves the conversion of a CAA (encoding Glu) codon to UAA (encoding a stop codon), resulting in a truncated version of the protein. This form, known as apoB48, is roughly half the length of the full version of the protein. The latter is expressed, unedited, in the liver of humans. ApoB is expressed in both the small intestine and liver of mammals, but the location of editing varies among different species. Humans and some other mammals selectively edit apoB in the small intestine but not in the liver. In contrast, rodents such as mice and rats edit the apoB transcript in both the liver and the small intestine [56]. ApoB plays an important role in lipid transport and has been implicated in the disease process leading to atherosclerosis in humans. Intriguingly, in species that edit apoB in both the small intestine and liver, atherosclerosis is never seen [56].

Editing of apoB involves a complex of trans-acting factors termed the apoB editing complex (APOBEC). A number of homologs of this complex have been identified [47], so this complex is now termed APOBEC-1. The complex includes apobec-1, the cytidine deaminase that catalyzes the actual deamination reaction. However, it cannot edit the apoB transcript in the absence of a collection of complementary and stimulating factors [9]. Of these, APOBEC-1 complementary factor (ACF) appears to be involved in binding the mRNA substrate and positioning apobec-1 so that the deamination reaction can occur [57]. APOBEC-1 stimulating factor (ASF) seems to improve efficiency of editing [58] and other components of the complex may participate in edit site selection [11].

In addition to these trans-acting factors, evidence has accrued that sequence based signals help identify the exact cytidine that should be edited. The first signal associated with C-to-U editing of apoB was the identification of a mooring sequence located downstream of the edit site [59]. Additional mutational analyses of the region have revealed sequence-based signals. The first is a required upstream signal believed to influence the efficiency of editing [60], and a downstream AU-rich region between the edit site and the mooring sequence [53]. The introduction of these signals into other positions within the apoB transcript can induce editing *in vitro* [9, 60], suggesting that these sequence based signals are associated with editing of a given site. The known *cis*-signals for apoB editing are shown in Fig. (2). There are three other documented instances of C-to-U editing in humans, but all are the result of aberrant editing. Perhaps for this reason, none seem to share the conserved sequence signals associated with apoB editing [54].

At least two instances of aberrant C-to-U editing have been documented in human tumors: editing of a CGA to UGA in the neurofibromatosis type 1 (NF1) gene [24] and multiple editing of a transcript known as NAT1 (novel APOBEC-1 transcript 1) [62]. Both of these instances appear to contribute to tumorigenesis and are not observed in normal tissue. A more recent example of C-to-U editing in humans has been the discovery of an edit site in interleukin-12 (IL-12) transcripts. IL-12 is a potent immune system regulator, and editing of its transcript leads to atopy, a form of allergic hypersensitivity [43].

While only four instances of C-to-U editing have been identified through experimental analysis, there are over a

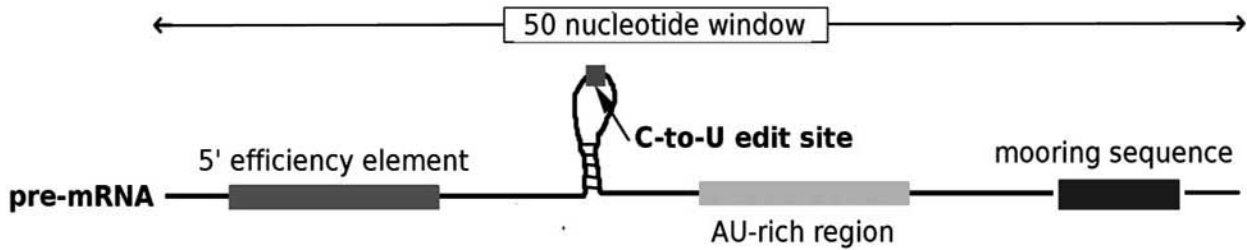


Fig. (2). The sequence-based signals known to influence C-to-U edit site selection in the apoB protein. Derived from [1, 54, 61].

dozen homologs of APOBEC-1 in humans, expressed in a variety of tissues [47]. Thus, the identification of new C-to-U edit sites seems to be only a matter of time, and an area where computational efforts may indeed have significant impact on our current knowledge and understanding of this process.

5.2.2. C-to-U Editing in Plant Organelles

While data for C-to-U editing in humans and other mammals have been relatively limited, a plethora of data exists from plant organellar systems. Many plant mitochondrial and chloroplast transcripts are C-to-U edited [2], and the data from several mitochondrial genomes has been used to develop computational methods for identifying putative sites of C-to-U editing (Section 6.1.2).

As with C-to-U editing in mammals, there appear to be requirements for sequence motifs upstream and downstream of edit sites in plant organellar transcripts. Flanking sequences seem to be located in a region extending approximately 30 nucleotides upstream of the edit site and 10 nucleotides downstream of the edit site [63, 64]. Interestingly, this window of roughly 40 nucleotides is similar to the 50 nucleotide critical window around the C-to-U edit site in apoB (Fig. (2)). Other similarities include the apparent requirement for certain nucleotides in close proximity to the edit site [65, 66], the possible presence of a secondary structure around the edit site [51] and a strong AU-bias in the editing region [64]. Furthermore, in a review of the experimentally identified edit sites in the mitochondrial genome of *Arabidopsis thaliana*, researchers noted that the nucleotide immediately upstream of the edited C was a pyrimidine (C or U in the mRNA) in 93% of cases. In addition, it was observed that the edited C was in the second codon position in over half the instances [36].

Given these experimental findings, it seems reasonable to expect that these flanking sequences will have some consistent pattern. Earlier experimental work by Choury and colleagues argued against a clear consensus signal, but did ob-

serve some preferences and biases in nucleotide composition [66]. When a clear consensus is not present in a sequence, it is sometimes possible nevertheless to use nucleotide bias as a statistical measure of the likelihood that a signal exists. Mulligan and colleagues used information theory approaches to computationally test the hypothesis that the regions flanking edit sites contain signals required for editing to occur at a given C [52]. They found evidence for preferred nucleotides in a region extending approximately 20 nucleotides upstream of the edit site and to the nucleotides in positions +1 and +2 downstream of the edit site. While individual nucleotides were not conserved, certain di- and tri-nucleotide combinations were over-represented in the region (Table 1). This was especially true for signals upstream of the edit site, although some preferences for the dinucleotide immediately downstream of the edit site were also observed [52].

Despite the similarities in cis-signals associated with C-to-U editing, plant organellar C-to-U editing may not share the same mechanism for C-to-U editing observed in mammalian nuclear transcripts. No homologs to APOBEC-1 or ACF have been identified in plants. Nine homologs of cytidine deaminases are encoded in the nuclear genome of the model plant organism *Arabidopsis thaliana*, but these do not seem to actually participate in editing of organellar transcripts. While conclusive proof is still pending, the general perception is that the nuclear encoded cytidine deaminases are not part of the C-to-U editing mechanism in plant organelles [2].

