Available online at www.sciencedirect.com
SCIENCE DIRECT®
C. R. Biologies 328 (2005) 882–899
COMPTES RENDUS BIOLOGIES
http://france.elsevier.com/direct/CRASS3/

Review / Revue

Protein variety and functional diversity: Swiss-Prot annotation in its biological context

Brigitte Boeckmann [a,*], Marie-Claude Blatter [a], Livia Famiglietti [a], Ursula Hinz [a], Lydie Lane [a], Bernd Roechert [a], Amos Bairoch [a,b]

[a] *Swiss Institute of Bioinformatics, Centre médical universitaire, 1, rue Michel-Servet, 1211 Genève 4, Switzerland*
[b] *Department of Structural Biology and Bioinformatics, Centre médical universitaire, 1, rue Michel-Servet, 1211 Genève 4, Switzerland*

Received 13 May 2005; accepted after revision 5 June 2005

Available online 28 July 2005

Presented by Stuart Edelstein

Abstract

We all know that the dogma 'one gene, one protein' is obsolete. A functional protein and, likewise, a protein's ultimate function depend not only on the underlying genetic information but also on the ongoing conditions of the cellular system. Frequently the transcript, like the polypeptide, is processed in multiple ways, but only one or a few out of a multitude of possible variants are produced at a time. An overview on processes that can lead to sequence variety and structural diversity in eukaryotes is given. The UniProtKB/Swiss-Prot protein knowledgebase provides a wealth of information regarding protein variety, function and associated disorders. Examples for such annotation are shown and further ones are available at http://www.expasy.org/sprot/tutorial/examples_CRB. ***To cite this article: B. Boeckmann et al., C. R. Biologies 328 (2005).***

Résumé

Un gène, plusieurs protéines : l'annotation de Swiss-Prot dans le contexte biologique. Il est maintenant évident pour tout le monde que le dogme « un gène, une protéine » est obsolète. Au cours de la synthèse d'une protéine fonctionnelle, le transcrit et la chaîne polypeptidique peuvent être modifiés de multiples façons. Ces modifications ont une incidence directe sur la fonction biologique de la protéine et dépendent non seulement de l'information génétique, mais également des conditions dans lesquelles se trouve la cellule : un nombre limité d'isoformes protéiques est produit dans une cellule donnée, à un moment précis. Cet article dresse un bref inventaire des processus biologiques impliqués dans la formation de protéines différentes à partir d'un même gène chez les eucaryotes, ainsi qu'une description des diversités structurelle et fonctionnelle qui en découlent. La banque de connaissances sur les protéines UniProtKB/Swiss-Prot est particulièrement riche en informations décrivant l'origine des différences entre les séquences de protéines dérivées d'un même gène, les modifications post-traductionnelles, ainsi que les conséquences de cette variabilité sur leur(s) fonction(s) et, le cas échéant, les maladies associées. De nombreux exemples

* Corresponding author.
E-mail address: brigitte.boeckmann@isb-sib.ch (B. Boeckmann).

1631-0691/$ – see front matter
doi:10.1016/j.crvi.2005.06.001

d'annotation sont décrits et d'autres sont disponibles sur le site http://www.expasy.org/sprot/tutorial/examples_CRB. ***Pour citer cet article : B. Boeckmann et al., C. R. Biologies 328 (2005).***

Keywords: Protein database; Annotation; Protein synthesis; Sequence variety; Post-translational modification; Protein–protein interaction; Disease

Mots-clés : Banque de données sur les protéines ; Annotation ; Synthèse protéique ; Diversité des séquences ; Modifications post-traductionnelles ; Interaction protéine–protéine ; Maladie

1. Introduction

Despite an abundant biodiversity, all living beings are based on a similar cellular system, which is run by a population of self-organized molecules: proteins. Proteins catalyze, regulate and control most procedures that occur in a cell for the benefits of the whole organism. If we wish to understand how living systems work, it is important to understand how proteins function. The way life evolved facilitates this task in that we can compare not only organisms but also physiological processes and their components. Consequently, conclusions can be drawn from the findings and applied from one system to another. In the past decades, numerous methods – e.g., sequence comparisons, protein family prediction, detection of functional domains, conserved domains and altered amino-acid positions within these regions, motif searches, structure prediction or phylogenetic studies – and databases have been developed for the efficient prediction of a protein's function. However, such methods require the 'correct' amino-acid sequence as input; the retrieval of such input is still a challenging task [1]. In eukaryotes especially, the formation of the nascent amino-acid sequence implies the possible creation of a huge number of isoforms. Without further experimental evidence it is impossible to predict the existing and biologically relevant proteins from the total of all possible variants. The same is true for most of the other alterations in a protein's structure. What is more, there is still a long way to go to understand how polypeptides interact with each other in order to accomplish a specific task. In this respect, the study of inheritable diseases can be very informative and demonstrate how the smallest of deviations from a protein's structure can lead to its dysfunction and be the cause of severe disorders.

This article gives an overview on cellular processes underlying sequence variety and structural diversity. We have chosen to focus on eukaryotes to narrow down our discussion. The UniProtKB/Swiss-Prot protein knowledgebase [2,3] aims to record all protein variations and their functional impact. Throughout the text, examples of corresponding Swiss-Prot annotation are given and the reader is encouraged to look at further examples when the primary accession number is indicated (e.g. P12345). Complete Swiss-Prot examples are provided at http://www.expasy.org/sprot/tutorial/examples_CRB. Swiss-Prot entries can also be retrieved from the ExPASy server [4] by building the URL, e.g., http://www.expasy.org/uniprot/P12345.txt for the raw format, http://www.expasy.org/uniprot/P12345 for the NiceProt format, or by entering the accession number in the quick search at http://www.expasy.org/ (NiceProt format). Further examples of Swiss-Prot entries can be found via the relevant keywords. Details on the format of Swiss-Prot entries are provided in the user manual at http://www.expasy.org/sprot/userman.html.

2. Formation of the nascent amino-acid sequence

The vast majority of eukaryotic proteins are nuclear-encoded, as are most of the proteins which are the products of DNA-containing organelles. Mitochondria and plastids generate only a small fraction of their own proteome [5–10]. Regulatory mechanisms during protein synthesis can influence the concentration, destination, sequence variety, structural diversity and thus the functional options of the resulting protein. Out of a broad variety of possible mRNAs and protein sequences that can be generated from a single gene (Fig. 1), only one or a few are created at a time, dependent on the type of tissue or the stage of development.