

## Protein Identification and Analysis Tools on the ExPASy Server

Elisabeth Gasteiger, Christine Hoogland, Alexandre Gattiker, Séverine Duvaud, Marc R. Wilkins, Ron D. Appel, and Amos Bairoch

### 1. Introduction

Protein identification and analysis software performs a central role in the investigation of proteins from two-dimensional (2-D) gels and mass spectrometry. For protein identification, the user matches certain empirically acquired information against a protein database to define a protein as already known or as novel. For protein analysis, information in protein databases can be used to predict certain properties about a protein, which can be useful for its empirical investigation. The two processes are thus complementary. Although there are numerous programs available for those applications, we have developed a set of original tools with a few main goals in mind. Specifically, these are:

1. To utilize the extensive annotation available in the Swiss-Prot database (*1*) wherever possible, in particular the position-specific annotation in the Swiss-Prot feature tables to take into account posttranslational modifications and protein processing.
2. To develop tools specifically, but not exclusively, applicable to proteins prepared by two-dimensional gel electrophoresis and peptide mass fingerprinting experiments.
3. To make all tools available on the World-Wide Web (WWW), and freely usable by the scientific community.

In this chapter we give details about protein identification and analysis software that is available through the ExPASy World Wide Web server (*2*).

Analysis tools include Compute pI/Mw, a tool for predicting protein isoelectric point (pI) and molecular weight (Mw); ProtParam, to calculate various physicochemical parameters; PeptideMass, a tool for theoretically cleaving proteins and calculating the masses of their peptides and any known cellular or artifactual posttranslational modifications; PeptideCutter, to predict cleavage sites of proteases or chemicals in protein sequences; ProtScale, for amino acid scale representation, such as hydrophobicity plots.

Protein identification tools include TagIdent, a tool that lists proteins within a user-specified pI and Mw region, and allows proteins to be identified through the use of short “sequence tags” up to six amino acids long; AACompIdent, a program that identifies proteins by virtue of their amino acid (AA) compositions, sequence tags, pI, and Mw; AACompSim, a program that matches the theoretical AA composition of proteins against the Swiss-Prot database to find similar proteins; MultiIdent, a combination of

other tools mentioned above that accepts multiple data types to achieve identification, including protein pI, Mw, species of interest, AA composition, sequence tag, and peptide masses; and Aldente, a powerful peptide mass fingerprinting identification (PMF) tool.

Protein characterization tools in the context of PMF experiments include FindMod, to predict posttranslational modifications and single-amino acid substitutions; GlycoMod, a tool to predict the possible compositions for glycan structures, or compositions of glycans attached to glycoproteins; FindPept, to predict peptides resulting from unspecific proteolytic cleavage, protease autolysis, and keratin contaminants; and BioGraph to visualize the results of the ExPASy identification and characterization tools.

The tools described here are accessible through the ExPASy WWW server, from the tools page, <http://www.expasy.org/tools/> (see **Fig. 1**). In addition to the tools maintained by the ExPASy team, this page contains links to many analysis and prediction programs provided on Web sites all over the world. The “local” ExPASy tools can be distinguished by the small ExPASy logo preceding their name. They are continually under development and thus may change with time. We document new features of tools in the “What’s new on ExPASy” Web page at <http://www.expasy.org/history.html>. Feedback and suggestions from users of the tools is very much appreciated and can be sent by e-mail to [tools@expasy.org](mailto:tools@expasy.org). Detailed documentation for each of the programs is available from the Web site.

## 2. The Swiss-Prot Database

The identification tools described below all work directly and exclusively with the Swiss-Prot protein knowledgebase and its automatically annotated supplement TrEMBL (*1*). Since the maintainers of Swiss-Prot and TrEMBL (the Swiss Institute of Bioinformatics and the European Bioinformatics Institute) joined forces with the PIR group at Georgetown University to form the UniProt consortium (*3*), Swiss-Prot and TrEMBL are also known as the “UniProt Knowledgebase.”

In order to make the most of the tools, it is helpful to understand a number of concepts applied in Swiss-Prot and TrEMBL. The Swiss-Prot user manual (<http://www.expasy.org/sprot/userman.html>) provides a detailed description of the database format and scope, and complements the information in this section.

### 2.1. Annotation Quality


Swiss-Prot is known for its extensive manual annotation, whereas the vast majority of TrEMBL entries are unannotated or automatically annotated. This has a number of implications for the user of proteomics tools.

Identification results usually show the description (DE) line of protein entries matching the experimental data (sometimes, this description may be truncated if it is longer than the space available in the output tables). Whereas all Swiss-Prot description lines are manually created and verified to list the most common name and synonyms used for a protein, enforcing standardized nomenclature, TrEMBL DE lines usually consist of the phrase typed in by the submitter of the underlying nucleotide coding sequence, or of a protein name inferred by automatic annotation procedures. As far as keywords (KW lines) and feature tables (FT lines) are concerned, the situation is similar: all


[ExpPASy Home page](#)
[Site Map](#)
[Search ExpPASy](#)
[Contact us](#)
[Swiss-Prot](#)
[PROSITE](#)
[SWISS-2DPAGE](#)

Hosted by [SIB Switzerland](#)
Mirror sites: [Australia](#) [Bolivia](#) [Canada](#) [China](#) [Korea](#) [Taiwan](#) [USA](#)

for



## ExpPASy Proteomics tools

The tools marked by  are local to the ExpPASy server. The remaining tools are developed and hosted on other servers.

