

EXPRESSIVITY OF PROTEIN FAMILIES BY CODON ADAPTATION INDEX ANALYSIS: THE FAMiCOD ANALYSER PROJECT

M. Ramazzotti, G. Manao G. Ramponi and D. Degl'Innocenti

Dipartimento di Scienze Biochimiche, Università degli Studi di Firenze, viale Morgagni 50 50134 Firenze, Italy.
matteo.ramazzotti@unifi.it

INTRODUCTION

All the organisms that have been studied so far have shown a largely different usage of synonymous codons. These differences seem to be due to the cellular tRNA abundance and therefore to a different regulation of tRNA and aminoacyl tRNA-synthetase transcription and to different aminoacyl tRNA-synthetase activity. The codon usage seems not to be an evolutionary constraint since it has been found large differences among strictly related organisms. As a result, highly expressed proteins tend to be coded by species-specific "optimized" coding sequences composed by the most abundant codons. The basic meaning of this behaviour is to minimize the risk of tRNA depletion during intense translation and misincorporation of amino acids from rare codons (1). The analysis of the codons used in the coding sequences of proteins may therefore be an index of protein expression. The codon usage can be measured according to a number of criteria, spanning from raw count of the used codons to complex multivariate analysis. The most simple and sufficiently confident method seems to be the Codon Adaptation Index (CAI), which measures the variability of the codon usage in a gene in respect to the variability of a reference set of genes (2). This reference set must be composed of highly expressed genes in order to give a convenient correlation between CAI and gene expression. It is easy to understand that some proteins are conserved at high level into a cell, such as some ribosome associated protein or some basic transcription factor. It may be an underlying assumption that all organisms present such proteins at high levels. Many researchers have confirmed through microarray analysis an effective correlation between gene expression and codon usage. Here we present a simple and effective method to predict a set of highly expressed genes to be used in estimation of protein expression. Thanks to this approach a genome-wide analysis of gene expression may be performed and, thanks to protein classification databases, protein families expression may be estimated. This leads to the evaluation of the effective role of some interesting pathways in different organisms.

METHODS AND ALGORITHMS

The Family Codon (FAMiCOD) Analyser Project is a package of programs dedicated to the codon usage analysis and basically to the retrieval of highly expressed genes from whole genome CDS data without the need of experimental resources. The basic computational method of the programs is a recursive Codon Adaptation Index comparison which allows the estimation of coding sequences index of expression with respect to the codon usage table built on a set of reference coding sequences. The CAI value, which mathematically may range from 0 to 1, ranges from 0,1 to 0,9 when applied to true proteins. The primary goal of the system is to improve the quality of the reference dataset and in the end to define it as the set of highly expressed genes for an organism. According to this set, the program is able to calculate the CAI for a given coding sequence (or even a protein name, if a coding sequence is associated to it in EMBL database) or for all the proteins coded in a genome. Repeating this approach for a number of organisms, one is able to compare the expression rates of the protein families.

Obtaining family members

The FAMiCOD FamiFind program is a Swiss-Prot automatic retrieval system which allows to enter a verbose protein family query and to retrieve all the correspondance in the database. For each protein ID, the program downloads the protein sequences from Swiss-Prot, its coding sequence from EMBL and checks for the correctness of the CDS. If the CDS is valid, the program retrieves the owner organism name and downloads, from CUTG database, the corresponding genome-based codon usage table. Thanks to this the gCAI (genomic CAI) is calculated. The scope of the program is to provide the checked CDS for all the protein family members and the list of organisms on which to retrieve the highly expressed genes. The gCAI serves solely as a non-randomness checker at this stage.

