33 Mizuguchi, K. *et al.* (1998) Getting knotted: a model for the structure and activation of Spätzle. *Trends Biochem. Sci.* 23, 239–242
34 Armstrong, P.B. and Quigley, J.P. (1987) *Limulus* α2-macroglobulin: first evidence in an invertebrate for a protein containing an internal thiol ester bond. *Biochem. J.* 248, 703–707
35 Armstrong, P.B. *et al.* (1998) α2-Macroglobulin does not function as a C3 homologue in the plasma hemolytic system of the American horseshoe crab *Limulus*. *Mol. Immunol.* 35, 47–53
36 Levashina, E.A. *et al.* (2001) Conserved role of a complement-like protein in phagocytosis revealed by dsRNA knockout in cultured cells of the mosquito, *Anopheles gambiae*. *Cell* 104, 709–718
37 Adema, C.M. *et al.* (1997) A family of fibrinogen-related proteins that precipitates parasite-derived molecules is produced by an invertebrate after infection. *Proc. Natl. Acad. Sci. U. S. A.* 94, 8691–8696
38 Jiang, H. and Kanost, M.R. (2000) The clip-domain family of serine proteases in arthropods. *Insect Biochem. Mol. Biol.* 30, 95–105
39 Muta, T. *et al.* (1990) Proclotting enzyme from horseshoe crab hemocytes: cDNA cloning, disulfide locations, and subcellular localization. *J. Biol. Chem.* 265, 22426–22433
40 Iwanaga, S. *et al.* (1998) New types of clotting factors and defense molecules found in horseshoe crab hemolymph: their structures and functions. *J. Biochem.* 123, 1–15
41 Thielens, N.M. *et al.* (1999) Structure and functions of the interaction domains of C1r and C1s: keystones of the architecture of the C1 complex. *Immunopharmacology* 42, 3–13
42 Patthy, L. (1990) Evolution of blood coagulation and fibrinolysis. *Blood Coagul. Fibrinolysis* 1, 153–166
43 Stenflo, J. (1999) Contributions of Gla and EGF-like domains to the function of vitamin K-dependent coagulation factors. *Crit. Rev. Eukaryotic Gene Expr.* 9, 59–88
44 Celie, P.H.N. *et al.* (2000) Hydrophobic contact between the two epidermal growth factor-like domains of blood coagulation factor IX contributes to enzymatic activity. *J. Biol. Chem.* 275, 229–234
45 Patthy, L. *et al.* (1984) Kringles: modules specialized for protein binding: homology of the gelatin-binding region of fibronectin with the kringle structures of proteases. *FEBS Lett.* 171, 131–136
46 Rich, T. *et al.* (2000) How low can Toll go? *Trends Genet.* 16, 292–294
47 Pujol, N. *et al.* (2001) A reverse genetic analysis of components of the Toll signaling pathway in *Caenorhabditis elegans*. *Curr. Biol.* 11, 809–821
48 Davie, E.W. *et al.* (1991) The coagulation cascade: initiation, maintenance, and regulation. *Biochemistry* 30, 10363–10370

What does it mean to identify a protein in proteomics?

Juri Rappsilber and Matthias Mann

The annotation of the human genome indicates the surprisingly low number of ~40 000 genes. However, the estimated number of proteins encoded by these genes is two to three orders of magnitude higher. The ability to unambiguously identify the proteins is a prerequisite for their functional investigation. As proteins derived from the same gene can be largely identical, and might differ only in small but functionally relevant details, protein identification tools must not only identify a large number of proteins but also be able to differentiate between close relatives. This information can be generated by mass spectrometry, an approach that identifies proteins by partial analysis of their digestion-derived peptides. Information gleaned from databases fills in the missing sequence information. Because both sequence databases and experimental data are limited, a certain ambiguity often remains concerning which sequence variant(s) and modification(s) are present. As the common denominator of all the isoforms is a gene, in our opinion, it would be more accurate to state that a product of this particular gene rather than a certain protein has been identified by mass spectrometry.

With the completion of the human genome project, it has become clear that organism complexity is generated more by a complex proteome than by a complex genome. The proteome is defined here as the time- and cell-specific protein complement of the genome; that is, it encompasses all proteins that are expressed in a cell at one time, including isoforms and protein modifications. Whereas the genome is constant for one cell, largely identical for all cells of an organism, and does not change very much within a species, the proteome is very dynamic with time and in response to external factors, and differs substantially between cell types. Protein analysis methods such as antibody binding or mass spectrometry have been key techniques to study both individual proteins and entire proteomes. However, proteins are usually 'identified' using these methods with the concept that one gene encodes one protein. The protein diversity that can result from a single gene locus demands a more precise concept taking into consideration the results obtained using different tools.

