Use of mass spectrometry-derived data to annotate nucleotide and protein sequence databases

Matthias Mann and Akhilesh Pandey

Mass spectrometry-based proteomic methodologies can be used to annotate both nucleotide and protein sequence databases. Because such data have to be derived from proteins, they can be used to identify coding regions of the genome as well as to predict the complete primary sequence of proteins and their expression patterns and post-translational modifications.

Complete sequences of several genomes including Saccharomyces cerevisiae, Caenorhabditis elegans and Drosophila melanogaster as well as a draft of the human genome have already been reported. This has led to an information explosion and we now have more data available to us than we can manage. For example, take the issue of the total number of genes in the human genome. For years it was presumed to be approximately 80,000–100,000 genes. It was expected that once the human genome was fully sequenced, it would be trivial to calculate the exact number of genes. Paradoxically, as we approach the finishing line with respect to sequencing, several groups have re-estimated the total number of genes and their latest speculations now range from 35,000 to 120,000 (Refs 6,7). Where exactly did we go wrong? Given the stark reality that the prediction accuracy rate of gene prediction programs is approximately 40%, it is clear to see where we might have erred (see results of the Drosophila Genome Annotation Assessment Project in Ref. 8). The mouse sequencing project, which will be available to the public this year, is estimated to help tremendously in annotating coding regions in the human genome. It should be noted that prediction of genes is still a much simpler task than predicting, for example, a phosphorylation site or the function of a domain or a protein itself.

http://tibs.trends.com 0968-0004/01/S – see front matter © 2001 Elsevier Science Ltd. All rights reserved. Pt: S0968-0004(01)00172-6-6
Proteomics is any large-scale analysis of proteins and the proteome has largely been ignored. The term 'proteomics' is any large-scale analysis of proteins and has been applied to studies ranging from identification of spots from 2D electrophoretic gels to characterization by treatment of a specific cell line, by a specified growth factor for a given time. See Fig. 1 for a representation of the various scenarios in which mass spectrometric analysis interfaces with biological experiments. For example, inherent in the identification of a phosphorylation site in a peptide is the information that this protein was phosphorylated by treatment of a specific cell line, by a specified growth factor for a given time. See Fig. 1 for a representation of the various scenarios in which mass spectrometric analysis interfaces with biological experiments.

**Prediction of exons from genomic databases**

Until now, identification and characterization of proteins was difficult for a variety of reasons. First, the various scenarios in which mass spectrometric analysis interfaces with biological experiments. For example, inherent in the identification of a phosphorylation site in a peptide is the information that this protein was phosphorylated by treatment of a specific cell line, by a specified growth factor for a given time. See Fig. 1 for a representation of the various scenarios in which mass spectrometric analysis interfaces with biological experiments.

**Box 1. Major primary sequence databases**

- **Nucleotide sequence databases of known/predicted proteins (e.g. GenBank and EMBL)**
  - Contain nucleotide sequences of all cloned and characterized genes. In addition, these databases contain cDNA sequences corresponding to proteins that have not yet been characterized (generally deposited by cDNA sequencing consortia). A recent inclusion is proteins that are 'totally hypothetical' because they are generated by artificially 'stitching together' the exons that are predicted from the genomic DNA by genefinding programs (this information must be used with extreme caution). GenBank includes 'finished' high-throughput genomic sequences in this category.

- **Protein sequences of known proteins (e.g. SWISS-PROT)**
  - Contain protein sequences of all cloned and characterized proteins. In addition, they contain translations of open reading frames of cDNAs that are predicted to code for proteins. The level of annotation of these depends upon the particular database being used. Curated databases created by human reading of the literature and subsequent entry into a database also exist (i.e. the yeast protein database; YPD).

- **Database of expressed sequence tags (ESTs) (e.g. dbEST)**
  - Contain single-pass nucleotide sequence reads from 5' and 3' ends of cDNA inserts.

- **Genomic databases (e.g. high-throughput genomic sequence; HTGS)**
  - Contain nucleotide sequences from genomic DNA obtained by a variety of strategies. HTGS contains 'unfinished' genomic sequence deposited by major sequencing centers.

**Box 2. Limitations of EST databases**

- **Short stretches of cDNAs**
  - They might not contain coding sequences and generally do not contain the full coding sequence.

