Recode-2: new design, new search tools, and many more genes

Michaël Bekaert\textsuperscript{1}, Andrew E. Firth\textsuperscript{2}, Yan Zhang\textsuperscript{3}, Vadim N. Gladyshev\textsuperscript{3}, John F. Atkins\textsuperscript{2,4} and Pavel V. Baranov\textsuperscript{5,*}

\textsuperscript{1}School of Biology and Environmental Science, University College Dublin, \textsuperscript{2}BioSciences Institute, University College Cork, Ireland, \textsuperscript{3}Division of Genetics, Department of Medicine, Brigham & Women’s Hospital and Harvard Medical School, Boston, MA 02115, USA, \textsuperscript{4}Human Genetics Department, University of Utah, UT 84112 USA and \textsuperscript{5}Biochemistry Department, University College Cork, Ireland

Received August 15, 2009; Accepted September 4, 2009

ABSTRACT

‘Recoding’ is a term used to describe non-standard read-out of the genetic code, and encompasses such phenomena as programmed ribosomal frameshifting, stop codon readthrough, selenocysteine insertion and translational bypassing. Although only a small proportion of genes utilize recoding in protein synthesis, accurate annotation of ‘recoded’ genes lags far behind annotation of ‘standard’ genes. In order to address this issue, provide a service to researchers in the field, and offer training data for developers of gene-annotation software, we have gathered together known cases of recoding within the Recode database. Recode-2 is an improved and updated version of the database. It provides access to detailed information on genes known to utilize translational recoding and allows complex search queries, browsing of recoding data and enhanced visualization of annotated sequence elements. At present, the Recode-2 database stores information on approximately 1500 genes that are known to utilize recoding in their expression—a factor of approximately three increase over the previous version of the database. Recode-2 is available at http://recode.ucc.ie

INTRODUCTION

The term ‘translational recoding’ describes the utilization of non-standard decoding during protein synthesis and encompasses such processes as ribosomal frameshifting, codon redefinition, translational bypassing and StopGo (1–7). What is often considered as a decoding error—e.g. a frameshifting error or mistranslation of a particular codon—may occasionally benefit the organism by increasing its fitness and survival. In such instances the propensity for the decoding ‘error’ may be selected for during evolution, leading to the formation of a particular sequence context that elevates the frequency of the ‘error’. To discriminate such cases of programmed decoding ‘misbehaviour’ from promiscuous translational errors or translational noise, the term recoding is used. The position within an mRNA where a recoding event takes place is termed the ‘recoding site’. Sequence elements responsible for increasing the efficiency of recoding events are termed ‘recoding stimulatory signals’, and a minimal sequence fragment that allows recoding to take place at the natural efficiency (i.e. relative to the level of standard decoding at the recoding site) is termed a ‘recoding cassette’.

Recoding can benefit gene expression in a number of ways. It can regulate gene expression by being part of a sensor for particular cellular conditions. Prominent examples include ribosomal frameshifting in bacterial release factor 2 (RF2) and eukaryotic antizyme mRNAs. In both instances, ribosomal frameshifting is required for the production of the corresponding active full-length protein products. In the RF2 mRNA, the efficiency of frameshifting is negatively regulated by the cellular concentration of its product, RF2, providing an auto-regulatory circuit for its biosynthesis (8–10). In the antizyme mRNA, the efficiency of frameshifting is modulated by cellular levels of polyamines, whose concentration in turn is controlled by antizyme (11,12). Thus, this mechanism ensures the maintenance of antizyme production at the levels required to support physiologically appropriate concentrations of polyamines. Recoding can also be used for the diversification of protein products encoded by a single gene. An illustrative example is in bacterial dnaX mRNA, where frameshifting allows synthesis of two different protein subunits—sharing the same N-terminal part—from a single open reading frame (ORF) in its mRNA (13–15). A presumed constant ratio of frameshifting in dnaX ensures a fixed stoichiometric balance between these two subunits (16). This balance,
then, is independent of the absolute levels of *dnaX* transcription and translational initiation on its mRNA. Similarly, in many viruses recoding is responsible for setting a ratio between protein products (such as those encoded by gag–pro–pol genes in retroviruses) produced from a single mRNA (17). Recoding also provides RNA viruses with a mechanism for the translation of downstream ORFs on polycistronic RNAs [other mechanisms include leaky scanning, shunting, reinitiation, IRESs and the production of subgenomic RNAs (18)] and may also be involved in global regulation mechanisms, such as mediating the switch between translation and replication on the same genomic RNA (19). Finally, recoding provides a way for the incorporation of non-standard amino acids—e.g. amino acids that share their codons with termination signals (the most prominent example of which is selenocysteine, encoded by UGA) (20–22). For further information on the diverse variety of recoding functions, see recent reviews (1,3,7,23,24).