Rather, it appears that a different class of proteins, known as pentatricopeptide repeat proteins (PPR), appear to be critical for RNA editing to occur in the organelles [2]. Recent evidence suggests that PPR proteins are translocated into the organelles and at least one is absolutely required for editing in the chloroplast [67]. Initially, PPR proteins were believed to serve a function similar to ACF, namely, to help with identification of the C to be edited. More recent evidence suggests some of the PPR proteins may also contain a functional cytidine deaminase encoded in a unique catalytic domain [68]. If this is the case, then a PPR protein would be

Table 1. Sequence Based Signals for C-to-U Edit Sites in Plant Mitochondrial Genomes

Position	-17	-16	-15	-14	-12	-11	-10	-6	-5	-4	-2	-1	0	+1	+2
Dominant	U	A	C	A	A	U	G	C	C	A	U	U	C	G	G
Nucleotide		G				A		A	U	G	C	C			U
												A			

Table 1: Sequence-based signals that appear to be strongly correlated with C-to-U edit sites in plant mitochondria based on an information theoretic approach. The edited C is shown in bold at position 0. The most significant nucleotide in each position flanking C-to-U edit sites in two plant species (*Arabidopsis* and *Oryza*) is shown. Derived from [52].

able to both select sites for editing and carry out the requisite deamination reaction. Since PPR proteins are members of a large family of proteins and represent as much as 1% of the plant nuclear genome content, some have speculated that there may be a unique PPR protein to edit each individual edit site in the organellar genes [2, 68].

6. COMPUTATIONAL METHODS FOR IDENTIFYING EDIT SITES

The complexity of the signals involved in substitution RNA editing may in part reflect the myriad roles that this process plays within the cell. However, the complexity of the entire process need not deter efforts to uncover specific signals and aspects of the process in individual species. In the ensuing discussion, I review some of the computational approaches that have been developed to identify sites of A-to-I and C-to-U editing. Interestingly, some of these methods have also provided perspectives into the underlying mechanisms of the process. As with most other areas of computational analysis of biological phenomena, the primary goal of computational approaches in RNA editing has been to document candidate edit sites. Such wide-scale surveys at the genomic level can then provide a basis for further experimental investigation. However, an intriguing trend has developed in the RNA editing realm in which computational approaches have begun to highlight aspects of the signals associated with edit sites. These signals have, in turn, shed some light on possible mechanistic aspects of the RNA editing process. Thus, in this particular field, computational efforts are contributing not only to the knowledge base by predicting candidate edit sites but also proposing testable hypotheses of the signals involved in the editing process.

One of the earliest efforts to computationally predict sites of RNA editing was developed by Ralf Bundschuh for a form of RNA editing in which nucleotides are inserted or deleted from transcripts. This is in contrast to the substitution editing that is the focus of this review, but his work is nevertheless instructive. Bundschuh proposed a computational approach to identify sites of insertional and deletion editing in the mitochondrial transcripts of the slime mold, *Physarum polycephalum*. By comparing *Physarum* mitochondrial transcript sequences at the protein level to known, unedited homologs in other species, he was able to predict the sites where nucleotides had to be inserted or deleted in order to generate the correct protein sequence [69]. This approach proved to be very successful and has spawned numerous similar efforts in other species including in plant mitochondrial genomes [70]. I refer to such methods as “comparative approaches” in the ensuing discussion since they rely on sequence homology to guide the prediction of edit sites.

Others have tried to discern the features associated with edit sites through statistical or machine learning approaches. Such methods have been developed both for A-to-I and C-to-U substitution editing, although much of the work has been developed in plant mitochondrial genomes rather than in mammalian systems. In these approaches, which I term “feature-based” methods, there are two phases of analysis. In the first or training phase, known edit sites are used to derive features that correlate well with editing. In the testing phase, the method is evaluated on other data to assess performance. Needless to say, for these methods to work, sufficient data is

required to generate a training dataset and an independent test dataset. The training-testing procedures often require large amounts of data of high quality; in this particular context, it is desirable to have experimentally verified data to use for the training phase.

The data on human A-to-I and C-to-U edit sites is relatively limited (see Section 7), but there appears to be a wealth of data on C-to-U edit sites from plant mitochondrial genomes (see [36, 37, 38] among others). Thus, the majority of feature-based methods have been developed for plant mitochondrial genomes. Each method has incrementally improved on the predictive ability of its predecessors. The most recent methods are surprisingly accurate and offer some insights into the mechanisms that may underlie RNA edit site selection in plant mitochondrial transcripts [71, 72]. These advances are of particular value to those in the plant community who study RNA editing in organellar systems. Yet, computational biologists are generally keen to extend their approaches to as broad a base as possible. It remains to be seen if any of these methods or the insights they have stimulated will carry over into the work on predicting sites of A-to-I or C-to-U editing in humans. In Section 7.3, I explore the possibility that these approaches might one day contribute to a generalized computational model of substitution RNA editing.

6.1. Predicting C-to-U Edit Sites in Plant Mitochondrial Genomes

The presence of a dataset as well documented as that for C-to-U editing in plant mitochondrial genomes has spawned a number of computational efforts in the past five years [41, 51, 70-72]. The majority of these approaches are *ab initio*, although at least two methods utilize comparative approaches. In fact, the methods with the highest accuracy tend to rely on comparative approaches [70, 72]. This is because comparison to the well studied mitochondrial protein sequences is the simplest means of identifying likely edit sites. However, comparative approaches are stymied in that they are less likely to predict edit sites that result in synonymous substitutions (where the encoded amino acid does not change) or editing that occurs in non-coding regions such as introns.

Feature-based methods have the advantage that they can potentially identify edit sites regardless of location or whether they cause a synonymous or nonsynonymous substitution. However, these methods in general have lower accuracy than comparative approaches, perhaps because not all of the signals associated with edit site selection are captured by these models. As discussed in Section 5, efforts to discern the *cis*-signals associated with edit site selection have yielded some insights. However, it is not clear if other, subtle signals exist in the sequence or its secondary structure that might drive edit site selection. Further experimental work will be needed to discover such features if they exist. The ensuing sections describe one comparative method, several *ab initio* methods and one method that combines both approaches to predicting C-to-U editing in plant mitochondrial genomes. It is possible to compare the methods developed to date for predicting C-to-U edit sites in plant mitochondrial genomes because of the use of a common dataset. This is described in more detail in Section 6.1.4.

6.1.1. Comparative Approaches

Comparative approaches rely on the fact that changes to the transcript will be reflected in altered protein sequences. Therefore, comparing the known amino acid sequence with what is encoded in the genome will highlight discrepancies that are likely to be sites of RNA editing. This approach works especially well for instances of RNA editing that dramatically alter the protein sequence from what would have been encoded based on the genomic transcript. In the specific case of plant mitochondrial transcripts that undergo C-to-U editing, most edit sites result in a change in the encoded amino acid [36]. Thus, any mismatches between the genomically encoded version of the protein and the known protein sequence can be highlighted. The transcript nucleotide sequence can then be examined to determine if a C-to-U or other substitution editing process might yield the functional version of the protein sequence.

This approach was applied with great success in plant mitochondrial genomes as part of a program known as PREP-Mt [70]. PREP-Mt takes two inputs: the identity of the mitochondrial protein believed to be encoded by a given DNA sequence and the genomic DNA sequence for the predicted coding region. The sequence is then translated and aligned at the protein level with a database of aligned sequences of the relevant mitochondrial protein from a variety of species. Any mismatches between the translated sequence and the aligned sequences are evaluated to discern whether a C-to-U edit would correct the mismatch. In other words, if a C-to-U edit of the codon would generate the amino acid observed in the aligned sequences, then the site is marked as a likely edit site. A scoring scheme and an associated system of resolving ambiguous sites enable automated assessment of all candidate mismatches. The algorithm preferentially minimizes the number of edit sites required to bring the translated protein sequence into alignment with its counterparts in the aligned sequence database. This ensures a more parsimonious approach, in that an amino acid mismatch is only associated with RNA editing if other explanations, such as mutations and speciation, cannot account for the observed differences [70].

