3 Protein Identification in Proteomics

PATRICIA HERNANDEZ, PIERRE-ALAIN BINZ, AND MARC R. WILKINS

Abstract

Identifying proteins contained in a biological sample represents a central task in a majority of proteomics projects. The present chapter discusses various attributes that can be used to identify proteins: the species of origin, the isoelectric point, the amino acid composition, sequence tags and mass values. Then, the chapter focuses on protein identification using mass spectrometry data. It introduces the mass spectrometry technology, presents specific instrumentations and describes the major identification approaches, including peptide mass fingerprinting, peptide fragment fingerprinting and de novo sequencing. Available software tools are described, and for each approach, a concrete example of identification results is given. Particular care is attributed to the identification of proteins carrying post-translational modifications.

3.1 Introduction

Protein identification plays a central role in proteomic research. It is the essence of projects that aim to catalogue all proteins present in a biological sample. It is inherent to protein-expression profiling, which seeks to discover and identify differentially expressed proteins. In association with efforts to better understand molecular mechanisms and pathways, it is also linked with the mapping of protein co- and post-translational modifications.

The number of proteins to be identified in proteomics projects is massive. It is therefore essential to use automated identification techniques that generate unequivocal results. Currently, the most widespread method for protein identification is the correlation of spectra of proteins and peptides obtained by mass spectrometry (MS) with protein sequence data stored in databases. These mass spectra are called 'fingerprints' because they represent a unique key for a protein that can be used for its identification. When sequence databases are searched, a number of additional attributes can be used in association with spectral data. For example, taxonomic information about the origin of the sample can be used to restrict a search to proteins
from one or more relevant species. Annotations in entries of the Swiss-Prot database, such as information on subcellular compartments, may be of further help. When available, other attributes like the protein molecular weight and isoelectric point, as well as knowledge of its sequence or amino acid composition, are also useful. These last attributes are discussed briefly in the following sections. Then, the chapter presents details of protein identification using MS.

3.2 Attributes of Proteins Useful for Their Identification

3.2.1 Species of Origin

Protein species of origin is obviously not, by itself, a sufficiently powerful attribute to allow the unequivocal identification of any protein; however, it can be used to limit a database search to proteins from relevant species. Usually, identification tools allow the choice from a number of kingdoms, phyla and one or more genera. When an organism under study is not listed in an identification tool, the search can be extended to related species whose proteins have high sequence similarity. This process is called cross-species identification.

3.2.2 Protein Isoelectric Point

The isoelectric point (pI) of a denatured protein is a function of its amino acid composition, N- and C-terminal amino acids, and any post-translational modifications. On two-dimensional electrophoresis (2-DE) polyacrylamide gels (see Chap. 2), the experimental (apparent) pI of many proteins can be estimated at the same time. Estimation is typically done by first calculating the theoretical pI of ten or more proteins on a 2-DE gel, using a tool such as Compute pI/MW (1994). Theoretical pI values for these proteins can then be used with image-analysis software to create a pI grid, allowing the pI of other proteins on the gel to be estimated. The accuracy of these estimates will depend largely on the care that is taken in the construction of the grid, and the type of sample being studied.

3.2.3 Protein Mass

In database searches, the molecular mass of a protein can be used to dramatically reduce the number of sequences that might be the identity of a protein. Protein mass can be theoretically determined by summing the mass of all amino acids of a protein together with the mass of any post-translational