- **Sequencing errors**
  - This limits the use of ESTs in establishing the reading frame that codes for the protein. Often we have found that the correct protein sequence must be obtained by 'stitching together' the putative translation products from all three reading frames of the EST.

- **Untranslated regions**
  - Because the ESTs are biased towards the 3' end of the gene, they might only contain the 3' UTR, precluding their use in searches based on protein homology.

- **Abundance**
  - The occurrence of most ESTs correlates somewhat with the abundance of mRNA level of the given gene. Thus transcripts of medium-to-low abundance might not be present in the EST databases.

- **Difficulty in alignment**
  - It can be difficult to align ESTs derived from the same gene due to alternative splicing or gaps in the sequence due to the fact that only the ends of cDNAs are sequenced. This is one of the reasons why the number of unique EST clusters might represent fewer genes than the number of clusters.
**Box 3. Limitations of genomic databases**

- **Interrupted coding regions**
  Because the coding exons in most eukaryotes are interrupted by introns, the prediction of coding regions is not straightforward. Human genes contain the most introns and the largest introns of most mammals studied, making precise gene predictions especially difficult.

- **No reading frame**
  Because the intron–exon junctions do not have any relation to codons, there is no continuous open reading frame as is seen in cDNAs. This is because splitting of exons by introns does not preserve reading frames.

- **Non-directional**
  The strand of genomic DNA that codes for a gene is not fixed. Genes can be transcribed from both strands in opposite directions.

- **Repeat elements**
  The presence of repeated elements can create serious problems in alignment and analysis. Repetitive sequences are excluded when an 'effective size' of the human genome is calculated.

Edman sequencing, which was commonly used for identification, required large amounts of purified protein [Science's Signal Transduction Knowledge Environment (STKE): http://www.stke.org/cgi/content/full/OC_sigtrans;2000/37/pl1]. Second, even when such amounts were available, the relevant proteins could often not be identified because of a blocked N terminus. Mass spectrometry does not suffer from these drawbacks – it is a sensitive method for identifying proteins (femtomole levels of proteins from gels)\(^1\), can analyze mixtures (therefore, the protein does not have to be pure) and is not affected by a blocked N terminus.

Spots excised from 2D gels, bands excised from 1D gels or a complex mixture of proteins in solution, can be digested by proteases such as trypsin and the peptides sequenced by tandem mass spectrometry. These peptide sequences can then be used to search the genomic databases\(^2\) to identify coding regions. Although ESTs have been used successfully to identify novel proteins\(^3\), \(1,13\), it was found that the use of EST data alone does not improve overall gene prediction\(^8,15\). Some of the factors that contributed to this were problems with alignment, the occurrence of paralogs and the low quality of EST sequences. In addition, ESTs led to an increased prediction of smaller genes, thereby reducing specificity\(^15\). However, when any data based on proteins were included, a better prediction accuracy rate was obtained\(^15\). This observation makes mass spectrometry-derived peptide sequence data even more valuable as a complement to bioinformatics tools for gene prediction. Because different gene prediction programs predict different exon–intron structures, such peptide sequences will help establish which ones are ‘real’. Of course, this method cannot normally be used to predict every exon, as it is difficult to achieve 100% coverage for every protein by mass spectrometry at the low levels usually available in biological experiments. The major reasons for this are: (i) some of the peptides derived from a given protein might not ionize well in the mass spectrometer and thus result in very small peaks; (ii) the peptides might be too large or too small for analysis; or (iii) the peptides might give rise to complex fragmentation spectra that are hard to interpret. Nevertheless, by using two different enzymes, close to 100% of the protein primary sequence can often be verified if sufficient material is present. Besides prediction of coding exons, it is possible to find alternatively spliced variants at the protein level based on peptide sequences that ‘span’ different exons in the genome. Several instances showing how mass spectrometry-derived data can be used to annotate genomic sequence are shown in Fig. 2.