PREP-Mt was reported to have an accuracy of between 97% and 99% for a variety of mitochondrial genomes [70], although others have disputed these claims (see Section 6.1.4 and [41, 71]). Even with modifications to the calculation of performance metrics such as accuracy, PREP-Mt has an accuracy of 84% or higher [41]. Indeed, the key advantage of comparative methods such as PREP-Mt is the ability to yield highly accurate predictions of edit sites. The potential shortcoming, of course, is that such methods can only be applied when there is extensive knowledge of the proteins encoded by the genomes of interest. A second potential concern with this specific approach is that it can only identify edit sites which result in an amino acid change. For those edit sites that result in synonymous substitutions and those in non-coding regions, such an approach is less effective. Nevertheless, comparative approaches offer the highest accuracy for those instances where prior knowledge of protein sequence or homology from other species exists.

6.1.2. Feature-Based Approaches

To address some of the weaknesses of the comparative approach, a number of feature-based methods have been developed to identify sites of C-to-U editing in plant mitochondria. As described in Section 5.2.2, the original experimental description of editing sites in *Arabidopsis* noted that the majority of edited Cs were preceded by a pyrimidine (C or U in the mRNA) and that the edited C was in the second position of the codon in roughly 50% of the instances. All but 15 of the 441 instances of editing were within coding regions [36]. In addition, it has been proposed that RNA secondary structure, perhaps the formation of a stem-loop of the region flanking the edit site, plays a role in selection of edit sites [51]. These features all have been utilized by various feature-based approaches to develop methods that can predict the sites where editing is likely to occur in a given gene or genomic sequence. The first of these, proposed by Cummings and Myers, utilized relatively straightforward statistical measures to identify features associated with editing. This group focused on a 41 nucleotide window surrounding each known edit site. They randomly selected a set of unedited Cs and the surrounding sequences to serve as a null set (true negatives). Nucleotide composition in these sequence windows as well as free energy calculations for the RNA secondary structure of the sequence window were used in tree-based statistical methods to develop a predictive algorithm. The overall accuracy of predictions ranged from 71% to 74% depending on the statistical method used. Combining data from multiple genomes yielded accuracy rates as high as 85% [51].

There were two interesting findings of this work that may provide insights into the mechanisms that underlie edit site selection. First, the authors were able to confirm the earlier observation that the nucleotide immediately upstream of the edit site is a pyrimidine in most cases, and that the nucleotide in this position is of critical importance in ensuring accurate prediction of the edit site [51]. The second and perhaps more striking finding was that the use of folding free energy as a measure of possible RNA secondary structures in the region dramatically boosted predictive ability (from 70.5% to 84.8%). Thus, the researchers were able to highlight the possible presence of secondary structures as part of the signal associated with edit site recognition. This finding stood in sharp contrast to earlier experimental work that suggested secondary structure did not play a role in edit site recognition [73]. However, more recent experimental efforts to identify consensus sequences around known edit sites have highlighted the lack of strong consensus signals while demonstrating the importance of overall nucleotide composition in the region [66]. This may be indirect evidence that secondary structure does play a role in edit site recognition. Many conserved RNA secondary structures have low consensus at the sequence level, but do show strong nucleotide biases and organization [74]. Thus, the presence of nucleotide bias in the region flanking the edit site may be subtle evidence of selective pressure to maintain a secondary structure associated with edit site recognition by the RNA editing machinery.

In any case, the key advantage of feature-based methods is precisely that they can provide insights into those features that drive edit site selection and thereby highlight aspects of

the mechanisms that underlie the process of C-to-U editing. In this sense, computational approaches that take an *ab initio* approach have the ability to both improve our ability to identify edit sites as well as further our understanding of the process itself. Of course, the lower accuracy level of such approaches limits their immediate use in reliably predicting sites in new genomes or sequences. However, in the past few years, a number of other feature-based methods have been proposed, and the latest ones are quite competitive with comparative approaches in terms of accuracy and other measures of performance.

The first of these feature-based methods to improve on the Cummings and Myers approach was proposed by Thompson and Gopal. We used a machine learning approach known as a genetic algorithm (GA) to evaluate several features associated with edit site recognition. The goal was to ascertain which of these features were most important for accurate edit site prediction. By extension, we assumed that such features might also be of importance for *in vivo* edit site selection. As described in [41], we selected six features to study based on the data available from several mitochondrial genomes. These included assessing the composition of the nucleotide immediately upstream of the edited C (in the -1 position) as well as the nucleotide immediately downstream (in the +1 position). We also evaluated the tendency of certain codons and their encoded amino acids to be preferentially edited to ascertain whether editing in plant mitochondria is driven toward modifying the amino acid content or juxtaposition within the resulting proteins. We randomly selected a similar sized dataset of known, unedited Cs to serve as our null set in training and testing of the genetic algorithm.

These features were used as variables to the genetic algorithm, which attempted to optimize weights based on classification accuracy. The results, after correction for an error in data processing, were similar to those of Cummings and Myers. Accuracy ranged from 77% to 86%, with the latter being calculated for those predictions that had high confidence. As with the Cummings and Myers work, we were able to discern features that have a strong correlation with edit site selection. One of the key findings was that editing seems to preferentially target codons that will yield more hydrophobic amino acids after editing. We speculate that this may be a means of improving protein stability, since large numbers of hydrophobic residues in the cores of some proteins improve their stability [75]. In addition, we confirmed the earlier findings that the nucleotide composition immediately upstream and downstream of the edit site is a key component in accurate edit site prediction [41].

A second machine learning approach was recently applied to the same dataset by Du and colleagues. They used a support vector machine (SVM) combined with a measure of the trinucleotide composition of the region immediately upstream (-3 to -1) and downstream (+1 to +3) to predict sites of editing. The result is a feature-based method with accuracy levels of 83% to 85%, comparable to PREP-Mt, the comparative approach described in Section 6.1.1. The advantage of this approach over PREP-Mt is that it is not limited to nonsynonymous edits of coding regions; it can predict edit sites regardless of whether they occur in coding or noncod-

ing regions, as well as those that cause synonymous and nonsynonymous changes to the protein sequence. Furthermore, the authors were able to demonstrate that the triplets flanking the edit site are strongly biased and can be used to improve predictive accuracy (from 82% to 85%) [71]. This supports other computational and experimental analyses that found that certain di- and tri-nucleotides are preferred in the regions flanking edit sites [52, 66]. Thus, as with earlier efforts, this approach has both improved predictive ability and furthered our understanding of the signals that are associated with edit site selection in plant mitochondrial C-to-U editing.

6.1.3. Methods Combining Both Approaches

While recent advances in feature-based approaches have brought performance to roughly the level of comparative approaches, it is still the case that comparative approaches tend to have higher predictive accuracy. Thus, a very recent approach has tried to blend the best of both worlds and create a hybrid method that draws on both protein and nucleotide sequence homology with feature-based selection to improve on edit site prediction [72]. As with PREP-Mt (Section 6.1.1), the Cytidine-to-Uridine Recognizing Editor (CURE) starts with an alignment of the input sequence to known homologs of the encoded protein. Identified mismatches are then evaluated to assess whether they have features of known edit sites, such as nucleotide triplet bias in the flanking sequence [71]. The combined approach has a reported accuracy of 98% for a variety of mitochondrial genomes [72], although as discussed in Section 6.1.4 below, the actual accuracy is closer to 85%.