Recoding cassettes may be composed of a variety of diverse sequence elements. For example, primary nucleotide sequences may promote re-arrangements of tRNA molecules relative to their codons in mRNA inside the ribosome or affect recognition of tRNAs or release factors in the ribosomal A-site. On the other hand, many recoding signals act in the form of RNA secondary structures, such as simple stem-loops, or more complex pseudoknots, kissing stem-loops and other structures that involve interactions between considerably distant RNA regions (19,25–28). *Trans*-acting RNA signals affecting ribosomal decoding through complementary interactions with ribosomal RNA (29–32), or through the nascent peptide acting within the ribosome exit tunnel (6,33,34), are also known. Some recoding events—such as selenocysteine insertion—require the presence of additional specialized machinery such as selenocysteine tRNAs, selenocysteine-specific translation factors and several other components of the selenocysteine biosynthesis and insertion pathway (20,35–37). Recent reviews on stimulatory signals involved in the modulation of recoding events and molecular mechanisms of recoding provide further details (7,25,27,38,39).

Despite considerable progress in the development of computational tools for the prediction of protein coding genes in sequenced genomes, the identification and annotation of recoded genes lags far behind. The hurdle lies not so much in the fact that recoded genes do not obey standard rules of genetic readout but, rather, in the considerable diversity of recoded genes and sequence elements responsible for recoding. Even among evolutionarily related genes, all utilizing recoding, the diversity of recoding signals can be considerable. An extreme example is when orthologous genes utilize recoding at different stages of gene expression to achieve the same goal. An example is in *dnaX*, where ribosomal frameshifting is employed by enterobacteria, but transcriptional slippage is used in *Thermus thermophilus* (40). A similar situation occurs in bacterial insertion sequence (IS) elements, where a certain group of IS elements utilizes transcriptional slippage to produce ORFA–ORFB fusions, while many other IS elements utilize ribosomal frameshifting for the same purpose (41). The diversity of recoding functions, combined with the wide spectrum of unrelated sequence elements involved in recoding, makes the design of a uniform model of recoding intractable. Nonetheless, in recent years, we have witnessed the development of specialized models and computational tools for the identification of particular subsets of recoding cassettes, or tools that are specific to recoding events in particular groups of homologous genes (42–45).

These developments, at least partially, were facilitated by the availability of a compiled dataset of known recoded genes collected together in the Recode database (http://recode.genetics.utah.edu), which was initially launched 9 years ago (46,47). To facilitate further development of computational tools for the prediction of recoded genes in the ever faster growing body of sequence data, as well as to provide bench researchers with up-to-date information on recoding, an efficient means of Recode database population and annotation are now required. In this article, we describe the incarnation of the database, Recode-2. The major advances of Recode-2 (hosted in a new location http://recode.ucc.ie) over previous versions include a new web design allowing enhanced visualization of stimulatory signals, a uniform RecodeML format for the annotation of recoded genes, and a significantly larger number of entries—including many recently identified cases—that altogether have more than doubled the size of the database since its last published update.

**DATABASE ORGANIZATION AND USAGE**

The data are stored in a local PostgreSQL database that is queried by PHP scripts embedded in the web interface. The schema of the PostgreSQL database is shown in Figure 1. The database stores information on individual genes that utilize recoding, the mechanisms and stimulatory signals involved, and references to the original literature sources that describe the recoding events. In order to facilitate the uniform annotation of recoding events, we have designed an XML-based format for the annotation of recoded genes, RecodeML. The document type definition for RecodeML is available in the Recode-2 web site at http://recode.ucc.ie/DTD The extensibility of the RecodeML format will allow incorporation of new annotation, if required, for newly discovered types of recoding, and the associated features, as they are being discovered. The database handles batch importation of properly designed RecodeML entries into the PostgreSQL database, thus facilitating rapid population of the database with new data.