**Protein identification and characterization**

- [!\[\]\(b4572a044582c68c9e6e6b6b9b95c325\_img.jpg\) AACompldent](#) - Identify a protein by its amino acid composition
- [!\[\]\(a4848a7f290c3fd14bf2276a5b09747a\_img.jpg\) AACompSim](#) - Compare the amino acid composition of a Swiss-Prot entry with all other entries
- [!\[\]\(c1b926fcee63cdc7f15c603966ea3d76\_img.jpg\) Multident](#) - Identify proteins with *pI*, *Mw*, amino acid composition, sequence tag and peptide mass fingerprinting data
- [!\[\]\(8bca340a5031aa30dac36202fe939cc1\_img.jpg\) PeptIdent](#) - Identify proteins with peptide mass fingerprinting data, *pI* and *Mw* Experimentally measured, user-specified peptide masses are compared with the theoretical peptides calculated for all proteins in Swiss-Prot, making extensive use of database annotations
- [!\[\]\(3e3e4303b0632a7f1c9ba6d71d06a491\_img.jpg\) TagIdent](#) - Identify proteins with *pI*, *Mw* and sequence tag, or generate a list of proteins close to a given *pI* and *Mw*
- [!\[\]\(92ce96eaea0d7284690928c496458da0\_img.jpg\) FindMod](#) - Predict potential protein post-translational modifications and potential single amino acid substitutions in peptides. Experimentally measured peptide masses are compared with the theoretical peptides calculated from a specified Swiss-Prot entry or from a user-entered sequence, and mass differences are used to better characterize the protein of interest.
- [!\[\]\(247994d2ad855c6281a2ea7e46db0b1d\_img.jpg\) GlycoMod](#) - Predict possible oligosaccharide structures that occur on proteins from their experimentally determined masses (can be used for free or derivatized oligosaccharides and for glycopeptides)
- [!\[\]\(84605d36fdd34af39a387412552dd1b3\_img.jpg\) GlycanMass](#) - Calculate the mass of an oligosaccharide structure
- [!\[\]\(fd013da321a40c586e1c377ff7c7451e\_img.jpg\) FindPept](#) - Identify peptides that result from unspecific cleavage of proteins from their experimental masses, taking into account artefactual chemical modifications, post-translational modifications (PTM) and protease autolytic cleavage
- [!\[\]\(35793e973846d66cd30a9059235083d1\_img.jpg\) PeptideMass](#) - Calculate masses of peptides and their post-translational modifications for a Swiss-Prot or TrEMBL entry or for a user sequence
- [!\[\]\(affe54e245ead3d9a054c9a6dfac0526\_img.jpg\) PeptideCutter](#) - Predicts potential protease and cleavage sites and sites cleaved by chemicals in a given protein sequence
- [PepMAPPER](#) - Peptide mass fingerprinting tool from UMIST, UK
- [Mascot](#) - Peptide mass fingerprint, sequence query and MS/MS ion search from Matrix Science Ltd., London
- [PepSea](#) - Protein identification by peptide mapping or peptide sequencing from Protana, Denmark
- [PeptideSearch](#) - Peptide mass fingerprint tool from EMBL Heidelberg
- [ProteinProspector](#) - A variety of tools from UCSF (MS-Fit, MS-Tag, MS-Digest, etc.) for mining sequence databases in conjunction with mass spectrometry experiments [Mirrors at [UCL-Ludwig](#), UK / [Ludwig Institute Melbourne](#) (Australia)]
- [PROWL](#) - Protein chemistry and mass spectrometry resource from Rockefeller and NY Universities [or from [Genomic Solutions](#)]
- [PFMUTS](#) - Shows the possible single and double mutations of a peptide fragment from MALDI peptide mass fingerprinting

Fig. 1. The ExpPASy tools page, <http://www.expasy.org/tools/>. All underlined text represents hypertext links, which, when selected with a computer mouse, take the user to the corresponding page for the chosen tool. The tools whose names are preceded by a small ExpPASy logo are maintained by the ExpPASy team; all other links lead to external servers.

Swiss-Prot entries are assigned a comprehensive list of keywords as part of the manual annotation process; TrEMBL, however, has very few, but automatically assigned keywords. Even more importantly for identification tools, feature tables, which contain information about known position-specific events in the sequence, such as posttranslational modifications and processing, or sequence variants, are very complete in Swiss-Prot and scarce in TrEMBL. Finally, the sequences themselves are carefully checked in Swiss-Prot and much less likely to contain errors (e.g., frameshifts) than in TrEMBL.

## **2.2. Alternative Splicing**

Many proteins exist in more than one isoform, one cause of which is alternative (differential) splicing. Splice isoforms may differ considerably from one another, with potentially less than 50% sequence similarity between isoforms. In the Swiss-Prot database, one sequence (usually that of the longest isoform) is displayed for each protein. Known variations of this sequence are recorded in the feature table (using the VARSPLIC key), together with the name(s) of the isoform(s) in which each variant occurs. Unique and stable identifiers have been assigned to all alternative splice isoforms, and the sequences of these isoforms are distributed with Swiss-Prot. The unique splice isoform identifiers (of the form P19491-2, where P19491 is the accession number of the “original” Swiss-Prot entry, and “-2” denotes the second annotated splice isoform in that entry) can be submitted to the ExPASy analysis tools. For identification tools, the databases that constitute the search space include the alternative splice isoform sequences annotated in Swiss-Prot and TrEMBL in addition to the canonical sequences contained in those databases. For each isoform, the ExPASy server provides a page displaying the complete sequence of that isoform, with direct links to submission forms of the analysis tools described in this chapter.