6.1.4. Comparing Results

With the plant mitochondrial edit site data, there is a nearly unique opportunity in the field to compare the results of multiple computational approaches. While each method uses a different subset of the known mitochondrial genomes and edit sites, cross comparison is nevertheless possible. For example, all the methods to date report performance on *Arabidopsis thaliana*. As the model organism for studies in plants, *Arabidopsis* provides a particularly important yardstick in evaluating the relative merits and potential weaknesses of each method. In addition, as one of the first mitochondrial genomes to have been thoroughly evaluated for RNA editing [36], it remains the gold standard for C-to-U editing in plant mitochondria. As such, *Arabidopsis* is an excellent test subject for the methods described here. Table 2 shows the comparative performance of each method described in Sections 6.1.1 and 6.1.2 on the *Arabidopsis* mitochondrial genome. There are 441 documented C-to-U edit sites in *Arabidopsis* [36], although each method used a subset of these edit sites and at least two methods include edit sites not originally reported [71, 72]. For each method, the number of cytidines that were edited constitutes the set of as true positives (TP), and the number of cytidines known to be unedited represent true negatives (TN). False positives (FP) are known unedited Cs that are predicted to be edited while false negatives (FN) are known edited Cs that are predicted to be unedited.

Using these values, as reported by each group, we can then assess performance using standard summary statistics. There are three summary measures of performance: accuracy, the most commonly used metric, sensitivity and speci-

Table 2. Performance of Methods to Predict C-to-U Edit Sites in Plant Mitochondrial Genomes

Method	Accuracy	Sensitivity	Specificity	CC
Tree based statistics ^a	0.74	0.70	0.81	0.51
Support Vector Machine (SVM) ^b	0.85	0.86	0.85	0.71
REGAL ^c	0.88	0.91	0.85	0.76
PREP-Mt ^d	0.83	0.79	0.86	0.81
CURE ^e	0.85	0.86	0.84	0.85

Table 2: Comparison of performance by various computational approaches on the known C-to-U edit sites in the *Arabidopsis thaliana* mitochondrial genome. ^a Reported performance is for the best tree-based method random forest models [51]. ^b Support vector machine approach as reported in [71]. ^c Reported performance measures for REGAL are based on predictions with a confidence value of 90% or higher. See [41] for details. ^d Specificity was recalculated as PPV for these data because the datasets were biased with many times more true negatives than true positives [70]. ^e Specificity was recalculated as PPV for these data because of the use of a highly biased dataset [72].

ficity. Sensitivity is the proportion of true positives (known edited Cs in this instance) that are correctly predicted to be edited. Sensitivity is calculated as shown in Equation 1.

$$S_n = TP / (TP + FN) \quad (1)$$

where TP are the true positives (known edit sites predicted to be edited) and FN are false negatives (known edit sites predicted to be unedited)

Specificity is the proportion of true negatives (known unedited Cs) that are correctly predicted to be unedited. The calculation for specificity is shown in Equation 2.

$$S_p = TN / (TN + FP) \quad (2)$$

where TN are the true negatives (known unedited Cs predicted to unedited) and FP are false positives (known unedited Cs predicted to be edited).

Accuracy is the mean of the two measures as shown in Equation 3.

$$Acc = (TP + TN) / (TP + TN + FP + FN) \quad (3)$$

These measures produce reliable estimates of the performance of a method under certain standard assumptions. Key among these is the assumption that the proportion of true positives is roughly equal to the proportion of true negatives in the dataset [76]. When there are many more of one or the other category, these measures tend to be skewed to artificially high values that are not representative of true performance. Specifically, when there are many more known negatives (unedited Cs in this instance) than known positives (edited Cs), specificity will be high regardless of how well the classifier actually performs. This is because the very large number of negatives assessed ensures that even random classification will yield a large number of correct assignments. Since accuracy is the mean of sensitivity and specificity, it will be skewed in favor of a high specificity value. As a result, the overall performance of the classifier will seem deceptively high. The high values, however, are misleading since they do not inform us of actual performance but reflect instead a bias in the composition of the dataset.

To address situations in which a biased dataset exists, as is the case in the C-to-U edit sites in plant mitochondrial genomes, there are three solutions. The first is to selectively assemble a dataset so that there are an equal number (or nearly equal number) of true positives and true negatives.

This approach was used by several of the methods described above [41, 51, 71]. Rather than consider every unedited C as a true negative, these approaches each randomly selected a subset of known, unedited Cs as members of the true negative pool. Two of the approaches further constrained the set of unedited Cs to better match the edited Cs in terms of specific features or properties of the flanking sequence [51, 41] although this was more a consequence of the features selected for analysis rather than a requirement to address dataset bias.

A second approach to address concerns of a biased dataset is to use different metrics to assess performance. For example, a more representative performance measure is the Matthew's correlation coefficient (Equation 4). This measure is less susceptible to skew and is the generally accepted metric for comparing multiple methods' performance [77].

$$CC = \frac{(TP * TN) - (FP * FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \quad (4)$$

A third remedy to concerns of biased dataset composition is to eschew the use of sensitivity and specificity, which are more likely to yield erroneous values, in favor of two other metrics. If there are many more true positives than true negatives, one can use negative predictive value (NPV) as a measure of performance in the place of sensitivity. This is less often the case than the alternative: many more true negatives than true positives, as is the situation here. Under these conditions, it is generally advisable to use positive predictive value (PPV, Equation 5) in place of specificity [77, 78].

$$PPV = TP / (TP + FP) \quad (5)$$

In Table (2), the performance of each method is shown based on the reported numbers of true positives, true negatives, false positives and false negatives for each method. As described in the above equations, sensitivity, specificity and accuracy are calculated based on the reported values. However, in instances where a biased dataset was used, PPV is used in place of specificity. In addition, the correlation coefficient is listed for each method. Readers are encouraged to focus on the correlation coefficient (CC) over other metrics in comparing the methods and their performance.

6.2. Predicting Sites of A-to-I Editing

The progress that has been made in predicting sites of C-to-U editing in plant mitochondria was predicated on the availability of a large dataset of known edit sites. Progress in

predicting A-to-I edit sites has been hampered by precisely this limitation; until recently, very few experimentally validated edit sites were available for computational analysis. A second challenge of predicting sites of A-to-I editing is that the vast majority of known edit sites are in non-coding regions, such as upstream of genes or in introns [18]. This hampers the ability to computationally identify edit sites based on comparative approaches, as has been used with great success in plant C-to-U edit site prediction [70, 72]. Nevertheless, the challenges of predicting A-to-I edit sites have also been the motivation for developing computational methods to both discover new instances of A-to-I editing as well as extract features associated with edit site recognition and selection [3].

As with the case of computational methods to predict C-to-U editing in plant mitochondria, the current methods fall broadly into "comparative" and "feature-based" categories. However, all the methods share a common first phase. Given the paucity of experimentally verified edit sites, all of the methods described below begin with a genome-wide survey of candidate edit sites [18, 19, 20, 48, 79]. All mismatches in which a genomic A is replaced by a transcript G (I is read as G by cellular and sequencing machinery) are highlighted for further analysis. It is in the downstream processing of these mismatches that the approaches differ.