For genomic annotation, we propose to use a special format of mass spectrometry in which crude protein mixtures are first reduced to peptides and the peptides then fed into the mass spectrometer after 1- or 2D chromatography\(^16,19\). Thousands of peptides can be sequenced in this approach in a fully automated fashion and with high sensitivity. We then suggest using database search software that can directly find the sequenced peptide in genomic databases, without the need for translation into the six possible reading frames (B. Küster et al., unpublished). This analysis can be done on relatively large amounts of starting material, enabling the analysis of low-copy-number proteins.

**Low-molecular-weight proteins**

A major subset of proteins that are unlikely to be identified from the genomic sequence solely by gene prediction are genes that encode proteins containing less than 100 amino acids\(^20\). This is because gene-prediction algorithms deliberately avoid prediction of such short genes by having a minimum gene size of ~100 amino acids. If such a limit is not imposed, then the number of false positives (i.e. number of potential coding regions that are actually noncoding) will increase sharply. As a result, these proteins will have to be identified by alternative methods. A major fraction of these small polypeptides are likely to be cytokines and growth factors that are secreted by cells – such proteins can be enriched from supernatants of cell lines or obtained from body fluids such as serum, urine or cerebrospinal fluid. Thus, direct sequencing of small secreted proteins by mass spectrometry will allow the discovery of small open reading frames (ORFs) in the genome. Another method is to use cDNA-based techniques to specifically isolate secreted molecules, such as the signal-sequence trap method\(^20\) or other generic methods that isolate less abundant cDNAs such as serial analysis of gene expression (SAGE)\(^22\).
Box 4. Analysis of proteins by mass spectrometry

- Peptide sequencing by tandem mass spectrometry
Proteins are generally digested by trypsin to generate peptides that are then sequenced by mass spectrometry. The method used to sequence peptides is termed tandem mass spectrometry, in which the peptide to be sequenced is first selected from the entire peptide mixture and then fragmented by collision with an inert gas to obtain a partial or complete sequence.

- Database searching using mass spectrometry-derived data
Mass spectrometry of tryptic digests by matrix-assisted laser desorption/ionization time-of-flight (MALDI-TOF) provides a highly accurate list of peptide masses. This list of peptide masses is used to search a theoretical tryptic digest of all the proteins in the databases to identify the protein.

The data obtained by tandem mass spectrometry (MS/MS) generally comprise the partial sequence of a peptide. This partial sequence is referred to as a ‘peptide sequence tag’. It generally contains enough information to identify the protein from databases as large as the human genomic database. If a peptide match for a given peptide sequence tag is not found, it might mean that the peptide belongs to an unknown protein or that the peptide is modified (see below).

- Quantitation of protein levels
Proteins or peptides are difficult to quantify in any mass spectrometer. However, if two samples are being compared, relative quantitation can be obtained by modifying the peptides in both samples (such as by adding a ‘tag’ of known molecular mass). One of the samples is modified with a naturally occurring isotope and the other with a stable isotope analog, such as deuterium-labeled tag, followed by mixing of the samples before analysis. In this case, the corresponding peptides derived from the two samples can be identified easily because they have a constant difference in mass, which is the mass of the stable isotope substitution. The difference in peak intensities then correlates directly with the abundance of the protein levels.

- Identifying phosphopeptides and phosphorylation sites
The peptides containing phosphates can be identified either by observing a change in molecular weight upon treatment with a phosphatase, or in a mass spectrometer, by looking for a reporter ion that is characteristic of phosphate groups (parent or precursor ion scanning).

Identification of phosphorylation sites can be done in several ways by mass spectrometry. In a triple-quadrupole or quadrupole time-of-flight mass spectrometer, the location of phosphoserines or phosphothreonines can be identified in the positive mode if the mass difference between two successive fragments corresponds to phosphorylation (addition of HPO₃ or 80 Da) or to phosphorylated threonine residues (101 + 80 Da). These residues can also be identified by a loss of 98 Da resulting from a β-elimination reaction. Phosphothreonines are generally more stable, do not undergo the β-elimination reaction and are thus found only as a difference of 163 + 80 Da between two successive fragments. The masses of serine, threonine and tyrosine residues are 87, 101 and 163 Da, respectively. Fragmentation to obtain sequence can also be done in an ion-trap mass spectrometer, although the fragmentation pattern is slightly different from that described above.