Antibodies

Antibodies are raised against an antigen. If the antigen was a peptide derived from a protein or the entire protein itself, the antibody can be used to recognize that protein. Antibodies recognize the three-dimensional arrangement of charges, and the hydrophilic and/or hydrophobic properties of peptides or proteins. Such a physical landscape is not always unique to one amino acid sequence, resulting in non-specific binding of the antibody to proteins that were not used as antigen. In general, this cross-reactivity is more pronounced with polyclonal antibodies, which are a mixture of antibodies against the same antigen, than with monoclonal antibodies. Specific antibodies against protein isoforms can be raised with variable success. However, antibodies normally have the ability to bind to different protein isoforms regardless of differences caused by, for example, alternative splicing or protein modifications, and often bind even to homologues. This allows researchers to find the orthologue of a protein in another species whose genome is not sequenced, and to investigate protein families using single antibodies (as was the case for the Sm proteins [1,2], a group of human autoantigens involved in pre-mRNA splicing). By contrast, cross-reactivity can result in binding of antibodies to unrelated proteins. For example, the human homologue of the yeast splicing factor Prp6p [3,4] was first cloned as a protein recognized by an antibody raised against an epitope of the human NF-κB

http://tibs.trends.com 0968-0004/02/$ – see front matter © 2002 Elsevier Science Ltd. All rights reserved. PII: S0968-0004(01)02021-7