6.2.1. Comparative Approaches

The chief premise of the comparative approaches for the identification of A-to-I edit sites has been that edit sites will be evolutionarily conserved. That is, if an edit site is required for correct functioning of a protein, then one would expect closely related species to edit this site consistently. This appears to be the case for at least a subset of A-to-I edit sites in neuronal transcripts. In one study, for example, the sequence regions around an A-to-I edit site in a Na⁺ channel in *Drosophila melanogaster* were compared across several *Drosophila* species. In addition to identifying conserved sequence elements in a downstream intron that participate in the dsRNA structure required for A-to-I editing (see Section 5.1), Hoopengardner and colleagues also found conserved signals flanking the edited A. The authors then used this signature sequence to search for new targets, eventually finding and verifying 16 new A-to-I edit sites in *Drosophila* transcripts [48].

A second comparative approach used comparisons of putative edit sites between the mouse and human genomes to bolster confidence that a given site was A-to-I edited. The methodology involved aligning mouse ESTs to the mouse genome in the first phase of data generation. Each A-G mismatch between the genome and ESTs was highlighted for further analysis. For each of these putative edit sites, the same region of the human homolog was analyzed. For example, if mouse ESTs showed a G where the genomic sequence showed an A, the sequence was compared to the human homolog at both the genomic and transcript level. If human ESTs showed the same mismatch with the reference human genome, then the authors retained the site as a likely A-to-I edit site. In other words, mismatches that showed conservation between the two species were considered to be the most reliable indicators of A-to-I editing at that position. To reduce the complexity of the analysis, the authors focused

on those edit sites that result in a nonsynonymous change, what they term "re-coding" edit sites. By focusing on this subset, they could then use the known protein sequence to further verify the candidate edit site [79].

In the second phase of this analysis, the authors then used the highest confidence candidate edit sites to extract features of the sequence regions surrounding these putative edit sites. They used a log-odds score (LOD) to quantitatively assess the sequence composition of these edit sites against other A-G mismatches from the genome-transcript analysis. The group were able to discover a novel A-to-I edit site that is conserved in mouse and human, and they were then able to verify this edit site experimentally [79].

6.2.2. Feature-Based Approaches

In contrast to the feature-based approaches described for C-to-U editing in plant mitochondria (Section 6.1.2), the methods that determine A-to-I edit sites based on features in the sequence must first generate the dataset from which to discern these features. All of the methods described below begin with a comparison of the reference human genome to collections of transcripts from ESTs, cDNAs and many other resources. Having identified potential edit sites as mismatches between the transcript sequence and its genomic counterpart, each method then utilizes a series of features and criteria to evaluate the likelihood that the mismatch is a consequence of RNA editing [3].

The first of these approaches was published in 2004 by Levanon and colleagues. Using ESTs and cDNAs mapped to the reference human genome, the authors selected only those A-G mismatches that were encoded within regions likely to form dsRNA. By using this preliminary criterion, they eliminated many mutations and single nucleotide polymorphisms that would otherwise obscure and distract from the actual A-to-I edit sites. Several other criteria, such as the presence of clusters of mismatches, were used in "cleaning" the set of mismatches to identify putative edit sites. The final result was the identification of over 12,000 candidate A-to-I edit sites. A small subset of 30 sites were selected for experimental validation, and 26 have strong evidence of A-to-I editing [20].

This approach pioneered the use of the dsRNA as the primary feature for computational prediction of RNA editing sites. Subsequent methods have been essentially refinements on this basic procedure [18, 19]. For example, a different group utilized the same basic procedure, with the added constraint that they only considered A-G mismatches that occurred in clusters. It is known that some A-to-I edit sites tend to cluster in certain genes and specifically in certain regions of genes such as the untranslated regions (UTRs) and in introns. Therefore, it is reasonable to focus on those mismatches that show these distributions both in juxtaposition and in location. By using these additional criteria, this approach identified over 14,000 A-to-I edit sites. In addition, the group noted the overwhelming majority of such edit sites were either in or closely juxtaposed with Alu repeat regions. They therefore proposed that RNA editing may play a role in modulating the behavior of these repeat regions [18]. The association of Alu repeats and A-to-I editing has been confirmed by several other studies [17, 19], leading to specula-

tions about the role of RNA editing in regulating these repeat regions (see Section 2.3).

7. OPEN QUESTIONS AND ISSUES

The dramatic rise in computational methods for predicting sites of RNA editing and the subsequent verification of those predictions in the last few years is evidence that such approaches are meaningful contributors to our understanding of the prevalence and nature of substitution RNA editing. Yet, in many ways, such efforts are only in their infancy. There are essentially three areas where concerted, collaborative effort will be required to help further uncover the key players on the RNA editing stage. First is the need for high quality, experimentally verified data. Second, the need to identify all the contributors to the RNA editing process, from trans-acting protein factors to cis-acting sequence signals, is critical to building meaningful computational models of the process. Finally, one would like to be able to build computational models that are extensible and generalizable. While this may yet be a pipe dream, in the last section of this review, I explore how one might go about discovering the commonalities inherent in A-to-I and C-to-U editing as we currently understand these processes. The hope is that one day we will build a computational model that can accurately and effectively identify any site of substitution RNA editing. That goal, if achievable at all, will only occur as a result of close collaboration between experimental and computational scientists. Thus, the key challenge in extending our understanding of RNA editing is in the effective collaboration of scientists from a variety of backgrounds.

7.1. The Need for Data

The first concern in the efforts to understand RNA editing and the signals associated with this process is the lack of experimentally verified A-to-I or C-to-U edit sites. The number of experimentally verified edit sites for either A-to-I or C-to-U edits in higher eukaryotes lags behind computational predictions of candidate sites (see Section 6 for more on this). The best datasets of putative edit sites in humans are for A-to-I editing [18, 20], but only a handful have been experimentally validated [19]. Thus there is a need for high throughput experimental methods that can more rapidly assess the many candidate edit sites that have been identified computationally so far. Without a larger set of validated edit sites, it will be difficult to extend the current computational efforts or to discern the signals that may drive edit site selection in A-to-I editing.

The situation is all the more dire for instances of C-to-U editing in mammals. Only four instances of C-to-U editing in humans or other mammals have been confirmed experimentally [47, 43]. This has essentially stalled efforts to develop computational feature-based approaches. Comparative approaches might come to the rescue in this instance, were it not for the paucity of known edit sites in other mammals as well. Yet the tantalizing fact remains that there are a dozen homologs of the apobec-1 enzyme that edits the canonical C-to-U edit site in apoB. Some of these homologs may have roles in editing DNA rather than RNA [33, 31], but the number of homologs begs the question of whether there are also many more C-to-U editing targets that have yet to be discovered. The greatest challenge that lies ahead, for both computational and experimental biologists, will be in developing

the methods needed to identify and dissect C-to-U edit sites in mammals. Without more knowledge of such edit sites *in vivo*, it will be difficult to further our understanding of the signals that underlie this process.

In contrast to the limited set of known C-to-U edit sites in human, the plant mitochondrial and chloroplast genomes abound in such edit sites. Correspondingly, the number and variety of computational approaches used to analyze these data are impressive. From machine learning to parametric analyses to comparative methods (Section 6.1), these approaches have both refined our ability to identify C-to-U edit sites in these genomes as well as provided important insights into the signals associated with the editing process. The successes of computational approaches for these genomes stands as proof positive of what could be done if more data were available for edit sites in humans and other mammals. It is hoped that it will only be a matter of time before such data are made available for mammalian genomes so that the computational work done to date can be extended and deepened significantly.