- Other post-translational modifications
All modifications cause a shift in mass and therefore, in principle, are detectable by mass spectrometry. For example, the N terminus of a protein is indicated by an acetylated peptide – in this case, the N-terminal amino acid contained within the peptide is 42 Da greater than expected. Similarly, myristoylation causes an increase by 210 Da, and GalNac or GlcNAc sugars increase the mass of the corresponding peptide by 203 Da.

Initiation codon
The translational start sites in mRNAs are not well defined. When the mRNA is translated, the ribosomes generally choose the first set of nucleotides encoding a methionine (AUG). The sequences around these AUG sequences have been proposed to affect the efficiency of translation from a particular potential start site – these sequences have been formulated into a consensus sequence for translation initiation (Kozak’s consensus). However, most AUGs (ATGs in cDNA) are not in complete, or sometimes any, agreement with the consensus Kozak sequence. This means that determination of the initiator ATG that will be translated for a given gene might have to be done experimentally or other methods might have to be used. This could involve direct sequencing of the N terminus of the protein, prediction of initiator methionine by homology to known proteins, or guessed by the fact that there are multiple stop codons or no other ATGs upstream of a suspected initiator ATG. This implies that either a longer protein would be predicted, or conversely, a shorter version of a protein would be predicted when, in reality, an upstream ATG is being used.

Using mass spectrometry, it is possible to obtain peptides that are acetylated at the N terminus – this
establishes it as an N-terminal peptide. Another modification that can distinguish an N-terminal peptide from other internal peptides is a fairly frequent cleavage by aminopeptidases of the initiator methionine. In the post-genomic era, one obvious advantage of recognizing the initiator methionine is that the translational start of a gene can then be deduced from the corresponding entry in the genome database.

**Signal peptides**

Signal peptides are a hydrophobic stretch of amino acids that are found at the N termini of most membrane-bound receptors as well as soluble ligands. Although computer programs can predict roughly where the cleavage of this signal sequence can occur, the exact site of cleavage must be determined experimentally. Most experiments in biological mass spectrometry involve cleavage of the protein by proteases such as trypsin before analysis. Owing to the cleavage specificity of trypsin, this results in peptides that contain an arginine or a lysine residue at their C termini. This implies that when the sequence of a peptide is determined, the residue in the protein that immediately precedes the peptide must be an arginine or a lysine. An instance where this does not occur is when the peptide is derived from the N terminus of the protein. In the case of secreted proteins, the signal peptide is cleaved and the 'mature' protein is found in the extracellular environment. Therefore, when a secreted protein is being analyzed and a peptide that corresponds to the N terminus is found, the site of cleavage is automatically assigned. This information is especially critical when recombinant growth factors are manufactured for human use in certain heterologous expression systems, because inclusion of amino acids that are actually cleaved in the mature protein might elicit an immune response.

**Prediction of correct reading frame and stop codon**

If a very short EST sequence is present in databases, it is not immediately clear whether it contains any open reading frame (coding region) and, if it does, what that reading frame is. A peptide sequence obtained by mass spectrometry is found to correspond to an EST sequence, it might be used to assign the correct reading frame to that clone. Further, as explained above, a tryptic peptide sequence that does not end in an arginine or lysine suggests that it is

---

**Fig. 1.** Overview of various types of experiments that can be coupled to mass spectrometric analysis. The data that can be used to annotate databases include information on the 'upfront' experiment (such as growth factor used or the antibody used in the immunoprecipitation step), as well as the data obtained by mass spectrometry. Abbreviation: TGF-β, transforming growth factor β.
Fig. 2. Outputs of gene prediction programs compared with a real protein. The figure shows how mass spectrometry-derived data can be used for annotation of genomic databases. Any tryptic peptide whose sequence matches a genomic region localizes a coding exon. In addition, some peptides might be sequenced that ‘span’ two exons in the genome, thereby predicting two exons adjacent to each other (exons 6 and 7). A peptide with an acetylated N terminus or corresponding to a cleaved methionine indicates the N terminus of the protein. Similarly, a peptide that does not contain an arginine or lysine at its C terminus might indicate the C terminus of the protein.