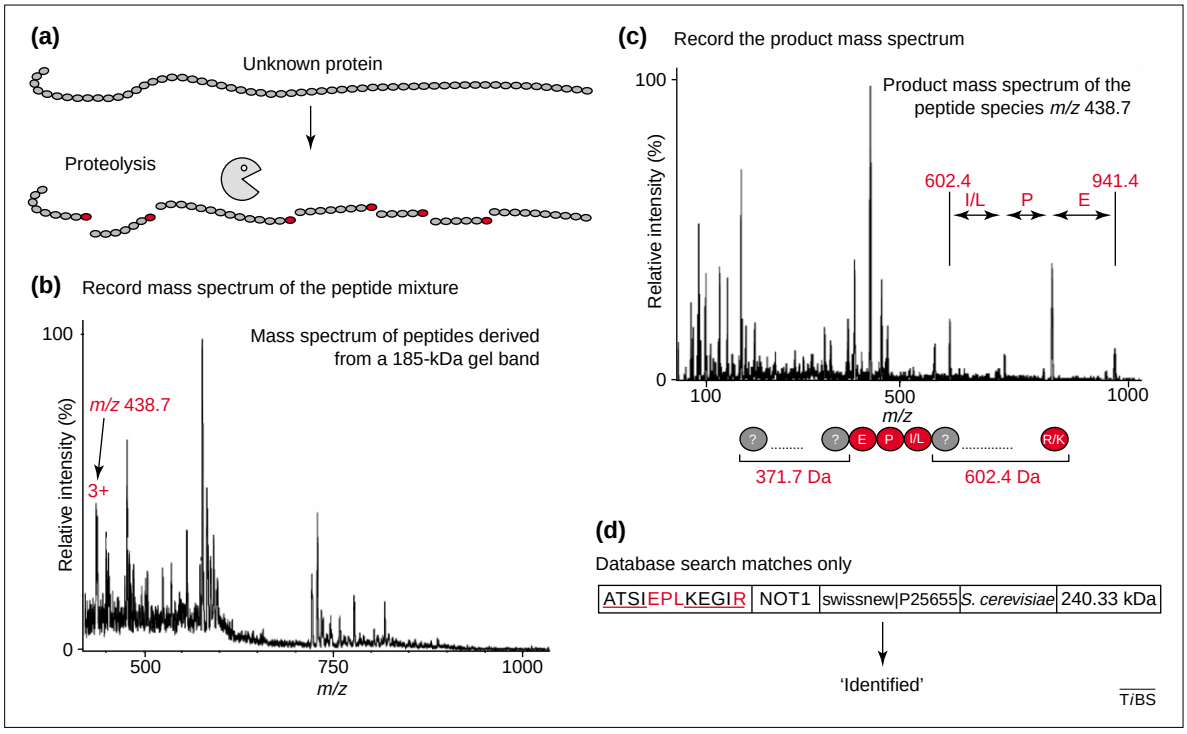

Fig. 1. Mass spectrometric identification of a protein. A protein is digested using a highly specific protease (typically trypsin) (a) and the derived peptides are analysed in a mass spectrometer (b). One peptide species is selected for collision with an inert gas, such as argon, in the mass spectrometer. The derived fragments of this peptide are measured to give the product mass spectrum (c). Some of these fragments differ in their mass by individual amino acids – leucine and isoleucine having identical masses. Part of the sequence can therefore be read out from a series of peaks in the spectrum. This sequence information is placed in the peptide by the mass of the fragments, and is used in the 'peptide sequence tag' in conjunction with the mass of the peptide and the specificity of the protease (tryptic digest results in K or R at the C terminus of the peptide) to search for a match in the database. A single peptide sequence tag is usually sufficient to unambiguously link a database entry with the investigated protein (d). Alternatively, fragment masses are measured and automatically compared with the predicted fragment masses of all peptides derived from a database to find the best match. In any case, the confidence increases with the number of fragmented peptides matching the same entry. Typically 2–70% of the sequence of the entry is covered by the experimental data, depending on the sample amount. Abbreviations: *m/z*, mass : charge ratio; *S. cerevisiae*, *Saccharomyces cerevisiae*.

p65 subunit involved in signal transduction [5] and, so far, there is no indication of a functional link between these two proteins.

Although antibodies have been extremely useful tools, their binding is not necessarily proof that a product of a certain gene is present. To understand the detailed biological process in which a protein functions, it is essential to know not only the individual gene whose product is involved but also the exact state (e.g. splice variant, processed state and modification) of the gene product.

Mass spectrometry

This information can be obtained by mass spectrometry. This technique identifies a protein not by analysing it directly but by analysing the peptides derived from digestion of the protein with a protease, usually trypsin. The problem of identifying a protein is thus reduced to the problem of identifying peptides. The advantage of such an approach lies in, among other things: (1) the ease with which peptides can be obtained by digestion and elution from a gel-purified protein, compared with the difficulty of eluting the intact protein from the gel; (2) the higher sensitivity for smaller molecules by mass spectrometry; and (3) the better fragmentation behaviour of peptides when compared with proteins. The obtained peptides can be separated using liquid chromatography or directly analysed by mass spectrometry [6,7]. Within the mass spectrometer, individual peptides are separated and fragmented by collision with an inert gas. The masses of the derived fragments are then measured. This produces a read out of sequence stretches that can be followed by database searches – the 'peptide sequence tag' approach (Fig. 1). Alternatively, databases can be searched using the fragment mass data itself. The information obtained from fragmenting a peptide in a mass spectrometer is sufficient to unambiguously identify peptides in large sequence databases, such as the human genome, provided that the operator is an educated user.

It is important to note that the entire sequence of the protein is not usually determined by mass spectrometry. Instead, the information from a subset of its peptides is used to find a matching database entry. The sequence information contained in the database entry then fills those sequence stretches for which no experimental data was obtained. Protein identification via peptide sequencing is thus a combination of experimental data and data deposited in sequence databases.

Does the identification of one or several peptides allow the conclusion that a protein has been identified? Yes, in organisms in which one gene gives

Juri Rappsilber
Matthias Mann
Protein Interaction Laboratory in the Center of Experimental Bioinformatics, Dept of Biochemistry and Molecular Biology, University of Southern Denmark, Campusvej 55, DK-5230 Odense M, Denmark.
e-mail: rappsilber@bmb.sdu.dk; mann@bmb.sdu.dk