7.2. The Need for Biological Understanding

While additional data on edit sites is crucial for the development of methods to predict novel sites of RNA editing, such approaches will always lag behind unless we can further our understanding of the myriad factors that influence edit site selection and participate in the actual editing of the transcript. A case in point is the development of methods to identify sites of A-to-I editing in the human brain. Earlier work in dissecting the key trans-acting factors laid the groundwork for the computational approaches that followed. Specifically, the experimental work of identifying the ADAR enzymes [10] and their reliance on dsRNA as the key signal for edit site selection [5], allowed computational biologists to develop algorithms that utilize dsRNA as one of the key features in predicting likely sites of A-to-I editing [18, 19, 20]. By the same token, the lack of knowledge on other signals that might contribute to the selection of the individual adenine that is edited has limited the ability of these methods to use a solely features-based approach. Rather, the current methods all localize their searches for putative edit sites to those regions where a mismatch occurs between a transcript and the reference genome. In other words, given further experimental analyses into the cis-acting factors that enable selection of individual adenines, computational approaches might one day be able to detect A-to-I edit sites even when transcript-genome mismatches are not available.

Similarly, a wealth of experimental knowledge has accrued on the C-to-U editing of apoB, yet this has not yet translated into a computational method that can survey a genome and report sites of likely C-to-U editing. Indeed, this is still the area in which computational methods are weakest. Efforts to develop methods based on features of the apoB edit sites and the few other known instances of C-to-U editing have yet to yield any significant metric or approach to wide-scale identification of C-to-U edit sites. Furthermore, the intriguing evidence that some homologs of the APOBEC-1 complex edit DNA [33] or as serve as inhibitors of retroviral infection [30], offers an opportunity to explore what may be a novel mechanism for cellular defense. In this context, it is all the more pressing that experimental and

computational scientists combine their efforts to extend our knowledge of C-to-U editing in mammals. Progress in this and other aspects of substitution RNA editing will be critical if we are one day to create meaningful and accurate computational models of RNA editing.

7.3. Long Term Aspirations

From a computational perspective, one of the most desirable aspects of an algorithm is the ability to generalize the approach to many different variations of the same fundamental problem. The urge to generalize in biology can lead to disastrous results, yet it is nevertheless intriguing to consider whether the different forms of substitution RNA editing might one day be identified by a common computational approach. At this point, such an exercise is purely speculative. Yet, such high level investigations can nevertheless provide some intriguing avenues for future exploration.

The mechanisms and trans-acting factors in different forms of RNA editing are quite distinct, yet it may be possible to glean some common themes in the cis-signals associated with various edit sites. For example, a key feature of both forms of deamination editing seems to be double-stranded RNA structures around the edit site. In A-to-I editing, as discussed in Section 5.1, the dsRNA structure is recognized by the ADAR enzymes as part of the process of selecting the individual adenine to edit. Similarly, the apoB transcript region around the edited C is predicted to form a stem loop structure, suggesting that dsRNA may play a role in selecting the edit site here too [50]. Furthermore, at least one computational analysis of C-to-U edit sites in plant mitochondrial genomes found strong evidence that secondary structures in RNA play a role in edit site selection [51]. It should be noted that in both the case of apoB and in the plant mitochondrial sites, the region of dsRNA is quite small averaging about fifty nucleotides compared with A-to-I editing where the dsRNA spans many hundreds of nucleotides. Nevertheless, the formation of a dsRNA element, regardless of overall length, may be one common theme for all nucleotides sites undergoing deamination editing.

A second common theme is the preference for AU-rich sequences in the regions flanking the edit site. This is seen in C-to-U edit sites in both apoB and plant mitochondrial sites [53, 64, 66], and a looser form of this preference may exist for A-to-I edit sites as well [80]. In fact, this preference for AU-rich regions may extend to other forms of RNA editing that do not involve deamination reactions [35].

While it might seem that these are at best tenuous commonalities, it is possible that the themes described here are the result of a shared biological process. For example, there is some evidence that a single enzyme can catalyze both C-to-U and A-to-I deaminations, leading to speculation that an ancestral editing enzyme might have encompassed both forms of editing [81]. Indeed, there is some evidence that C-to-U and A-to-I editing might occur on the same transcript. Such an instance was recently documented in a tRNA from the trypanosomatids, where RNA editing was originally discovered [82]. Thus, it may be possible at some future point to consider developing a computational method that can identify sites of both C-to-U and A-to-I editing based on common features such as those described here. In the meantime, close collaborations between computational and ex-

perimental biologists offer the best chance to uncover the fundamental principles that drive the intriguing process of substitution RNA editing in eukaryotic cells.

8. ACKNOWLEDGEMENTS

I would like to thank Eric Foster, Saria Awadalla and Gary Skuse for meaningful conversations in the past on aspects of RNA editing.

REFERENCES

- [1] Keegan LM, Gallo A, O'Connell MA. The many roles of an RNA editor. *Nat Rev Genet* **2001**; 2: 869-78.
- [2] Takenaka M, Verbitskiy D, van der Merwe JA, Zehrmann A, Brennicke A. The process of RNA editing in plant mitochondria. *Mitochondrion* **2008**; 8: 35-46.
- [3] Levanon EY, Eisenberg E. Algorithmic approaches for identification of RNA editing sites. *Brief Funct Genomic Proteomic* **2005**; 5: 43-45.
- [4] Stuart K, Allen TE, Heidmann S, Siewert SD. RNA editing in kinetoplastid protozoa. *Microbiol Mol Biol Rev* **1997**; 61: 105-20.
- [5] Polson AG, Crain PF, Pomerantz SC, McCloskey JA, Bass BL. The mechanism of adenosine to inosine conversion by the double-stranded RNA unwinding/modifying activity: a high-performance liquid chromatography-mass spectrometry analysis. *Biochemistry* **1991**; 30: 11507-14.
- [6] Powell LM, Wallis SC, Pease RJ, Edwards YH, Knott TJ, Scott J. A novel form of tissue-specific RNA processing produces apolipoprotein-B48 in intestine. *Cell* **1987**; 50: 831-40.
- [7] Kawahara Y, Ito K, Sun H, Aizawa H, Kanazawa I, Kwak S. Glutamate receptors: RNA editing and death of motor neurons. *Nature* **2004**; 427: 801.
- [8] Gardiner K, Du Y. A-to-I editing of the 5HT2C receptor and behavior. *Brief Funct Genomic Proteomic* **2006**; 5: 37-42.
- [9] Blanc V, Davidson NO. C to U RNA editing: Mechanisms leading to genetic diversity. *J Biol Chem* **2003**; 278: 1395-98.
- [10] Bass BL. RNA editing by adenosine deaminases that act on RNA. *Annu Rev Biochem* **2002**; 71: 817-46.
- [11] Smith HC, Sowden MP. Base modification mRNA editing through deamination - the good the bad and the unregulated. *Trends Genet* **1996**; 12: 418-24.
- [12] Nishikura K. Editor meets silencer: crosstalk between RNA editing and RNA interference. *Nat Rev Mol Cell Biol* **2006**; 7: 919-31.
- [13] International Human Genome Sequencing Consortium. Finishing the euchromatic sequence of the human genome. *Nature* **2004**; 431: 931-45.
- [14] Brett D, Pospisil H, Valcarcel J, Reich J, Bork P. Alternative splicing and genome complexity. *Nat Genet* **2002**; 30: 29-30.
- [15] Barlati S, Barbon A. RNA editing: a molecular mechanism for the fine modulation of neuronal transmission. *Acta Neurochir Suppl* **2005**; 93: 53-57.
- [16] Wang Q, O'Brien PJ, Chen CX, Cho DS, Murray JM, Nishikura K. Altered G protein-coupling functions of RNA editing isoform and splicing variant serotonin2C receptors. *J Neurochem* **2000**; 74: 1290-1300.
- [17] Eisenberg E, Nemzer S, Kinar Y, Sorek R, Rechavi G, Levanon EY. Is abundant A-to-I RNA editing primate specific? *Trends Genet* **2005**; 21: 77-81.
- [18] Athanasiadis A, Rich A, Maas S. Widespread A-to-I RNA editing of Alu-containing mRNAs in the human transcriptome. *PLOS Biology* **2004**; 2: e191.
- [19] Levanon EY, Hallegger M, Kinar Y, et al. Evolutionarily conserved human targets of adenosine to inosine RNA editing. *Nucleic Acids Res* **2005**; 33: 1162-68.
- [20] Levanon EY, Eisenberg E, Yelin R, et al. Systematic identification of abundant A-to-I editing sites in the human transcriptome. *Nat Biotech* **2004**; 22: 1001-5.
- [21] Hartner JC, Schmittwolf C, Kispert A, Muller AM, Higuchi M, Seeburg PH. Liver disintegration in the mouse embryo caused by a deficiency in the RNA editing enzyme ADAR1. *J Biol Chem* **2004**; 279: 4894-4902.
- [22] Yamanaka S, Balestra ME, Ferrell LD, et al. Apolipoprotein B mRNA-editing protein induces hepatocellular carcinoma and dysplasia in transgenic animals. *Proc Natl Acad Sci USA* **1995**; 92: 8483-87.

