25. Codon usage analysis based on web services

Denis Shestakov, Tapio Salakoski, Mauno Vihinen.

Presenter (PhD student): Denis Shestakov
Affiliation: Turku Centre for Computer Science
Country: Finland

The standardized interface to sequence analysis tools is highly desirable to integrate bioinformatics applications. We implemented a SOAP (Simple Object Access Protocol) server and web services providing a program-friendly interface to codon usage analysis. All data are managed in XML-based formats to make exchanging and reusing information more simple.

Keywords: codon usage, web services, sequence analysis workflow

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 codon usage analysis workflow can be divided into three steps: (1) selection of DNA sequences to analyze; (2) computation of codon frequencies and other codon bias indices (such as effective number of codons, codon adaptation index, etc.) for selected sequences; and (3) analysis of values calculated at step 2 using statistical methods (for instance, using correspondence analysis). Currently existing software packages and programming toolkits (BioPerl, BioJava, etc.) used for codon usage analysis have several drawbacks. One of them is limited support for data exchange and, in particular, passing data from step 2 to statistical module at step 3 usually involves needless mapping from one format to another. Next limitation is a complexity of reusing codon usage analysis results, which are tables of codon usage frequencies as well as statistical result data obtained at the end of step 3. Even those stand-alone programs combining all three above-mentioned steps are able to calculate only predefined codon bias measures and to perform exactly some certain type of statistical analysis. Additionally, these programs are not flexible to analyze codon position, e.g., codon usage of the first ten codons after the start codon for a list of coding regions. In this work, we address these issues and present a web services-based facility for codon usage analysis. Our motivation was to simplify programmatic access to analysis tools involved in the codon usage analysis pipeline, and to provide to the community an XML format specifically for codon usage data. We implemented a SOAP (Simple Object Access Protocol) server and web services providing a program-friendly interface to codon usage analysis. All data including query requests to the web services are managed in XML formats. We use the Bioinformatic Sequence Markup Language (BSML) as a main data container, StatDataML (XML format for statistical data) to represent statistical data, and CUData XML Schema (XML format developed by us) to represent calculated codon frequencies and codon bias indices. The web service responsible for statistical analysis provides an interface to the R statistical environment and can perform analysis using several statistical methods (including correspondence analysis and principal component analysis). Output of the web services is in XML format and, thus, can be stored in local XML databases.