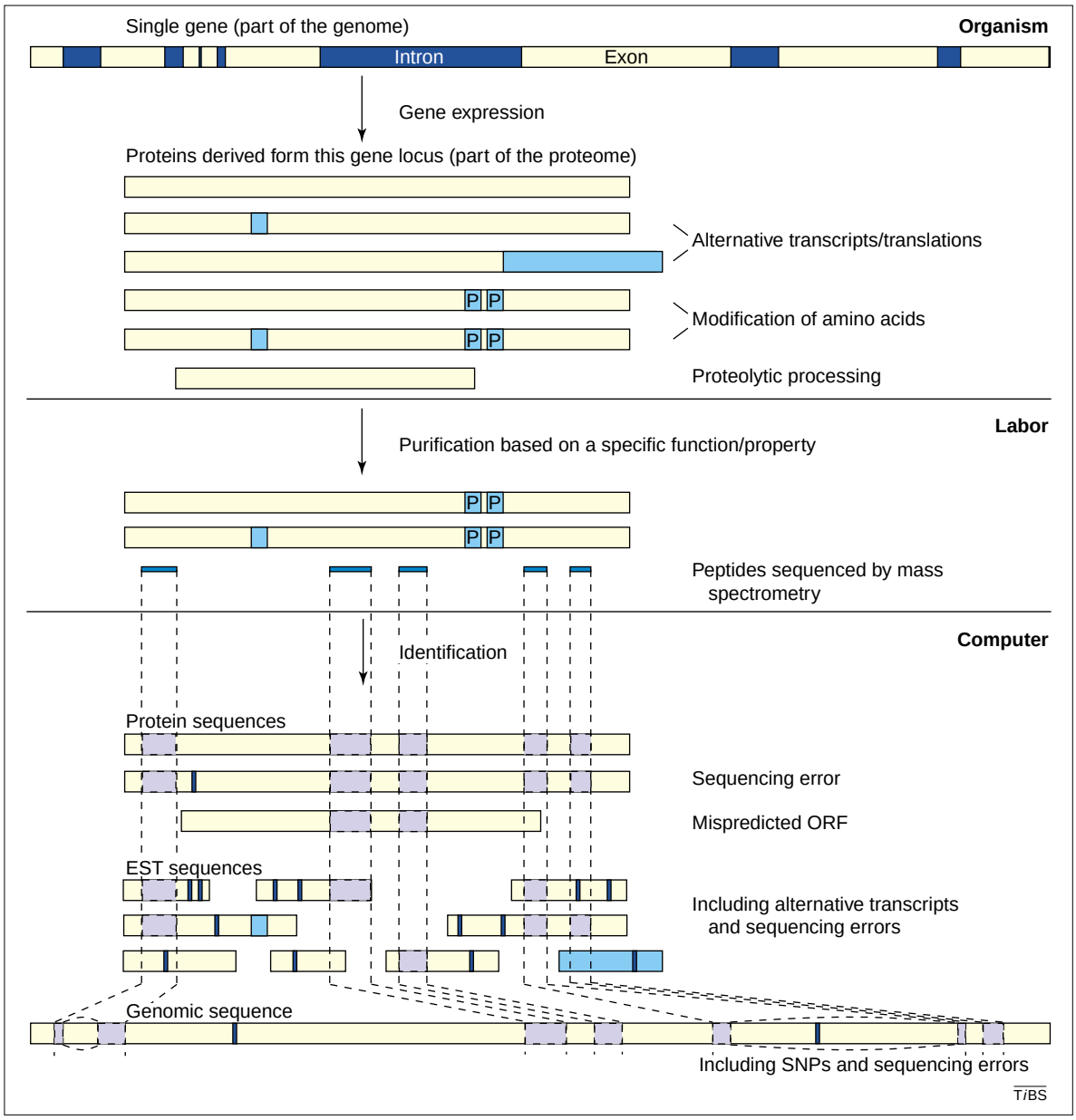

Fig. 2. Exemplary view of the sequence diversity that can be found in the cell, in laboratory experiments and in databases, and how this diversity is addressed by mass spectrometric peptide data. In an organism, all protein isoforms are encoded by the genome. In lower organisms, proteins are encoded in one continuous stretch whereas in higher organisms, coding regions (exons) are disrupted by non-coding regions (introns). Gene expression, which is highly regulated by numerous mechanisms, often results in a cell- and time-specific occurrence of several protein isoforms. Purification in the laboratory is based on specific properties such as binding to a protein complex or response in a functional assay. This is not always sufficient to resolve the natural occurring sequence diversity, and more than a single protein might be obtained. The difference between the purified isoforms and the resolution of the detection methods determine whether the mixture is detected at this point of the analysis. In the next step, mass spectrometry and databases are used in combination to identify the protein(s). Mass spectrometry offers experimental data on the sequence of a set of peptides derived from the protein(s). Sequences in the databases are then used to fill the gaps left by the incomplete experimental data. However, sequence databases offer an incomplete and somewhat incorrect representation of the naturally occurring sequence diversity. There are three types of sequence database, each having respective advantages. (1) The protein sequence databases, containing the translated results of cDNA sequencing efforts, provide the highest quality data. The protein sequence is available in a continuous stretch from its N to C terminus. Annotation sometimes includes observed modifications and sequence isoforms. However, sequence errors do occur and open reading frames (ORFs) are sometimes mispredicted. (2) Expressed sequence tag (EST) databases contain the results of high-throughput single-read cDNA sequencing efforts. Because of their abundance, ESTs are a valuable but also error prone source to detect sequence alternatives. Neither of these two databases contain all possible protein sequences. (3) In principle, genomic sequence databases contain, once the genomic sequence of the investigated species (e.g. human) is completely sequenced, all possible variants that are based on sequence alternatives. However, if information is not easily read out, for example, because of intronic stretches, it does not allow for filling the sequence gaps left by the peptide data. When prediction-based annotation remains problematic, the focus can be reversed – mass spectrometric data can be used to help the annotation, for example, when peptides are sequenced that bridge two exons. These three databases are the only source of information available to fill in the gaps left by the mass spectrometric peptide data. Whatever natural diversity is not covered by the experimental data or the databases remains unknown. As the result of the protein analysis shown here, the protein sample was connected with the correct gene, some protein sequence alternatives in the databases were not supported, and some sequence diversity – the modification of amino acids – remained undetected. Abbreviations: P, represents a modified amino acid; SNP, single nucleotide polymorphism.

rise only to a single protein. This is generally the case for many of the viral, archea, prokaryotic and lower eukaryotic proteins. However, higher eukaryotes tend to have more, very similar proteins (Fig. 2).

