Cryptically Simple Proteins Sequences in Eukaryotic Proteomes

M. Mar Alba[1], Michelle Simon[2], and John M. Hancock[2]

[1] Evolutionary Genomics - Genome Bioinformatics Research Group in Biomedical Research (GRIB), Dept of Experimental and Health Sciences (CEXS), Universitat Pompeu Fabra (UPF). Passeig Marítim, Barceloneta 37-49, 08003 Barcelona, Spain.
malba@imim.es

[2] Bioinformatics Group, MRC Mammalian Genetics Unit, Harwell, Oxfordshire OX11 0RD, U.K.
m.simon@har.mrc.ac.uk, j.hancock@har.mrc.ac.uk

Keywords. Proteomics, Protein sequence analysis, Simple sequences, Amino acid repeats

Introduction

Eukaryotic genomes contain a rich complement of repetitive sequences, ranging from transposable elements to simple sequence repeats such as mini- and microsatellites. Tautz et al [1] defined a class of simple sequence repeats which they called cryptically simple sequences (CSS). These correspond to regions of low complexity and contain concentrations of one or a few short motifs at above random expectation.

Analyses of CSS at the genome level [2, 3] show that they are present in all organisms but are much more common in eukaryotes than prokaryotes, apparently reflecting different mutational processes in the different types of organism. CSS in ribosomal RNA genes of insects show similar mutational patterns to microsatellites [4] suggesting they also mutate primarily by replication slippage.

Tandem repeated sequences are generally rare in genes, probably due to selection against frameshifts due to gain and loss of repeated units. Tandem repeats of codons are less rare and have been shown to occur in 7.6% of *S. cerevisiae* proteins. Tandem repeats in some human proteins are associated with neurological diseases but also provide the opportunity for rapid evolutionary diversification of protein function [5].

A number of studies have suggested associations of polyglutamine repeats, which are associated with many of the human diseases, with transcription factors in particular. In *S. cerevisia*e not only transcription factors but also protein kinases have been shown to be significantly enriched in polyglutamine repeats [6].

Here we describe the application of the SIMPLE algorithm, used to detect CSS, to protein sequences [7] and consider the distribution and functional associations of this novel class of amino acid sequence in *S. cerevisiae* and mammalian (human and mouse) proteins.

The SIMPLE Algorithm

SIMPLE scores windows within a sequence for recurrences of the motif at the centre of the window. The ratio of the average window score for a sequence to the average score obtained for a number of random sequences of the same length and composition provides a global measure of the repetitiveness of the sequence. Individual windows with unexpectedly high scores can also be identified, allowing repeated motifs and their locations to be identified. The application of this procedure to protein sequences [7] is illustrated in the Figure:

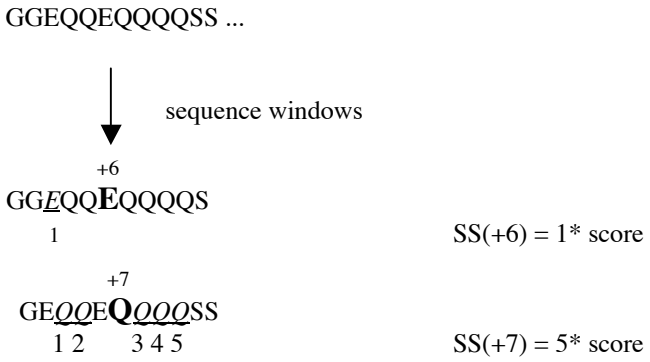

Fig. 1 Calculation of simplicity scores for protein sequences.
The number of occurrences of a repeat starting at the central reference element in the sequence window is counted and multiplied by the score for a repeat of that length (user-defined). If repeats of different length are considered simultaneously the scores are added together. Numbers above the sequence indicate the relative position of the reference element, which is indicated in bold. Numbers below the sequence indicate the numbers of reiterations of the reference element within the sequence window.

Results

Frequencies of cryptic amino acid repeats have been calculated for the yeast (*S. cerevisiae*) and two mammalian (human and mouse) proteomes. The relative frequencies of these repeats are similar to those of tandem amino acid repeats but cryptic repeats are more common. Associations with gene function also show similarities and differences from those seen for tandem repeats [6]. Results will be presented in more detail at the conference.

References

[1] D. Tautz, M. Trick and G. A. Dover, Cryptic simplicity in DNA is a major source of genetic variation. *Nature*, 322: 652-656, 1986.

[2] J. M. Hancock. The contribution of slippage-like processes to genome evolution. *J. Mol. Evol.* 41: 1038-1047, 1995.

[3] J. M. Hancock. Genome size and the accumulation of simple sequence repeats: Implications of new data from genome sequencing projects. *Genetica* 115: 93-103, 2002.

[4] J. M. Hancock and A. P. Vogler. How slippage-derived sequences are incorporated into rRNA variable region secondary structure: implications for phylogeny reconstruction. *Mol. Phyl. Evol.* 14: 366-374, 2000.

[5] J. M. Hancock. Evolution of sequence repetition and gene duplications in the TATA-binding protein TBP (TFIID). *Nuc. Acids Res.* 21: 2823-2830, 1993.

[6] M. M. Alba, M. F. Santibáñez-Koref and J. M. Hancock. Amino acid reiterations in yeast are overrepresented in particular classes of proteins and show evidence of a slippage-like mutational process *J. Mol. Evol*, 49: 789-797, 1999.

[7] M. M. Alba, R. A. Laskowski and J. M. Hancock. Detecting cryptically simple protein sequences using the SIMPLE algorithm. *Bioinformatics* 18: 672-678, 2002.