## **2.3. Posttranslational Modifications**

Posttranslational modification annotation (4,5) in Swiss-Prot, particularly in the feature table, is currently undergoing a major overhaul and standardization process. Controlled vocabularies are introduced for the feature descriptions corresponding to the feature keys MOD\_RES (used for processes like phosphorylation, acetylation, sulfation, and so on), LIPID (for palmitoylation, farnesylation, geranyl-geranylation, and so on), CROSSLNK (for thioether, thioester, and other bonds) and DISULFID. This facilitates the task of reliably parsing out information about posttranslational modification events and applying the corresponding mass corrections to affected peptides. A database of modifications, containing the biological mechanism, and the conditions for occurrence (taxonomy, type of amino acid, position within the sequence) for each stored modification, is being built and will be made available via ExPASy, extensively linked to Swiss-Prot entries and proteomics tools.

It should be noted that while mass calculations can take into account known posttranslational modifications if they consist in the addition of simple groups (e.g., phosphorylation, acetylation), the algorithm used for the calculation of isoelectric points (and used by many of the tools described later) does not.

## **2.4. Swiss-Prot-Related Conventions for the ExPASy Tools**

Unless otherwise stated, the ExPASy tools use Swiss-Prot annotations to process polypeptides to their mature forms before using them for calculations or protein identification procedures. Thus, protein signal sequences and propeptides are removed where found, and precursor molecules processed into their resulting chains.

The characterization and analysis tools described in this chapter all accept Swiss-Prot/TrEMBL identifiers (including splice isoform identifiers) as well as raw sequences as input.

When entering sequence data into text boxes for the tools, note that any spaces, newline (return) characters, and numbers will be ignored. This allows sequences in other formats, for example GCG format, to be used directly in the programs without first removing any numbering or other formatting. When using FASTA format, the first (header) line should be removed before submitting to the server.

The numbering used by the tools for amino acids in protein sequences refers to the Swiss-Prot entry. If proteins are processed to mature forms, the number of the N-terminal amino acid will remain the same as it was in the unprocessed protein sequence.

## **2.5. Stability of Swiss-Prot Entry Names Is Not Guaranteed**

In 2004, the format of Swiss-Prot entry names (ID) will be extended from 4letters/underscore/5letters to at most 5letters/underscore/5letters. We have never claimed that Swiss-Prot IDs are stable, and have always strongly recommended the use of primary accession numbers instead. The months following the publication of this chapter will see a particularly large number of ID changes, as a result of this format change. Here, we identify all Swiss-Prot entries by their ID and AC, but would like to insist that the only identifiers whose stability we can guarantee are the accession numbers.

# **3. Single-Protein Analysis Tools on the ExPASy Server**

## **3.1. Compute pI/Mw Tool**

This tool ([http://www.expasy.org/tools/pi\\_tool.html](http://www.expasy.org/tools/pi_tool.html)) calculates the estimated pI and Mw of a specified Swiss-Prot/TrEMBL entry or a user-entered AA sequence (*see* **Notes 1, 2**). These parameters are useful if you want to know the approximate region of a 2-D gel where a protein may be found.

To use the program, enter one or more Swiss-Prot/TrEMBL identification names (e.g., LACB\_BOVIN) or accession numbers (e.g., P02754) into the text field, and select the “click here to compute pI/Mw” button. If one entry is specified, you will be asked to specify the protein’s domain of interest for which the pI and mass should be computed. The domain can be selected from the hypertext list of features shown, if any, or by numerically specifying the domain start and end points.

If more than one Swiss-Prot/TrEMBL identification name is entered, all proteins will automatically be processed to their mature forms, and pI and Mw values calculated for the resulting chains or peptides. If only fragments of the protein of interest are available in the database, no result will be given and an error message will be shown to highlight that the pI and mass cannot be returned accurately. Some database entries



have signal sequences or transit peptides of unknown length (e.g., Q00825; ATPI\_ODOSI). In those cases, an average-length signal sequence or transit peptide is removed before the pI and mass computation is done (*see Note 3*). In Swiss-Prot release 42.6 of 28-Nov-2003, the average signal sequence length is 22 amino acids for eukaryotes and viruses, 26 amino acids for prokaryotes and bacteriophages, and 31 for archaeobacteria. Transit peptides have an average length of 57 amino acids in chloroplasts, 34 for mitochondria, 34 for microbodies, and 65 for cyanelles.

If your protein of interest is not in the Swiss-Prot database, you can enter an AA sequence in standard single letter AA code into the text field, and select the “click here to compute pI/Mw” button. The predicted pI and Mw of your sequence will then be displayed. A typical output from the program is shown in **Fig. 2A**.

Alternatively to the verbose html output, the result for a list of Swiss-Prot/TrEMBL entries can also be retrieved in a numerical format, with minimal documentation. A file containing four columns—ID, AC, pI, and Mw—is generated and can be loaded into an external application, such as a spreadsheet program. A typical file output is shown in **Fig. 2B**.

## 3.2. ProtParam Tool

### 3.2.1. Using ProtParam

ProtParam (<http://www.expasy.org/tools/protparam.html>) computes various physico-chemical properties that can be deduced from a protein sequence. No additional information is required about the protein under consideration. The protein can either be specified as a Swiss-Prot/TrEMBL accession number or ID, or in the form of a raw sequence. White space and numbers are ignored. If you provide the accession number of a Swiss-Prot/TrEMBL entry, you will be prompted with an intermediary page that allows you to select the portion of the sequence on which you would like to perform the analysis. The choice includes a selection of mature chains or peptides and domains from the Swiss-Prot feature table (which can be chosen by clicking on the positions), as well as the possibility to enter start and end position in two boxes. By default (i.e., if you leave the two boxes empty) the complete sequence will be analyzed (*see Note 4*).