Checking for non randomness

The FAMiCOD FamiStrap program is intended to verify the effective translatability of a protein, even if this aspect should be implicit in already verified

Retrieval of highly expressed genes

The FAMiCOD HighXP retriever program is a GUIDed interface to first approaching genome data. Thanks to direct FTP connection with the NCBI database one can choose an organism of interest and download three main files: a fasta formatted file containing all known CDS, a fasta formatted file containing all the corresponding proteins and a reference table with chromosome positions, gene names, protein descriptions and many other features. As a first step, the exact correspondance between gene and proteins is performed, since it has been observed that rare internal stop codons or non GATC characters (bases) in CDSs, which is not acceptable for triplet analysis. An error-free organism specific local database is rebuilt. At this stage, two main approaches have been implemented. One can choose to "seed" a protein family (e.g. ribosomal) and to build a temporary reference dataset (and therefore a temporary codon usage table) with the CDS corresponding to the protein extracted from the "description" feature of the downloaded reference table. This approach is not intended for not annotated genomes, for which the seeding procedure may not produce any results. The

other approach is to build a genomic codon usage table assuming all the CDSs as the reference dataset. From both approach, the next step is to apply the codon table to whole CDS dataset and to define a codon adaptation index threshold value. The program will produce the CAI for all the proteins of the organism, but will list only the proteins with an over threshold value. Since the most "organism adapted genes" emerges from all other genes, these have to be included in a list of "interesting genes". This new dataset is then used to build a new codon table and the whole genome CAI calculation is performed again. This time a smaller list of genes will emerge from other. Such genes should include the highly expressed ones, but it has to be refined (discarding some, including some other). The new set may be used to build a new codon table or to modify the existing one, depending on the number of the sequences. By repeating those improving rounds of selection and by raising slowly the CAI inclusion threshold, the highly expressed genes dataset can be obtained.

VALIDATION AND RESULTS

The most important program of the FAMiCOD Analyser Project is the HighXP Retriever which in the end determines the quality of the final results. Since our method contains many assumptions, we validated our strategy of highly expressed gene finding with the genes dataset produced by other authors [Carbone et al. 2003]. They validated their prediction method with known microarray experiments for prokaryotes such as *Escherichia coli*, *Bacillus subtilis* and *Haemophilus influenzae*, defining the reference sets containing 1% of the whole genome CDSs. We defined three different reference sets of CDS for each organism, containing more than 1%, 1% and less than 1% CDSs. We performed the CAI calculation on all the genes contained in the three genomes with codon usage tables derived from our datasets and from the other sets. We then observed the distribution of the complete genome CDS CAI values for each organism and we reported the average difference values between datasets. We found out that the best correlation is effectively obtained if the sets contain the 1% of genes with respect to the total number of genes in the genome. This means that the CAI values are strictly dependent from the dataset used. We therefore assumed the 1% is the best population for the reference set of highly expressed genes. We then observed that the discrepancies between our values and reference values were poorly dependent on the measured values, indicating that CAI is a good estimating mechanism. According to validation results, we performed the reference dataset creation for all the available organisms. In the end we calculated the CAI value for each acylphosphatase.

We tested all the FAMiCOD Analyser Project programs with the protein acylphosphatase family. We firstly retrieved all available family members by using "acylphosphatase" as a verbose family query in FamiFind program. We obtained 76 entries, some of them missing the coding sequence or the codon usage table for the corresponding organism, therefore unavailable for gCAI calculation. In about 5:30 minutes we performed the gCAI calculation of about 50 acylphosphatase members which shared an average index of 0.75, a non random associated value, as expected. We therefore collected 50 coding sequences with the corresponding organisms.

For each organism we downloaded the whole genomes CDS, Protein and Reference table files. Thanks to the iteration performed by HighXP Retriever program we were able to define the sets of highly expressed genes

REFERENCES:

1. Ikemura T. (1981) *J. Mol. Biol.* 146:1-21.
2. Sharp P.M. and Li W.H. (1987) *Nucleic Acids Res.* 15:1281-95.
3. Gasteiger E et al. (2003) *Nucleic Acids Res.* 31:3784-88.
4. Nakamura Y., Gojobori T. and Ikemura, T. (2000) *Nucl. Acids Res.* 28, 292.