BIOINFORMATICS APPLICATIONS NOTE Vol. 20 no. 13 2004, pages 2148–2149
doi:10.1093/bioinformatics/bth204

TIP: protein backtranslation aided by genetic algorithms

Andrés Moreira and Alejandro Maass*

Centro de Modelamiento Matemático and Departamento de Ingeniería Matemática FCFM, U. de Chile, Casilla 170/3-Correo 3, Santiago, Chile

Received February 18, 2004; revised and accepted on March 10, 2004
Advance Access publication April 1, 2004

ABSTRACT

Summary: Several applications require the backtranslation of a protein sequence into a nucleic acid sequence. The degeneracy of the genetic code makes this process ambiguous; moreover, not every translation is equally viable. The usual answer is to mimic the codon usage of the target species; however, this does not capture all the relevant features of the 'genomic styles' from different taxa. The program TIP ('Traducción Inversa de Proteínas') applies genetic algorithms to improve the backtranslation, by minimizing the difference of some coding statistics with respect to their average value in the target.

Availability: http://www.cmm.uchile.cl/genoma/tip/

Contact: anmoreir@dim.uchile.cl

INTRODUCTION

The backtranslation of a peptide into a DNA sequence may be required in several contexts: for instance, for the *in vivo* production of an artificial protein, or of a protein for which the gene is unknown. It may even happen that the gene is known, but a different gene is required in order to introduce it in a different species, with a different 'genomic style' [for instance, Smith *et al*. (1990), perform backtranslation to introduce a plant gene in *Escherichia coli*].

The existence of different genomics styles was noticed first in Grantham (1986) where the idiosyncratic nature of codon usage was recognized. The current approach to backtranslation in all commercial and non-commercial software—to the best of our knowledge—relies precisely on codon usage: the known codon frequencies of the target species are used as the probabilities of codon assignation for each amino acid, and thus the codon usage of the target species is reproduced; the only exception is White and Seffens (1998), where a neural network is allowed to see a small window of the amino acid sequence. However, codon usage (and local frequencies in general) are not a complete description of the different coding styles. We have reviewed, computed and discussed (Moreira, 2004; http://arxiv.org/abs/physics/0304016) the most important 'coding statistics' found in the literature, identifying those which distinguish the genomic styles of the species or larger taxa.

The codings statistics which have been found to characterize genomic styles are nucleotide frequencies (which, thanks to Chargaff's laws, can be described by the G+C content), codon usage, dinucleotide frequencies [termed 'genome signature' by Karlin and Mrázek (1997)], index of DNA homogeneity [IDH; defined in Miramontes *et al*. (1995)], Fourier spectra and autocorrelation functions. Only the first two of these properties are recovered when backtranslation is performed in the traditional way, and the latter two demand the consideration of a window larger than those used by White and Seffens (1998).

The purpose of our software is to improve on this, applying genetic algorithms to force the backtranslated sequence to mimic the target style, described by some of these coding statistics. In addition, the software can be easily modified to include other, user-defined properties.

SOFTWARE OVERVIEW

Source code (in standard C++) is available, apart from Windows binary files. There are two programs: FileEdit and Tip. FileEdit is just a complementary utility, used for filtering, extracting and converting sequences files into GenBank and FASTA formats. It also includes tests for discarding homologue or identical sequences. FileEdit is used for preparing the files needed in Tip, especially for the generation of species' profiles.

Tip is the main program, and works on sequence files in the FASTA format; it also generates some files in formats of its own, mainly for parameters and species profiles. It currently computes (and imitates) G + C level, relative codon usage, IDH, discrete Fourier transform and autocorrelations functions. The last three are applied to binary sequences, and are therefore combined with the three possible binary projections of the DNA alphabet.

A species profile is represented in a distance parameters file, which includes the evaluation of the different coding statistics. This file is generated by Tip, with the average of the

*To whom correspondence should be addressed.

evaluation of the statistics in some representative set of known coding sequences of the species. The distance between a given sequence and a given target species, for a given coding statistic, is then computed as the absolute value of the difference between its evaluation on the sequence, and the value contained in the distance parameters file (in the case of Fourier transform and autocorrelation functions, a weighted average along the curves is used).

Backtranslation proceeds as follows. A protein sequence is given, and a distance parameters file (i.e. a target species) is selected. A population of backtranslations ('guesses') is generated, in the traditional form, imitating codon usage. A genetic algorithm is then applied on succesive generations of the guesses population, leading to a minimization of a weighted average of the distances between the coding statistics of the guesses and those of the target species. The scheme used for the genetic algorithm is very standard, and can be customized by the user; all the weights in the different stages are customizable as well.

The program interacts with the user through the standard input (there is no GUI). The input is read in complete lines, so that the standard input can be redirected to an input file in order to use the program in a batch form. The source code was written in a modular way, to allow any programmer to add new properties which will be immediately incorporated to Tip's evaluation and optimization of sequences, after recompilation. Short tutorials for FileEdit, for Tip and for the modification of the code are provided in the website. More details about the coding statistics and their idiosyncratic features, as well as a better description of the genetic algorithm, can be found in Moreira (2004), (http://arxiv.org/abs/physics/0304016) and in other documents at the website. Supplementary material for that article can be found there too, including coding sequence files for 12 model species, extracted from GenBank release 131, some distance parameters files and examples of usage.

DISCUSSION

The definitive test for this backtranslation tool would be the *in vitro* generation of artificial genes (we expect it to perform better than the simpler approach). Meanwhile, an *in silico* experiment consisting in the backtranslation of a human protein into 'bacterial' style, and the comparison of the statistics of the resulting gene with those of an homologue bacterial gene (see Moreira, 2004), suggests that our approach is correct. Another possible use of this software is in the analysis of sequences: massive backtranslation of sequences under one or several criteria can illuminate the discussions on the relations and origins of features observed in coding statistics; it may also be useful for understanding the changes in genes after horizontal transfer to a new species, and for creating surrogate data in diverse applications.

ACKNOWLEDGEMENTS

The work was started during a visit to the GREG (Group de Recherche et d'Etude sur les Genomes) at the Institut of Mathematics at Luminy, University of Marseille, France. We were supported by CONICYT through the FONDAP program in Applied Mathematics.

REFERENCES

Grantham,R., Perrin,P. and Mouchiroud,D. (1986) Patterns of codon usage in different kinds of species. *Oxford Surv. Evol. Biol.*, **3**, 48–81.

Karlin,S. and Mrázek,J. (1997) Compositional differences within and between eukaryotic genomes. *Proc. Natl Acad. Sci., USA*, **94**, 10227–10232.

Miramontes,P., Medrano,L., Cerpa,C., Cedergren,R., Ferbeyre,G. and Cocho,G. (1995) Structural and thermodynamic properties of DNA uncover different evolutionary histories. *J. Mol. Evol.*, **40**, 698–704.

Moreira,A. (2004) Genetic algorithms for the imitation of genomic styles in protein backtranslation. *Theoret. Comput. Sci.*, **322**, 297–312.

Smith,A.T., Santama,N., Dacey,S., Edwards,M., Bray,R.C., Thorneley,R.N. and Burke,J.F. (1990) Expression of a synthetic gene for horseradish peroxidase C in *Escherichia coli* and folding and activation of the recombinant enzyme with Ca^{2+} and heme. *J. Biol. Chem.*, **265**, 13335–13343.

White,G. and Seffens,W. (1998) Using a neural network to backtranslate amino acid sequences. *Electronic J. Biotechnol.*, **1**, 3.