Finding genes by comparing genomes: the case of selenoproteins

Roderic Guigò

GENOME BIOINFORMATICS RESEARCH LAB
Grup de Recerca en Informàtica Biomèdica
Institut Municipal d'Investigació Mèdica - Universitat Pompeu Fabra

Although the genome sequence and gene content are available for an increasing number of organisms, eukaryotic selenoproteins remain poorly characterized. In these proteins, selenium (Se) is incorporated in the form of selenocysteine (Sec), the 21st amino acid. Selenocysteine is cotranslationally inserted in response to UGA codons (a stop signal in the canonical genetic code). The alternative decoding is mediated by a stem-loop structure in the 3'UTR of selenoprotein mRNAs (the SECIS element). Selenium is implicated in male infertility, cancer and heart diseases, viral expression and ageing. In addition, most selenoproteins have homologues in which Sec is replaced by cysteine (Cys). Genome biologists rely on the high-quality annotation of genomes to bridge the gap from the sequence to the biology of the organism. However, for selenoproteins, which mediate the biological functions of selenium, the dual role of the UGA codon confounds both the automatic annotation pipelines and the human curators. In consequence, selenoprotein are misannotated in the majority of genome projects. Furthermore, the finding of novel selenoprotein families remains a difficult task in the newly released genome sequences.

In the last few years, we have contributed to the exhaustive description of the eukaryotic selenoproteome (set of eukaryotic selenoproteins) through the development of a number of *ad hoc* computational tools. Our approach is based on the capacity of predicting SECIS elements, standard genes and genes with a UGA codon in-frame in one or multiple genomes. Indeed, the comparative analysis plays an essential role because 1) SECIS sequences are conserved between close species (eg. human-mouse); and 2) sequence conservation across a UGA codon between genomes at further phylogenetic distance strongly suggests a coding function (eg. human-fugu). Our analysis of the fly, human and fugu genomes have resulted in 8 novel selenoprotein families. Therefore, 19 distinct selenoprotein families have been described in eukaryotes to date. Most of these families are widely (but not uniformly) distributed across eukaryotes, either as true selenoproteins or Cys-homologues. The recent completion of the *Tetraodon nigroviridis* and *Fugu rubripes* genomes has allowed us to investigate the eukaryotic selenoproteome in a restricted and largely unexplored window within the vertebrate phylogeny. Our investigation has resulted in the identification of a novel selenoprotein family, currently under study, which appears to be restricted to actinopterygians among vertebrates.

The correct annotation of selenoproteins is thus providing insight into the evolution of the usage of Sec. Our data indicate a discrete evolutionary distribution of selenoproteins in eukaryotes and suggest that, contrary to the prevalent thinking of an increase in the number of selenoproteins from less to more complex genomes, Sec-containing proteins scatter all along the complexity scale. We believe that the particular distribution of each family is mediated by an ongoing process of Sec/Cys interconversion, in which contingent events could play a role as important as functional constraints. The characterization of eukaryotic selenoproteins illustrates some of the most important challenges involved in the completion of the gene annotation of genomes. Notably among them, the increasing number of exceptions to our standard theory of the eukaryotic gene and the necessity of sequencing genomes at different evolutionary distances towards such a complete annotation.