Post-translational modifications

The majority of proteins in cells undergo post-translational modifications. Extracellular and transmembrane proteins are frequently glycosylated, whereas intracellular signaling molecules can be basally or inducibly phosphorylated. It is thought that at any one time approximately one-third of all proteins in eukaryotic cells are phosphorylated. Although it is more difficult to detect and analyze post-translational modification-containing peptides, post-translational modifications are frequently discovered during large-scale proteomics experiments (reviewed in Ref. 30). At the peptide level, a peptide can be phosphorylated, acetylated or glycosylated (Box 4). However, it is more difficult to localize the exact residue that is modified, and in many instances it might not be possible at all. Nevertheless, knowledge of the relevant peptide carrying the modification might be adequate for the biologist even if the exact residue of modification is not located. Because mass spectrometry provides information about the exact mass of the post-translational modification in addition to its sequence, such information could be annotated as such in the databases for further use by the research community.

Expression data

When a report on a novel gene is published, there is usually some information on tissue expression by northern blotting to detect mRNA. Less commonly, the cell lines that express that protein are described. For investigators who are interested in studying that protein, it is often difficult to find a resource listing cell lines that express the protein. Because the samples for many mass spectrometry-based identification experiments are derived from a cell line, once the identity of the protein is determined, the entries for these proteins could be annotated in sequence databases giving the cell line of origin as the source of protein. Interesting avenues of investigation might be initiated in light of such expression information. For example, if a cytokine receptor known to be specifically expressed in hematopoietic cells is found in fibroblasts, its role in the newly discovered environment can then be tested. Newer technologies can quantify protein expression between two states using stable isotope incorporation into one of the states. Peak ratios between peptides with and without the isotope tag accurately quantify relative levels. In a special twist on this technique, using isotope-coded affinity tags (ICATs), hundreds of proteins can be quantified in a single experiment using only the cysteine-containing peptides. This quantitation is an improvement compared with mRNA data because mRNA levels have been found to correlate poorly with protein expression and might not be predictive of protein levels. Analogous to the data derived from mRNA expression patterns using microarrays that are made publicly available, protein expression data could be linked to the protein sequence itself (see below). Finally, because a minority of pseudogenes can be transcribed, if a protein is detected for a gene, then it can be presumed that it is not a pseudogene.

Protein localization

Proteins, especially novel ones, are often assigned subcellular localization based on homology to other proteins or by sophisticated computer programs. However, these predictions are generally not reliable and therefore the localization of a given protein must be confirmed by experimental methods. An interesting example is the S100 family of calcium-binding proteins that are involved in regulation of a plethora of intracellular activities such as protein phosphorylation, enzyme activities, cell proliferation and differentiation. Based on their sequence, they are not predicted to be secreted molecules. However, some members of this family were discovered as secreted proteins, either as part of proteomic approaches to study secreted molecules or as components of saliva. Similarly, proteins that are predicted to be nuclear based on the presence of nuclear localization signals might turn out to be cytosolic and vice versa. Mass spectrometric identification of specific fractions can be used to assist in annotating proteins localized to organelles such as mitochondria or the chloroplast.
Annotation with protein data - but how?

The submission of sequences of newly identified genes to databases is now routine. Standard web-based forms are to be filled out with the fields clearly defined by the agency that maintains the corresponding database. However, protein data cannot be submitted by users of GenBank, which is surprising given the amount of information being generated directly from proteins. This means that protein entries that are in the databases are largely "inferred" from the cDNA or genomic sequences by conceptual translation or prediction. GenBank currently only allows the submitter of a given sequence to modify the data submitted. There is hardly any provision for third-party correction, let alone third-party annotation. Therefore, the data acquired using mass spectrometry (or by other proteomic methods) cannot be added to the sequence databases under the current circumstances. Data derived from analysis of proteins are dependent upon several factors. For example, although identification of a secreted protein from a cell line could be added as a simple annotation of the protein, modifications such as phosphorylation would have to specify the exact experimental conditions (such as growth factor treatment being used). Besides, a vast amount of data can be accumulated on every protein. We favor an open system of annotation whereby the annotation is performed directly by the experimenter, as in a distributed annotation system being developed by Lincoln Stein (http://stein.cshl.org/das/). In this system, a single server is designated as a reference server that contains primary information about the DNA and protein sequences as well as authorship information. Several websites can then serve as annotation servers that provide additional information on the retrieved sequence. In this scenario, no attempt is made to resolve contradictions that exist between various annotation servers. The uniqueness of this system lies in the fact that the end-user can decide for him/herself which data to use. This would allow for true integration of data, unlike pointers to a weblink that contains related information on a particular sequence. One of the ways in which this could be accomplished would be to use systems similar to the software that is used to share music files over the internet. To illustrate this example, the popular Napster software has a central database and server that allows the user base to link to information that still resides on the hard disks of a large user community. In a more decentralized model, the Gnutella software disperses with the central server and connects the user community directly with each other. Applying either of these models to biology, the basic structure would be given by the genome and the gene names but massive amounts of information could be available dispersed over the internet. Obviously, highly sophisticated software would be needed to retrieve and integrate the disparate pieces of information.