### 3.2.2. The Calculated Parameters

The parameters computed by ProtParam include the molecular weight, theoretical pI, amino acid composition, atomic composition, extinction coefficient, estimated half-life, instability index, aliphatic index, and grand average of hydropathicity (GRAVY). Molecular weight and theoretical pI are calculated as in Compute pI/Mw. The amino acid and atomic compositions are self-explanatory. All the other parameters will be explained below.

#### 3.2.2.1. EXTINCTION COEFFICIENTS

The extinction coefficient indicates how much light a protein absorbs at a certain wavelength. It is useful to have an estimation of this coefficient for following a protein which a spectrophotometer when purifying it (*see Note 5*)

It has been shown (6) that it is possible to estimate the molar extinction coefficient of a protein from knowledge of its amino acid composition. From the molar extinction coefficient of tyrosine, tryptophan, and cystine (cysteine does not absorb appreciably at wavelengths >260 nm, while cystine does) at a given wavelength, the extinction

**A**

LACB\_BOVIN (P02754)

DE Beta-lactoglobulin precursor (Beta-LG) (Allergen Bos d 5).  
 OS Bos taurus (Bovine).

The parameters have been computed for the following feature:

FT CHAIN 17 178 BETA-LACTOGLOBULIN.

Considered sequence fragment:

```

      1          11          21          31          41          51
      |          |          |          |          |
1    |          |          |          |          |
61  ELKPTPEGDL EILLQKWENG ECAQKKIIAE KTKIPAVFKI DALNENKVLV LDTDYKKYLL 120
121 FCMENSAEPE QSLACQLVR TPEVDDEALE KFDKALKALP MHIRLSFNPT QLEEQCHI

```

Molecular weight: 18281.21

Theoretical pI: 4.83

**B**

ARS1_MOUSE	O54984	4.80	39065.08
ARSA_MOUSE_1	P50428	5.50	52173.26
ARSB_MOUSE	P50429	FRAGMENT	0.00
ARX_MOUSE	O35085	5.19	58504.37
ARY1_MOUSE	P50294	5.10	33713.36
ARY2_MOUSE	P50295	5.63	33701.41
ARY3_MOUSE	P50296	6.07	33685.69
ASAH_MOUSE_1	Q9WV54	6.11	13797.05
ASAH_MOUSE_2	Q9WV54	8.87	29017.27

Fig. 2. (A) Sample output from the Compute pI/Mw tool, where the program was requested to calculate the theoretical pI and Mw for the Swiss-Prot entry LACB\_BOVIN (P02754). Note that the Compute pI/Mw tool shows the sequence of the region of the protein that is under consideration. In this case, the sequence of the mature beta-lactoglobulin is shown, which results when the secretion signal sequence is removed from the precursor polypeptide. (B) Output file sample retrieved from the Compute pI/Mw tool, where the program was requested to calculate the theoretical pI and Mw for a list of Swiss-Prot/TrEMBL entries. Note that the numerical format is minimal, to be exported into an external application. If pI and Mw cannot be computed, a value of “0.00” appears in the Mw column, and the reason for this is displayed in the pI column in the form of a code, the meaning of which is as follows:

FRAGMENT Incomplete CHAIN/PEPTIDE: pI/Mw cannot be computed

UNDEFINED Unknown start- or endpoints: pI/Mw cannot be computed

XXX Sequence contains several consecutive undefined AA: pI/Mw cannot be computed

If a Swiss-Prot/TrEMBL entry has one or more mature chains/peptides documented, this is indicated by “\_1”, “\_2”, etc. appended to the ID. An appended “\_1,” “\_2,” and so on, indicates that the considered sequence is that corresponding to the first, second, and so on, CHAIN or PEPTIDE documented in the feature table.

coefficient of a denatured protein can be computed (*see Note 6*). Two tables are produced by ProtParam, the first one showing the computed values based on the assumption that all cysteine residues appear as half cystines, and the second one assuming that no cysteine appears as half cystine.

#### 3.2.2.2. IN VIVO HALF-LIFE

The half-life is a prediction of the time it takes for half of the amount of protein in a cell to disappear after its synthesis in the cell. The prediction is given for three organisms (human, yeast, and *E. coli*), but it is possible to extrapolate the result to similar organisms. ProtParam estimates the half-life by looking at the N-terminal amino acid of the sequence under investigation (*see Note 7*).

#### 3.2.2.3. INSTABILITY INDEX (II)

The instability index provides an estimate of the stability of your protein in a test tube. It can be predicted as described in **Note 8**. A protein whose instability index is smaller than 40 is predicted as stable; a value above 40 predicts that the protein may be unstable.

#### 3.2.2.4. ALIPHATIC INDEX

The aliphatic index of a protein is defined as the relative volume occupied by aliphatic side chains (alanine, valine, isoleucine, and leucine). It may be regarded as a positive factor for the increase of thermostability of globular proteins. **Note 9** details how the aliphatic index is computed.

#### 3.2.2.5. GRAND AVERAGE OF HYDROPATHY

The grand average of hydrophathy (GRAVY) value for a peptide or protein is calculated as the sum of hydrophathy (7) values of all the amino acids, divided by the number of residues in the sequence.

