Poster B-52

CUDXML: Codon Usage Database in XML Format

Authors:
Denis Shestakov (*Turku Centre for Computer Science*)
Tapio Salakoski (*University of Turku*)

Short Abstract: In this work, we present Codon Usage Database in XML Format (CUDXML), an XML-based database for codon usage data. Our motivation was to provide researchers involved in codon usage studies with a comprehensive database of codon usage patterns for all known complete protein coding sequences (calculated based on current GenBank release).

Long Abstract:

It has been found that there is the unequal use of synonymous codons coding for amino acids in many organisms, both prokaryotes and eukaryotes. The codon preference may vary considerably not only among different organisms but also among genes of the same organism. Numerous studies of codon usage bias done over the past two decades show that translation selection is responsible for the unequal usage of synonymous codons in protein coding genes in a number of species. In general, researchers have been successfully applying codon usage analysis in molecular evolution studies, particularly to identify highly expressed or horizontally transferred genes.

The most recognized repository containing the codon usage of all the full-length protein gene entries in the international DNA sequence databases is Codon Usage Database [1] (designated as CUTG) at the Kazusa DNA Research Institute. The purpose of the database is to provide users with a pre-calculated codon usage datasets to be used for codon usage-based analyses. Unfortunately, there are several limitations which complicate the use and access to codon usage data stored in the CUTG repository. Firstly, the database has restricted query capabilities, namely: (1) search for organism's codon usage by name of organism, and (2) search for codon usage patterns of sequences (within only one organism) whose features correspond to specified keywords. In this way, the CUTG does not support many typical queries essential for codon usage analysis (e.g., obtain codon usage patterns for all sequences having some certain features in several genomes of interest). Secondly, the data in the repository is represented in a plain text format which obviously complicates the programmatic access to data. From the perspective of a user on the other side, the format is user-friendly and requires additional reformatting for the convenient perception. At last, the database technically contains only codon frequencies for all complete protein sequences and does not store measures (e.g., relative synonymous codon usage) and indices (e.g., codon adaptation index) which are calculated based on codon frequencies and actually used in codon usage analysis. This all leads to the fact that many users prefer to write their own scripts or use existing tools for computation of codon usage rather that retrieve necessary pre-calculated patterns from the CUTG repository.

In this work, we address these issues and present Codon Usage Database in XML Format (CUDXML), an XML-based database for codon usage data. Our motivation was to provide researchers involved in codon usage studies with a comprehensive database of codon usage

patterns for all known complete protein coding sequences (calculated based on current GenBank [2] sequence database release). The key point of our approach is using the CUData XML Schema proposed by us for codon usage data. There are several pluses of putting data in XML-based format. In particular, data if in XML is self-describing and more reusable especially when utilized by dissimilar applications. Among other advantages of the representation in XML are simplified programmatic access to data and useful transformations via XSLT (data may be transformed to HTML, text or even PDF formats). In general, the CUData Schema is capable of representing occurrences of overlapping or non-overlapping words in nucleotide sequences. Note that counts for non-overlapping words of length of two, four and three bases are known as di-, tetranucleotide and codon usage patterns correspondingly. The proposed format is designed to deal with codon frequencies as well as the most popular codon usage-based measures and indices. The CUDXML is implemented using Berkeley DB XML database. The scripts for computing DNA sequence measures and data converting are written in Perl (with help of the BioPerl modules). The web interface to the CUDXML allows a user to browse codon usage patterns for the whole organisms as well as, unlike the CUTG repository, for each particular protein coding sequence. Additionally, the CUDXML provides extensive query capabilities. User queries formulated in such standard XML query languages as XPath and XQuery can be issued on the levels of organisms, sequences, and codons (e.g., obtain frequencies of codons coding glycine for all sequences in several genomes of interest). Since we expect the CUDXML will be accessed by both humans and software applications we provide an access to the CUDXML via a conventional web interface and via a program-friendly interface (based on SOAP protocol). Statistical calculations on codon usage patterns are unsupported in the current version of CUDXML but support of several statistical methods widely used in codon usage studies is a planned activity for the near future.

[1] Y. Nakamura, T. Gojobori, and T. Ikemura. Codon usage tabulated from international DNA sequence databases: status for the year 2000. Nucl. Acids Res., 28(1):292–, 2000.
[2] D. Benson, I. Karsch-Mizrachi, D. Lipman, J. Ostell, and D. Wheeler. Genbank. Nucl. Acids Res., 34(Database issue):D16-20, 2006.