The data in the database may be explored in two ways. They may be browsed by one of the three categories: kingdom (archaea, bacteria, eukaryotes and viruses), organism and type of recoding. The data may also be searched directly by key words that can be inserted into the search field. Searches that use regular expressions are allowed. The output of a database search is a list of Recode-2 entries in a short format that includes organism name, kingdom, genus, type of recoding event, status of
Figure 1. Entity-Relationship diagram for the Recode-2 database. The database schematic shows the relationships between the database tables. Each Recode-2 entry is composed of one entry in the Recode-2 table and a variable number of entries in related tables. PK (Primary Key) indicates the row selected as the unique identifier for a table. FK (Foreign Key) indicates rows of a table whose values match those in the primary key of a related table. (This ensures the constancy of the database.) The arrows show the connection relationships between the various tables (PK to FK).
The entry in the database and a link to the full database entry. The full description includes the following additional information: (i) the common name of the gene and the validation status of the recoding event; (ii) the organism description, giving the organism name and a link to the NCBI Taxonomy Browser (48); (iii) the sequence description, giving the Genbank (49) accession numbers for matching sequences (with links to Genbank) and links to detailed annotations of the sequences and to diagrams of RNA secondary structures involved in stimulation of the recoding event; (iv) information on the protein sequence generated as a result of recoding; (v) comments on the function of the recoding event and any additional notes and (vi) references to relevant literature (with links to corresponding abstracts in PubMed). The detailed sequence annotation appears in the form of text decorations that are described in the Help page of the Recode-2 database and are also illustrated within the Recode-1 logo itself (which can be used for rapid intuitive decoding of the text decorations and their associations with the mechanistic ways by which different sequences affect ribosome functions). To generate RNA secondary structure diagrams, PseudoViewer3 (50) is used, since it can handle complex pseudoknot RNA structures correctly. Figure 2 shows an example of sequence annotation for the human oaz1 gene, alongside a diagram of a stimulatory RNA secondary structure, and the Recode-1 logo.

NOVEL DATA

Unlike Recode-1, where all data on recoding events were introduced manually, Recode-2 also utilizes automated identification of recoding events by the recently developed computer programs ARFA (43) and OAF (44), that are able to identify and annotate +1 frameshifting events in mRNAs of bacterial RF2s and eukaryotic antizyme (OAZs), respectively. However, a significant source of recoding events remains to be serendipitous discoveries by experimental studies that sometimes are complemented...
by more systematic studies of large groups of similar genes (51,52). Therefore, a large proportion of new data are still populated manually or semi-manually. To ease manual population of recoding events, a special form has been designed that is available in the database upon user registration. User registration needs to be approved by one of the database contributors. The novel data in the database include 249 RF2 mRNAs identified by ARFA, 152 events identified by OAF, 200 new selenoprotein genes database include 249 RF2 mRNAs identified by ARFA, of the database contributors. The novel data in the database are designed that is available in the database upon user registration. User registration needs to be approved by one of the existing programs for recode annotation. We also plan to continue developing tools for the automated identification of recoding events from nucleotide sequences. As the field grows and the number of recoded genes progressively increases, it becomes harder to extract data from the relevant literature and a number of novel recoded genes may escape the database. Therefore, we encourage users and researchers in the field to submit their data directly to the Recode-2 database. We are also willing to provide help with the analysis of potential new recoding events.

ACKNOWLEDGEMENTS

We would like to express our appreciation to the colleagues who have contributed data for the previous versions of the database.

FUNDING

Science Foundation Ireland (SFI) grants (to P.V.B. and J.F.A.); National Institutes of Health grants (to J.F.A. and V.N.G.). Funding for open access charge: Science Foundation Ireland.

Conflict of interest statement. None declared.

REFERENCES