### 3.3. PeptideMass

The PeptideMass tool (<http://www.expasy.org/tools/peptide-mass.html>) is designed to assist in peptide-mapping experiments, and in the interpretation of peptide-mass fingerprinting (PMF) results and other mass-spectrometry data (8) (*see Note 10*). It cleaves *in silico* a user-specified protein sequence or a mature protein in the Swiss-Prot/TrEMBL databases with an enzyme or reagent of choice, to generate peptides. Masses of the peptides are then calculated and displayed. If a protein from Swiss-Prot has annotations that describe discrete posttranslational modifications (specifically acetylation, amidation, biotinylation, C-mannosylation, formylation, farnesylation,  $\gamma$ -carboxy glutamic acid, geranyl-geranylation, lipoyl groups, N-acyl glycerides, methylation, myristoylation, NAD, O-GlcNAc, palmitoylation, phosphorylation, pyridoxyl phosphate, pyrrolidone carboxylic acid, or sulfation), the masses of these modifications will be considered in peptide mass calculations (*see Note 11*). Post-translational modifications can also be specified along with a user-entered sequence that is not in Swiss-Prot or TrEMBL. Guidelines for the input format of posttranslational modifications (PTMs) are accessible directly from the PeptideMass input form (*see Note 12*). The mass effects of artifactual protein modifications such as the oxidation of methionine or acrylamide adducts on cysteine residues can also be considered. The program can supply warnings where peptide masses may be subject to change from protein isoforms, database conflicts, or mRNA splicing variation.

To use the program, enter one or more Swiss-Prot identification names (e.g., TKN1\_HUMAN) or any Swiss-Prot/TrEMBL accession number (e.g., P20366) into the text field, or enter a protein sequence of interest using the standard one-letter AA code. User-specified sequences should not contain the character X, but can contain the



character *J*, to represent either Ile or Leu, which are of the same mass. The enzyme or reagent to use to theoretically cleave the protein sequence should then be specified, and whether any missed cleavages should be allowed. You can select to exclude masses below a certain threshold (e.g., 500 Daltons) which might be too small to be visible in a mass spectrum. The PeptideMass output will include the portions of the sequence covered by only the fragments that are above that threshold. Special treatment (if any) of cysteine residues or oxidation of methionine should be selected, and whether results are desired as monoisotopic or average masses. Finally, click on the “Perform” button to send data to the program. **Figure 3** shows a typical output of the program PeptideMass, illustrating some of its features.

### 3.4. PeptideCutter

PeptideCutter (<http://www.expasy.org/tools/peptidecutter/>) predicts cleavage sites of proteases or chemicals in a protein sequence. Protease digestion can be useful if one wants to carry out experiments on a portion of a protein, separate the domains in a protein, remove a tag protein when expressing a fusion protein, or make sure that the protein under investigation is not sensitive to endogenous proteases. One or several reagents can be selected from a list of (currently) 33 proteases and chemicals.

The protein sequence can be entered in the form of a Swiss-Prot/TrEMBL accession number, a raw sequence, or a sequence in FASTA format, in one-letter amino acid code. Letters that do not correspond to an amino acid code (*B*, *J*, *O*, *U*, *X*, or *Z*) will cause an error message, and the user is required to correct the input. Please note that only one sequence can be entered at a time.

You have the possibility to select one or a group of enzymes and chemicals. Most of the cleavage rules for individual enzymes were deduced from specificity data summed up by Keil (9), and the rules are listed as part of the PeptideCutter documentation. You can also ask the program to consider only enzymes that cut the sequence a chosen number of times, which may be of particular interest if you have selected a large number of cleavage agents.

For the enzymes trypsin and chymotrypsin, enough experimental data were available to study cleavage by these enzymes at sites different from the otherwise widely used motifs “after K/R but not before P” for trypsin, and “after F/Y/W but not before P” for chymotrypsin. Keil (9) created probability tables for cleavage between all pairs of amino acids in the positions N- and C-terminal to the cleavage site. This more “sophisticated” model is available for use with PeptideCutter, and with this option, the output includes the cleavage probability for any potential site.

For the display of your results, there are three different output options:

1. For every selected enzyme, the number of cleavage sites and their positions are enumerated in a first table, in alphabetical order by enzyme name.
2. The second table (not displayed by default) displays all cleavage sites sequentially in the sequence, one site per line. For each position, the cleaving enzyme, the resulting peptide sequence, the peptide length, and its mass are listed (*see Note 13*). The peptides displayed are calculated based on the assumption that all chosen enzymes are present during digestion. If you want to have a list of peptides resulting from cleavage by only one enzyme or chemical, select only this enzyme and deselect all others.
3. Finally, there is also the possibility to show all results using a map. The entered protein sequence is marked with a “|” above an amino acid when there is a cleavage site between

You have selected [TKN1\\_HUMAN](#) (P20366) from Swiss-Prot:

Protachykinin 1 precursor (PPT) [Contains: Substance P; Neurokinin A (NKA) (Substance K) (Neurodin L); Neuropeptide K (NPK); Neuropeptide gamma; C-terminal flanking peptide]. Signal and propep in positions 1-56 have been removed.

- Peptide [SUBSTANCE P](#) at positions [58 - 68](#) [Theoretical pl: 11.00 / Mw (average mass): 1348.63]

mass	position	#MC	modifications	peptide sequence
1349.6362	58-68	0	<a href="#">AMID: 68</a> 1348.6515	RPKPQQFFGLM

100.0% of sequence covered (you may modify the input parameters to display also peptides < 500 Da):

```

      1          11          21          31          41          51
      |          |          |          |          |
1
61 PQQFFGLM                                     RPK 60

```

- Peptide [NEUROPEPTIDE K](#) at positions [72 - 107](#) [Theoretical pl: 8.40 / Mw (average mass): 3981.54]

mass	position	#MC	modifications	peptide sequence
1211.3653	86-96	0		ALYGHGQISHK
870.0077	100-107	0	<a href="#">AMID: 107</a> 869.0230	TDSFVGLM
864.8841	72-79	0		DADSSIEK
671.8577	80-85	0		QVALLK