Conclusions and outlook

It is becoming clear that protein-derived data is essential for annotation of nucleic acid and protein sequence databases. Mass spectrometry as the proteomic tool of choice is poised to contribute greatly to the wealth of these databases. Besides the advantage of high-throughput analysis, mass spectrometry provides experimental data and is thus a perfect complement to other annotation efforts. An international initiative and collaboration is, however, critical for the long-term viability of any open annotation system. We envision that, like most other web-based initiatives (electronic publication being one of them), such a system of annotation will become popular and indispensable in the not too distant future.

References

5 Claverie, J. M. (2000) Do we need a huge new central to annotate the human genome? Nature 403, 12
Life-or-death decisions by the Bcl-2 protein family

Jerry M. Adams and Suzanne Cory

In response to intracellular damage and certain physiological cues, cells enter the suicide program termed apoptosis, executed by proteases called caspases. Commitment to apoptosis is typically governed by opposing factions of the Bcl-2 family of cytoplasmic proteins. Initiation of the proapoptotic cascade requires assembly of certain caspase precursors on a scaffold protein, and the Bcl-2 family determines whether this complex can form. Its pro-survival members can act by sequestering the scaffold protein and/or by preventing the release of apoptogenic molecules from organelles such as mitochondria. Pro-apoptotic family members act as sentinels for cellular damage: cytotoxic signals induce their translocation to the organelles where they bind to their pro-survival relatives, promote organelle damage and trigger apoptosis.

Apoptosis, the stereotypic program of cellular suicide, removes unwanted cells throughout life, and its disrupted regulation is implicated in disorders ranging from cancer and autoimmune diseases to degenerative syndromes. The Bcl-2 family of cytoplasmic proteins plays a central regulatory role. Its interacting pro- and anti-apoptotic members integrate diverse upstream survival and distress signals to determine whether the cellular death warrant is issued. Many of the key players have been identified, and the spotlight is now on the stage where their ‘dance of death’ commences: the surface of organelles such as the mitochondria where Bcl-2 family members either reside or congregate during apoptosis. The hotly debated and still unresolved issue of how their control is exerted is the focus of this and other recent reviews.

The central pathway to death
The Bcl-2 family regulates an ancient path to cell death (Fig. 1), found in organisms as diverse as mammals, nematodes and fruit flies. The route culminates in the scission of critical target proteins by proteases of the caspase group, but only after traversing critical checkpoints. To preclude unscheduled cell suicide, each caspase is synthesized as a minimally active precursor and, its processing, at sites of caspase cleavage. The effector caspases are processed by ‘upstream’ caspases, but the apical caspase that sets up the execution must process itself. The autocatalysis requires multimerization, aided by an adaptor or scaffold protein such as the nematode protein CED-4, which primes autoactivation of caspase CED-3, or the mammalian CED-4 homolog Apaf-1, which primes autoactivation of procaspase-9 (Refs 4, 6, 7). The Bcl-2 family determines whether or not the multimeric scaffold/caspase complex, often termed an ‘apoptosome’, can assemble. Anti-apoptotic members such as Bcl-2 and its nematode counterpart CED-9 prevent apoptosome formation, but their life-saving activity can be foiled by other relatives such as Bim or EGL-1 (see below).

Caspases can also be controlled downstream of Bcl-2 (Fig. 1). The IAP (inhibitor of apoptosis) proteins appear to directly block caspase activity.