# Post-translational modifications: a challenge for proteomics and bioinformatics

Appel, Ron David; Bairoch, Amos Marc

# Editorial

## Post-translational modifications: A challenge for proteomics and bioinformatics

The Editors of this special issue:
Ron D. Appel (left) and Amos Bairoch (right)

There are now over one million different protein sequences available in the UniProt knowledgebase [1] (the combination of Swiss-Prot and TrEMBL). This wealth of sequence information is a bonanza to life scientists involved in studying life processes. Yet, the emperor is naked! The overwhelming majority of protein sequences are predicted from genomic data. The first consequence of this state of affairs is that some of these predictions are incorrect. This is especially relevant to eukaryotes where it is still difficult to build reliable gene models. Fortunately, full-length cDNA sequencing projects and comparative genomic studies are slowly helping towards the resolution of this shortcoming. The second and more insidious consequence of dealing with predicted sequences is that they do not represent the final active biological entity. Almost all proteins undergo some form of processing which generally has an impact not only on the chemical nature of the protein but also on its function, its subcellular location and its propensity to interact with other macromolecules.

Proteins are subject to three classes of protein modifications, pre-, co- and post-translational modifications. The majority of modifications are made when the protein is already folded, these are real post-translational modifications (PTMs). Some modifications are made while the polypeptide is still being synthesized on the ribosome, and these are called co-translational modifications. Finally two 'non-standard' amino acids (selenocysteine and pyrrolysine) can be incorporated into proteins by modification of some 'standard' amino acids while they are charged on special tRNAs, and these events are called pre-translational modifications. Post-translational modifications are defined [2] as the "series of chemical reactions whereby a newly synthesized polypeptide chain is converted to a functional protein". PTMs can themselves be classified into three categories: proteolytic cleavage of part of the sequence (removal of an initiator methionine, a signal sequence, a transit peptide, *etc.*), adjunction of a chemical group (acetylation, glycosylation, phosphorylation, *etc.*) and formation of inter- or intra-peptidic linkages (disulfide bonds, thioether links, *etc.*).

The world of protein modifications is therefore a very complex one. Current mass spectrometric methods can be used to identify and characterize a small percentage of these modifications. The characterization of the full spectrum of modifications requires a palette of experimental techniques that cannot all be currently applied in a high-throughput manner. It is therefore no anexaggeration to state that today we only see the tip of the iceberg in terms of PTMs. We therefore expect that the experimental detection of PTMs will be one of the major experimental challenges for proteomics in this decade.

For the field of bioinformatics, PTMs are also a challenge. There is an urgent need for software tools that efficiently predict the occurrence of various PTMs on an as yet uncharacterized protein sequence. To build such tools is not an easy task for at least two reasons. The first one is the lack of experimental data. The availability of a representative corpus of reliable "positive" and "negative" cases is required for the development of a good discriminative predictive method. While it is sometimes possible to obtain enough examples of experimentally

verified PTMs, it is far more difficult to derive a set of known protein sequences that have been shown not to harbour a specific PTM. Negative results are seldom published, yet they are essential to the comprehension of the scope of any life process! The second problem is one which is much more insidious. A protein sequence may well be potentially modified by a PTM, yet it may never have a chance to physically be in contact with the enzyme that could carry out the modification. Therefore the prediction of PTMs on a given protein is not only dependent on its sequence and its 3-D structure, but also on its subcellular localization, its developmental stage and its tissue specificity. The development of such bioinformatics tools is of the utmost importance not only for the precise prediction of potential PTMs, but also for identifying existing PTMs from experimental proteomic data.

This special issue of PROTEOMICS deals with two bioinformatics aspects of the world of protein modifications:

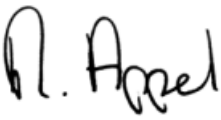
– The representation in databases of theses modifications

– Bioinformatics methods to predict PTMs in protein sequences

In the section on databases you will find two articles describing resources that cater for the chemical and biological description of PTMs (Garavelli; Creasy and Cottrell); an article describing how the UniProt knowledgebase deals with the storage and representation of PTMs (Farriol-Mathis *et al*.); and two articles that describe phosphorylation site databases (Hornbeck *et al.*; Wurgler-Murphy *et al.*).

In the section on predictions there are four articles describing methods to predict protein targeting signals (Schneider and Fechner; Small *et al.*; Reczko and Hatzigeorgiou; Gonnet *et al.*); two articles describing methods to detect lipid modifications (Eisenhaber *et al.*; Bologna *et al*.); two articles on the detection of glycosylation sites (Blom *et al.*; Joshi *et al.*); an article on the prediction of disulfide bonds (Martelli *et al.*) and an article on the quantitative estimation of modifications from 2-D gels (Kumar *et al.*).

We hope that you will find this special issue a useful survey of the state of a field that could be termed “PTMomatics”!



Ron D. Appel

Head of Proteome Informatics Group
Swiss Institute of Bioinformatics,
Geneva University and
Geneva University Hospital

Amos Bairoch

Head of Swiss-Prot Group
Swiss Institute of Bioinformatics and
University of Geneva

PS: Those of you who are either very sharp-eyed, as well as those who know Ron Appel or Amos Bairoch, will have noticed that the picture shown at the top of this editorial is a composite of the faces of both editors. Apart from the fact that we both always enjoy a good joke, we also wanted to symbolize the elusive nature of PTMs: a quick glance at them is not always enough to understand the full picture!

[1] Apweiler R., Bairoch A., Wu C. H., Barker W. C. *et al. Nucleic Acids Res.* 2004, *32*, D115–D119.
[2] Han K.-K., Martinage A., *Int. J. Biochem.* 1992, *24*, 19–28.