Several diverse mechanisms can result in the expression of many protein variants from the same gene locus in one species: single nucleotide polymorphisms, gene splicing, alternative splicing of pre-mRNA, RNA editing, translational frame shifts and hopping, co- and post-translational modification of amino acids, and proteolytic cleavage of the protein. Depending on the assay used to study a function, a mixture of isoforms can be purified in the laboratory but not noted as such, for example because of their similar migration in SDS–PAGE. The derived peptides are partially sequenced by mass spectrometry. Because the sequences of these peptides usually do not cover the entire sequence of the protein, one relies on the database to fill in the missing parts of the protein. Therefore, the experimental data generally do not allow for differentiating between sequence alternatives present in the extra information provided by the database. Also, if this extra information does not cover all isoforms, isoforms that are present might go undetected.

Even though the difference between two proteins can be very small – a single amino acid change or a single amino acid modification – it could be crucial for the function of the protein. Here, the identity of a protein is defined by its structural formula (connectivity of all atoms); that is, not only by the amino acid sequence but also by any modifications. One peptide is usually sufficient for identification of the gene encoding the investigated protein (a seven amino acid stretch is usually unique in the human genome), whereas several peptides give the gene product to near identity. Still, it is not necessarily known which alternative splice form or which modification is present, and it would be more accurate to state that a product of a certain gene, rather than a certain protein, has been identified. For example, a mass spectrometric analysis of a protein sample would not only identify a member of the caspase protease family, but also the species and particular member: human caspase 6 (from 14 known human caspases), isoform α (from two known splice variants), small subunit (from two known proteolytic variants). However, it might not be clear whether the N-terminal 13 amino acids have been cleaved off, and whether the protease is therefore in its active form. Although a protein sequence can be covered to 100% by mass spectrometric data, in routine work this is not practical. Large amounts of protein are required for such exhaustive analysis, and a lot more effort is required in the mass spectrometric analysis.

Natural sequence diversity, as well as sequenced fragments and sequence errors, lead to a host of database entries containing largely identical sequence information. Protein identification depends on the accuracy of the database entry and is not always able to differentiate between these largely identical database entries because of the reasons described above. In large-scale analyses where the peptides observed can vary extensively from experiment to experiment, one entry or the other will be reported preferentially. For different experiments, lists of identified entries will result that apparently contain different proteins, even though the date would match to all alternative entries. Grouping these alternative matches from different experiments into one final result is difficult. Therefore, databases that are as non-redundant and as accurately annotated as possible are a necessity for large-scale proteomics experiments.