91.7% of sequence covered (you may modify the input parameters to display also peptides < 500 Da):

```

      61          71          81          91          101          111
      |          |          |          |          |
61                                     DADSSIEKQ VALLKALYGH GQISHKrhkT DSFVGLM

```

- [NEUROPEPTIDE GAMMA 2ND PART](#) at positions [89 - 107](#) [Theoretical pl: 9.99 / Mw (average mass): 2135.43]

mass	position	#MC	modifications Peptide	peptide sequence
870.0077	100-107	0	<a href="#">AMID: 107</a> 869.0230	TDSFVGLM
863.9511	89-96	0		GHGQISHK

84.2% of sequence covered (you may modify the input parameters to display also peptides < 500 Da):

```

      61          71          81          91          101          111
      |          |          |          |          |
61                                     GH GQISHKrhkT DSFVGLM

```

Fig. 3. Sample output from the PeptideMass tool. The protein selected was TKN1\_HUMAN (P20366), and the program was requested to cleave with trypsin, show all peptides, sort peptides by mass, show all known modifications, use average masses, and display masses as  $(M + H)^+$ . The figures in the “modified mass” column in this case show the predicted masses of peptides known to be amidated. Note that there are three lists of peptides, which correspond to the cleavage of different products known to be created from the same initial polypeptide. Underlined text and numbers represent hypertext links in the output that, if selected, show either the Swiss-Prot entry for the protein (e.g., TKN1\_HUMAN), or the sequence of any portion of a protein specified with numbers (e.g., 58–68), or a relevant section of the online documentation (e.g., modifications). The feature table of the entry P20366 describes three more mature peptides, for which the program output is not shown here.

this amino acid and the neighboring amino acid in the C-terminal direction (i.e., directly on the “right side” of the marked amino acid). The sequence map is displayed in portions of 10 to 60 amino acids. The number of amino acids displayed per line can be modified, which may be particularly useful when printing out the map. If you have selected several enzymes and find the map too overloaded, it is possible to reduce the information and display only the cleavage sites of one enzyme by clicking on its name in the map.

### 3.5. ProtScale

ProtScale (<http://www.expasy.org/tools/protscale.html>) allows computation and representation (in the form of a 2-D plot) of the profile produced by any amino acid scale on a selected protein (*see Note 14*). ProtScale can be used with 50 predefined scales entered from the literature. The scale values for the 20 amino acids, as well as a literature reference, are provided on ExPASy for each of these scales. To generate data for a plot, the protein sequence is scanned with a sliding window of a given size. At each position, the mean scale value of the amino acids within the window is calculated, and that value is plotted for the midpoint of the window.

You can set several parameters that control the computation of a scale profile, such as the window size, the weight variation model, the window edge relative weight value, and scale normalization.

#### 3.5.1. Window Size

The window size is the length of the interval to use for the profile computation, i.e., the number of amino acids examined at a time to determine a point of hydrophobic character. When computing the score for a given residue  $i$ , the amino acids in an interval of the chosen length, centered around residue  $i$ , are considered. In other words, for a window size  $n$ , we use the  $i-(n-1)/2$  neighboring residues on each side of residue  $i$  to compute the score for residue  $i$ . The score for residue  $i$  is the sum of the scale values for these amino acids, optionally weighted according to their position in the window.

One should choose a window that corresponds to the expected size of the structural motif under investigation: A window size of 5 to 7 is appropriate for finding hydrophilic regions that are likely to be exposed on the surface and may potentially be antigenic. Window sizes of 19 or 21 will make hydrophobic, membrane-spanning domains stand out rather clearly (typically  $>1.6$  on the Kyte-Doolittle scale [7]).

#### 3.5.2. Relative Weight of the Window Edges

The central amino acid of the window always has a weight of 100%. By default, the amino acids at the remaining window positions have the same weight, but you can attribute a larger weight (in comparison with the other residues) to the residue at the center of the window by setting the weight value for the residues at the extremities of the interval to a value between 0 and 100%. The decrease in weight between the center and the edges will either be linear or exponential, depending on the setting of the weight variation model option. The ProtScale documentation includes graphic illustrations of the two available models.

#### 3.5.3. Scale Normalization

You can choose whether to use the unmodified selected scale values from the literature or to normalize the values so that they all fit into the range from 0 to 1. Normalization is useful if you want to compare the results of profiles obtained with different scales, and makes plots with a more uniform appearance.

### 3.5.4. Interpreting Results

The method of sliding windows, and hence ProtScale, only provides a raw signal and does not include interpretation of the results in terms of a score. When interpreting the results, one should consider only strong signals. In order to confirm a possible interpretation, one could slightly change the window size, or replace the scale by another similar one (e.g., two different hydrophobicity scales), and ensure that the strong signal is still present.

## 4. Protein Identification and Characterization Tools on ExPASy

### 4.1. TagIdent Tool

The TagIdent tool (<http://www.expasy.org/tools/tagident.html> [10,11]) serves two main purposes. Firstly, it can create lists of proteins from one or more organisms that are within a user-specified pI or Mw range (see **Note 15**). This is useful to find proteins from the database that may be in a region of interest on a 2-D gel. Secondly, the program can identify proteins from 2-D gels by virtue of their estimated pI and Mw, and a short protein “sequence tag” of up to six amino acids. The sequence tag can be derived from protein N-termini, C-termini, or internally, and generated by chemical- or mass-spectrometric sequencing techniques. As sequence tags are highly specific (e.g., there are 160,000 different combinations of four amino acid sequence tags) they represent a form of protein identification that is useful for organisms that are molecularly well defined and have a relatively small number of proteins (e.g., *Escherichia coli* or *Saccharomyces cerevisiae*). Interestingly, we have shown that C-terminal sequence tags are more specific than N-terminal tags (11); however, it remains technically more difficult to generate high-quality C-terminal protein sequence data. Thirdly, the sequence tag can be used, together with a very precise protein mass obtained from mass spectrometry, to identify a protein after peptide fragmentation. One can also specify terms to limit the search to a range of organisms or to a specific organism (species). Additionally, Swiss-Prot keywords can also be used in the identification procedure.