- [23] Yamanaka S, Poksay KS, Arnold KS, Innerarity TL. A novel translation repressor mRNA is edited extensively in livers containing tumors caused by the transgene expression of the apoB mRNA-editing enzyme. *Genes Dev* **1997**; 11: 321-33.
- [24] Skuse GR, Cappione AJ, Sowden M, Metheny LJ, Smith HC. The neurofibromatosis type I messenger RNA undergoes base modification RNA editing. *Nucl Acids Res* **1996**; 24: 478-486.
- [25] Mukhopadhyay D, Anant S, Lee RM, Kennedy S, Viskochil D, Davidson NO. C-U editing of neurofibromatosis I mRNA occurs in tumors that express both the Type II transcript and apobec-1 the catalytic subunit of the apolipoprotein B mRNA-editing enzyme. *Am J Hum Genet* **2002**; 70: 38-50.
- [26] Paz N, Levanon EY, Amariglio N, et al. Altered adenosine-to-inosine RNA editing in human cancer. *Genome Res* **2007**; 17: 1586-95.
- [27] Paddison PJ. RNA interference in mammalian systems. *Curr Top Microbiol Immunol* **2008**; 320: 1-19.
- [28] Liang H, Landweber LF. Hypothesis: RNA editing of microRNA target sites in humans? *RNA* **2007**; 13: 463-67.
- [29] Huang J, Liang Z, Yang B, Tian H, Ma J, Zhang H. Derepression of microRNA-mediated protein translation inhibition by apolipoprotein B mRNA-editing enzyme catalytic polypeptide-like 3G (APOBEC3G) and its family members. *J Biol Chem* **2007**; 282: 33632-640.
- [30] Li XY, Guo F, Zhang L, Kleiman L, Cen S. APOBEC3G inhibits DNA strand transfer during HIV-1 reverse transcription. *J Biol Chem* **2007**; 282: 32065-74.
- [31] Harris RS, Liddament MT. Retroviral restriction by APOBEC proteins. *Nat Rev Immunol* **2004**; 4: 868-77.
- [32] Neeman Y, Levanon EY, Jantsch MF, Eisenberg E. RNA editing level in the mouse is determined by the genomic repeat repertoire. *RNA* **2006**; 12: 1802-9.
- [33] Muramatsu M, Kinoshita K, Fagarasan S, Yamada S, Shinkai Y, Honjo T. Class switch recombination and hypermutation require activation-induced cytidine deaminase (AID) a potential RNA editing enzyme. *Cell* **2000**; 102: 553-63.
- [34] Harris RS, Peterson-Mahrt SK, Neuberger MS. RNA editing enzyme APOBEC1 and some of its homologs can act as DNA mutators. *Mol Cell* **2002**; 10: 1247-53.
- [35] Gott JM, Parimi N, Bundschuh R. Discovery of new genes and deletion editing in *Physarum* mitochondria enabled by a novel algorithm for finding edited mRNAs. *Nucleic Acids Res* **2005**; 33: 5063-72.
- [36] Giege P, Brennicke A. RNA editing in *Arapidopsis* mitochondria effects 441 C to U changes in ORFs. *Proc Natl Acad Sci USA* **1999**; 96: 15324-329.
- [37] Notsu Y, Masood S, Nishikawa T, et al. The complete sequence of the rice (*Oryza sativa L.*) mitochondrial genome: frequent DNA sequence acquisition and loss during the evolution ofowering plants. *Mol Genet Genomics* **2002**; 268: 434-45.
- [38] Handa H. The complete nucleotide sequence and RNA editing content of the mitochondrial genome of rapeseed (*Brassica napus L.*): comparative analysis of the mitochondrial genomes of rapeseed and *Arabidopsis thaliana*. *Nucl Acids Res* **2003**; 31: 5907-16.
- [39] Shields DC, Wolfe KH. Accelerated evolution of sites undergoing mRNA editing in plant mitochondria. *Mol Biol Evol* **1997**; 14: 344-49.
- [40] Tsudzuki T, Wakasugi T, Sugiura M. Comparative analysis of RNA editing sites in higher plant chloroplasts. *J Mol Evol* **2001**; 53: 327-32.
- [41] Thompson J, Gopal S. Genetic algorithm learning as a robust approach to RNA editing site prediction. *BMC Bioinformatics* **2006**; 7: 145. Correction in *BMC Bioinformatics* **2006**; 7: 406.
- [42] Carillo C, Bonen L. RNA editing status of nad7 intron domains in wheat mitochondrion. *Nucleic Acids Res* **1997**; 25: 403-9.
- [43] Kondo N, Matsui E, Kaneko H, et al. RNA editing of interleukin-12 receptor beta2 2451 C-to-U (Ala 604 Val) conversion associated with atopy. *Clin Exp Allergy* **2004**; 34: 363-68.
- [44] Morse DP, Bass BL. Long RNA hairpins that contain inosine are present in *Caenorhabditis elegans* poly(A)⁺ RNA. *Proc Natl Acad Sci USA* **1999**; 96: 6048-53.
- [45] Kempken F, Bolle N, Former J, Binder S. Transcript end mapping and analysis of RNA editing in plant mitochondria. *Methods Mol Biol* **1997**; 372: 177-92.
- [46] Furey TS, Diekhans M, Lu TA, et al. Analysis of human mRNAs with the reference genome sequence reveals potential errors, polymorphisms and RNA editing. *Genome Res* **2004**; 14: 2034-40.
- [47] Wedekind JE, Dance GSC, Sowden MP, Smith HC. Messenger RNA editing in mammals: new members of the APOBEC family seeking roles in the family business. *Trends Genet* **2003**; 19: 207-16.
- [48] Hoopengardner B, Bhalla T, Staber C, Reenan R. Nervous system targets of RNA editing identified by comparative genomics. *Science* **2003**; 301: 832-836.
- [49] Sherry ST, Ward MH, Kholodov M, et al. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **2001**; 29: 308-11.
- [50] Hersberger M, Patarroyo-White S, Arnold KS, Innerarity TL. Phylogenetic analysis of the apolipoprotein B mRNA-editing region. *J Biol Chem* **1999**; 274: 34590-597.
- [51] Cummings MP, Myers DS. Simple statistical models predict C-to-U edited sites in plant mitochondrial RNA. *BMC Bioinformatics* **2004**; 5: 132.
- [52] Mulligan RM, Chang KLC, Chou CC. Computational analysis of RNA editing sites in plant mitochondrial genomes reveals similar information content and a sporadic distribution of editing sites. *Mol Biol Evol* **2007**; 24: 1971-81.
- [53] Anant S, Davidson NO. An AU-Rich sequence element (UUUN[A/U]U) downstream of the edited C in apolipoprotein B mRNA is a high-affinity binding site for Apobec-1: binding of Apobec-1 to this motif in the 3' untranslated region of c-myc increases mRNA stability. *Mol. Cell Biol* **2000**; 20: 1982-92.
- [54] Davidson NO. The challenge of target sequence specificity in C-U RNA editing. *J Clin Invest* **2002**; 109: 291-294.
- [55] Maier RM, Zeltz P, Kossel H, Bonnard G, Gualberto JM, Girenberger JM. RNA editing in plant mitochondria and chloroplasts. *Plant Mol Biol* **1996**; 32: 343-365.
- [56] Greeve J, Altkemper I, Dieterich J-H, Greten H, Windler E. Apolipoprotein B mRNA editing in 12 different mammalian species: hepatic expression is reflected in low concentrations of apoB-containing plasma lipoproteins. *J Lipid Res* **1993**; 34: 1367-83.
- [57] Mehta A, Kinter MT, Sherman NE, Driscoll DM. Molecular cloning of APOBEC-1 complementation factor a novel RNA-binding protein involved in the editing of apolipoprotein B mRNA. *Mol Cell Biol* **2000**; 20: 1846-54.
- [58] Lellek H, Kirsten R, Diehl I, Apostel F, Greeve J. Purification and molecular cloning of a novel essential component of the apolipoprotein B mRNA editing enzyme-complex. *J Biol Chem* **2000**; 275: 19848-856.
- [59] Shah RR, Knott TJ, Legros JE, Navaratnam N, Greeve JC, Scott J. Sequence requirements for the editing of apolipoprotein B mRNA. *J Biol Chem* **1991**; 266: 16301-304.
- [60] Backus JW, Smith HC. Three distinct RNA sequence elements are required for efficient apolipoprotein B (apoB) RNA editing in vitro. *Nucleic Acids Res* **1992**; 20: 6007-14.
- [61] Hersberger M, Innerarity TL. Two efficiency elements flanking the editing site of cytidine 6666 in the apolipoprotein B mRNA support mooring-dependent editing. *J Biol Chem* **1998**; 273: 9435-42.
- [62] Yamanaka S, Poksay S, Arnold KS, Innerarity TL. A novel translational repressor mRNA is edited extensively in livers containing tumors caused by the transgene expression of the apoB mRNA-editing enzyme. *Genes Dev* **1997**; 11: 321-33.
- [63] Farre JC, Leon G, Jordana X, Araya A. cis Recognition elements in plant mitochondrion RNA editing. *Mol Cell Biol* **2001**; 21: 6731-37.
- [64] Chateigner-Boutin A-L, Hanson MR. Cross-competition in transgenic chloroplasts expressing single edit sites reveals shared cis elements. *Mol Cell Biol* **2002**; 22: 8448-56.
- [65] Williams MA, Kutcher BM, Mulligan M. Editing site recognition in plant mitochondria: the importance of 5'-flanking sequences. *Plant Mol Biol* **1998**; 36: 229-37.
- [66] Choury D, Farre JC, Jordana X, Araya A. Different patterns in the recognition of editing sites in plant mitochondria. *Nucl Acids Res* **2004**; 32: 6397-6406.
- [67] Kotera E, Tasak M, Shikania T. A pentatricopeptide repeat protein is essential for RNA editing in chloroplasts. *Nature* **2005**; 433: 326-30.
- [68] Salone V, Roderger M, Polsakiewicz M, et al. A hypothesis on the identification of the editing enzyme in plant organelles. *FEBS Lett* **2007**; 581: 4132-38.