Future developments

Mass spectrometric analysis of digestion-derived peptides can produce an incomplete description of the investigated protein. A way around this problem is to include data relating to the entire protein in the analysis. Ideally, this would be an accurate mass measurement of the protein. If the measured mass of the uncleaved protein, matched by peptide sequencing to a database entry, is also in agreement with the calculated mass deduced from the database entry, then the protein is identified – keeping in mind that some isoforms are isobaric (i.e. have the same mass : charge ratio), thereby limiting even this approach. Although such a measurement is often possible, the practicalities of protein weight measurements and the sensitivities of such measurements are currently limited. Also, the coupling of this measurement to fragmentation of intact proteins has not yet been established routinely as a means to identify the protein [8]. Of the current analytical approaches, separation techniques that are based on protein size and/or isoelectric point (pI), such as one- and two-dimensional gel electrophoresis, can separate protein variants to a certain extent. In-solution digestion of complex protein mixtures, followed by liquid chromatography–mass spectrometry techniques, relies exclusively on peptide sequencing and has, therefore, a limited potential to differentiate between protein variants.

We are developing methods to use the information in the databases to direct the mass spectrometric analysis towards detecting sequence variations and modifications. After the matching sequences have been retrieved from the database using fragmented peptides, these sequences are aligned and peptides predicted from unique stretches. The observed species corresponding to these predicted peptides are then preferentially sequenced during the remainder of the mass spectrometric experiment. These key peptides enable differentiation of isoforms that are represented in the database. This approach optimizes the information that can be obtained from mass spectrometry and even increases the efficiency with which unpredicted differences are found.

Conclusion

Methods are under continuous development to accommodate not only the complexity of large genomes but also complex proteomes. However, depending on the data that is available, one has to be careful as to what can be concluded from a protein analysis. Depending on the antibody used, very different information levels are accessible. For example, it might be possible to identify only the gene family to which the investigated protein belongs, or the detailed stage of the protein product with respect to a specific modification. In contrast to antibody approaches, mass spectrometric analysis of a human protein usually identifies a single gene. However, it is more difficult to say exactly which of the proteins (up to 10^6 [9]) possibly encoded by that particular gene is present. This can only be concluded after full characterization. Mass spectrometry can yield very detailed information about a protein sample, including an extensive analysis of protein modifications. However, it is not possible to state whether or not sequence alternatives or protein modifications other than those detected are present. For this reason, it would be useful to include the full set of mass spectrometric peptide data in the appendix of a publication. This clearly defines the level of information that could be obtained in that investigation, and allows researchers to judge the depth of the presented findings and the relation of the protein to data obtained in their own research with the same, or a very similar, protein.

Acknowledgements
We thank members of our institute for helpful discussions and critical reading of this manuscript. J.R. is a Marie Curie Fellow. This work was supported by a generous fund of the Danish National Research Foundation to the Center of Experimental Bioinformatics.

References

1 Lerner, M.R. and Steitz, J.A. (1979) Antibodies to small nuclear RNAs complexed with proteins are produced by patients with systemic lupus erythematosus. *Proc. Natl. Acad. Sci. U. S. A.* 76, 5495–5499
2 Rokeach, L.A. *et al.* (1992) Mapping of the immunoreactive domains of a small nuclear ribonucleoprotein-associated Sm-D autoantigen. *Clin. Immunol. Immunopathol.* 65, 315–324
3 Makarov, E.M. *et al.* (2000) The human homologue of the yeast splicing factor prp6p contains multiple TPR elements and is stably associated with the U5 snRNP via protein–protein interactions. *J. Mol. Biol.* 298, 567–575
4 Rappsilber, J. *et al.* (2001) SPF30 is an essential human splicing factor required for assembly of the U4/U5/U6 tri-small nuclear ribonucleoprotein into the spliceosome. *J. Biol. Chem.* 276, 31142–31150
5 Nishikimi, A. *et al.* (1999) A novel mammalian nuclear protein similar to *Schizosaccharomyces pombe* Prp1p/ Zer1p and *Saccharomyces cerevisiae* Prp6p pre-mRNA splicing factors. *Biochim. Biophys. Acta* 1435, 147–152
6 Mann, M. *et al.* (2001) Analysis of proteins and proteomes by mass spectrometry. *Annu. Rev. Biochem.* 70, 437–473
7 Aebersold, R. and Goodlett, D.R. (2001) Mass spectrometry in proteomics. *Chem. Rev.* 101, 269–295
8 Meng, F. *et al.* (2001) Informatics and multiplexing of intact protein identification in bacteria and the archaea. *Nat. Biotechnol.* 19, 952–957
9 Missler, M. and Südhof, T.C. (1998) Neurexins: three genes and 1001 products. *Trends Genet.* 14, 20–26

Pete Jeffs is a freelancer working in Paris, France.