#### 4.1.1. Use of TagIdent to List Proteins in a Defined pI and/or Mw Region

TagIdent can generate a list of proteins in a pI and Mw range of interest, which can be sent to the user by e-mail, if a valid e-mail address is specified, or displayed in the browser window. Queries can usually be dealt with within a few seconds in your browser window. However, if you wish to submit many different queries, you may prefer to receive the results by e-mail, for easier archiving, and also in order not to have to wait for the result of one query to come back before submitting the next one. When many queries are submitted and the results are requested to be sent by e-mail, the server schedules the execution of the different identification tasks in such a way that the server CPU remains easily accessible to other users. If desired, a name can be given to your query, which will appear as the subject of the e-mail message, or at the top of the result page. This is useful for archiving purposes or if many different queries are to be submitted to the program at the same time. You should then specify the pI and Mw regions within which you would like to search (e.g., pI of  $5.5 \pm 0.5$  units and Mw  $20,000 \pm 10\%$ ). If you would like to search using only one of the pI or Mw parameters, you can specify an unrestricted window to cover all possibilities for the other parameter (see **Note 16**). For example, a search where pI is set to  $7.0 \pm 5$  units but where a

Mw window of  $20,000 \pm 10\%$  is used, will return all proteins of sizes 18,000 to 22,000 Da, regardless of their pI. In the search, you can specify one or more terms matching those in the Swiss-Prot OS (species) or OC (classification) lines to limit the search to one organism, or a range of organisms. A document containing a full list of all Swiss-Prot species can be found at <http://www.expasy.org/cgi-bin/specelist>. Thus if you want to investigate proteins exclusively from *S. cerevisiae*, you can specify “cerevisiae.” This is better than specifying “yeast,” a word common to the classification of many yeasts, which includes not only proteins from *Saccharomyces cerevisiae*, but also those from *Candida albicans* and *Schizosaccharomyces pombe*. The same applies for *Homo sapiens*, where “sapiens” will search only for human proteins while “human” will include proteins from human viruses. If you would like to investigate proteins from a broader range of species, it is possible to specify a classification like “mammalia,” which will return all mammalian proteins within the specified pI and Mw region. Use of the word *ALL* will search all species in the database; however, this not recommended, given the size of protein databases, unless all other input data are extremely specific. If desired, searches can also be restricted through use of a Swiss-Prot keyword, such as *Plasmid* or *Alzheimer’s disease*. A document containing a full list of all Swiss-Prot keywords can be found at <http://www.expasy.org/cgi-bin/keywlist.pl>. By clicking on one of the keywords in this list, one obtains the definition of the keyword’s usage in Swiss-Prot, its mapping to GeneOntology terms (GO) (12) (if any), the keyword hierarchy and category, as well as a list of all Swiss-Prot entries annotated with that keyword. Keywords will be used only to restrict searches in Swiss-Prot. Any specified keyword will be ignored for TrEMBL, whose keyword annotation is only partial, and largely created by automated procedures without any manual intervention. Finally, select the “Start TagIdent” button to submit the request to ExPASy.

#### 4.1.2. Use of TagIdent to Identify Proteins From a 2-D Gel

TagIdent can identify proteins by matching sequence tags against proteins in Swiss-Prot from one or more species within a specified pI and Mw range (see **Note 17**). To use TagIdent for identification purposes, first specify the pI and Mw of the protein of interest as estimated from the 2-D gel. Then specify error margins that reflect the known accuracy of these estimates. (See **ref. 11** for an example of how pI and Mw ranges can be defined.) The species and keyword in the database to match against should then be specified (see **Subheading 4.1.1.**), the “Tagging” option selected by clicking in the small box, and the sequence tag entered in single-amino acid code in the “Tag” text box. Note that the sequence tag can contain one or more X to represent any unknown amino acid. Finally, you should specify the source of your protein sequence (N-, C-terminal, or internal), such that the program can show the protein area of interest in the search results. Thus, for example, if you have generated an N-terminal protein sequence tag by Edman degradation, you should request the program to show predicted protein N-termini. Finally, submit the search to the ExPASy server by selecting the “Start TagIdent” button. A typical output is shown in **Fig. 4**.