- [69] Bundschuh R. Computational prediction of RNA editing sites. *Bioinformatics* **2004**; 20: 3214-20.
- [70] Mower JP. PREP-Mt: predictive RNA editor for plant mitochondrial genes. *BMC Bioinformatics* **2005**; 6: 96.
- [71] Du P, He T, Li Y. Prediction of C-to-U RNA editing sites in higher plant mitochondria using only nucleotide sequence features. *Biochem Biophys Res Comm* **2007**; 358: 336-41.
- [72] Du, P, Li Y. Prediction of C-to-U RNA editing sites in plant mitochondria using both biochemical and evolutionary information. *J Theor Biol* **2008**; In press: Early access online.
- [73] Mulligan RM, Williams MA, Shanahan MT. RNA editing site recognition in higher plant mitochondria. *J Heredity* **1999**; 90: 338-44.
- [74] Zuker M. Calculating nucleic acid secondary structure. *Curr Opin Struct Biol* **2000**; 10(3): 303-10.
- [75] Van den Burg B, Dijkstra BW, Vriend G, Van der Vinne B, Venema G, Eijssink VG. Protein stabilization by hydrophobic interactions at the surface. *Eur J Biochem* **1994**; 220: 981-85.
- [76] Sokal RR, Rohlf JF. Biometry, third edition, W.H Freeman and Company, New York, 1995.
- [77] Baldi P, Brunak S, Chauvin Y, Andersen CA, Nielsen H. Assessing the accuracy of prediction algorithms for classification: an overview. *Bioinformatics* **2000**; 16: 412-24.
- [78] Burset M, Guigo R. Evaluation of gene structure prediction programs. *Genomics* **1996**; 34: 353-67.
- [79] Clutterbuck DR, Leroy A, O'Connell MA, Semple CAM. A bioinformatic screen for novel A-I RNA editing sites reveals recoding editing in BC10. *Bioinformatics* **2005**; 21: 2590-95.
- [80] Sixsmith J, Reenan RA. Comparative genomic and bioinformatic approaches for the identification of new adenosine-to-inosine substrates. *Methods Enzymol* **2007**; 424: 245-64.
- [81] Rubio MA, Pastar I, Gaston KW, *et al.* An adenosine-to-inosine tRNA-editing enzyme that can perform C-to-U deamination of DNA. *Proc Natl Acad Sci USA* **2007**; 104: 7821-26.
- [82] Rubio MA, Ragone FL, Gaston KW, Ibba M, Alfonzo JD. C to U editing stimulates A to I editing in the anticodon loop of a cytoplasmic threonyl tRNA in *Trypanosoma brucei*. *J Biol Chem* **2006**; 281: 115-20.

Received: June 16, 2008

Revised: June 25, 2008

Accepted: July 14, 2008