#### 4.1.3. Interpretation of TagIdent Results for Protein Identification

Accurate identification of proteins with sequence tags relies on all proteins from an organism being in sequence databases. In this manner, if only one protein within a given pI and Mw range is found to contain a certain N-, C-terminal, or internal sequence

```

Search performed in Swiss-Prot with following values:
  pI =          5.97
  delta-pI =    0.50
  Mw =          45098
  delta-Mw =    9019
  OS or OC =    ESCHERICHIA COLI
  KW keyword =   ALL
  Display the N-terminal sequence.
  Tag = MDQT
-----
Scan done on 11-Dec-2003.
Swiss-Prot Release 42.6 of 28-Nov-2003: 139947 entries
-----
462 proteins found in the specified pI/Mw ranges
Results with tagging: 1 found
-----
The number before the sequence indicates the position in the
mature protein where your tag MDQT has been found (first occurrence).
If the protein displayed results from the processing of a
precursor, the position of the tag in the precursor polypeptide
will be given in brackets.
The sequence tag itself is printed in lowercase.
---
DHE4_ECOLI (P00370)
  NADP-specific glutamate dehydrogenase (EC 1.4.1.4) (NADP-GDH).
  pI: 5.98, MW: 48581.37
1   mdqtYSLESFLNHVQKRDPNQTEFAQAVREVMTTLWPFLE...
---
Results without tagging: 461 found
(Printing the N-terminal sequence)
---
AAT_ECOLI (P00509)
  Aspartate aminotransferase (EC 2.6.1.1) (Transaminase A) (ASPAT).
  pI: 5.54, MW: 43573.36
MFENITAAPADPILGLADLFRADERPGKINLGIGVYKDET...
ABGA_ECOLI (P77357)
  Aminobenzoyl-glutamate utilization protein A.
  pI: 5.51, MW: 46588.22
MESLNQFVNSLAPKLSHWRRDFHHYAESGWVEFRTATLVA...
...

```

Fig. 4. Output of the TagIdent tool where it was used for protein identification. A protein from an *Escherichia coli* 2-D gel was uniquely identified by virtue of its N-terminal sequence tag, estimated pI, and mass. Although the program was requested here to display protein N-termini, it will show any protein that carries a specified tag in the “results with tagging” list, be it found at a protein N-terminus, C-terminus, or internally. Here the identification of the protein as DHE4\_ECOLI (P00370) is convincing not only because the tag is at the amino terminus, but because the tag was not found anywhere in the sequence of the other 462 proteins also within the specified pI and Mw window. The TagIdent output has been shortened for this figure. Note that this approach can also be used where the mass of an entire protein has been accurately determined by mass spectrometry. In such a case, the mass window used for searching can be quite small (e.g., mass  $\pm$  0.5%).

tag, one can be confident that there is no other, as yet undescribed protein that could otherwise match the tag (*see Note 18*). In fully sequenced organisms, the procedure is thus self-checking. Because of this, the TagIdent approach is very useful for organisms whose genomes are known, such as *Haemophilus influenzae*, *Mycoplasma genitalium*,



*Methanococcus jannaschii*, *Escherichia coli*, and even the eukaryote *Saccharomyces cerevisiae*. A TagIdent output for a protein from *E. coli* is shown in **Fig. 4**, and illustrates the specificity of the approach. Caution is advised when using TagIdent for the identification of proteins from poorly molecularly defined organisms, or organisms that contain large numbers of proteins (e.g., human) (*see Note 19*). A four-amino acid sequence tag (of which there are 160,000 different combinations) can be unique in microorganisms that have a total protein count of 500 to 6000, but less useful in human, for example, which has about 25,000 different known genes, and many more (possibly up to 100,000) different proteins resulting from alternative splicing. If protein identification results with TagIdent show more than one protein carrying the sequence tag in the expected region, the same sequence tag, pI, and Mw data can be used in conjunction with protein AA composition for identification with the AACompIdent tool (*see Subheading 4.2.*).

## 4.2. AACompIdent Tool

The AACompIdent tool (<http://www.expasy.org/tools/aacomp/> [13]) can identify proteins by their amino acid (AA) composition. The program matches the percent empirically measured AA composition of an unknown protein against the theoretical percent AA compositions of proteins in the Swiss-Prot/TrEMBL database. A score, which represents the degree of difference between the composition of the unknown protein and a protein in the database, is calculated for each database entry by the sum of the squared difference between the percent AA composition for all amino acids of the unknown protein and the database entry. All proteins in the database are then ranked according to their score, from lowest (best match) to highest (worst match). Estimated protein pI and Mw, as well as species of interest and keyword, can also be used in the identification procedure.

### 4.2.1. Basic Use of the AACompIdent Tool

After selecting the AACompIdent tool from the ExPASy Tools page, you must first choose the relevant AA constellation to use in matching. For AA compositions determined by standard methods, use Constellation 2. This constellation is for 16 AAs (Asx, Glx, Ser, His, Gly, Thr, Ala, Pro, Tyr, Arg, Val, Met, Ile, Leu, Phe, Lys), does not consider Cys or Trp, and calculates Asn and Asp together as Asx, and Glu and Gln together as Glx (*see Note 20*). You should specify the e-mail address to which the results should be sent, then scroll down to the “Unknown Protein” field. Here you should specify a name for the search that will appear as the subject of the e-mail message, the protein pI and Mw estimated from the 2-D gel, as well as error ranges that reflect the accuracy of these estimates. You should also specify a keyword if appropriate (*see Subheading 4.1.1.* and the Swiss-Prot list of keywords, <http://www.expasy.org/cgi-bin/keywlist.pl>), and one or more terms matching those in the Swiss-Prot OS (species) or OC (classification) lines to limit the search to one organism, or a range of organisms (*see Subheading 4.1.1.* and the Swiss-Prot list of species abbreviations, <http://www.expasy.org/cgi-bin/speclist>). Matching can also be done against all species in the database by specifying “ALL.” Finally, specify the experimentally determined AA composition of the protein, with compositional data expressed as molar percent. If you have analyzed a calibration protein in parallel with unknowns as part of your AA analysis procedure, the composition of this protein can be used to compensate for error

inherent to the AA analysis procedure (*see Note 21*). To do this, go to the “Calibration Protein” field, specify the Swiss-Prot ID name for the protein (e.g., ALBU\_BOVIN for bovine serum albumin) and enter the experimentally determined AA composition of the protein, with data expressed as molar percent. Finally, select the “Run AACompIdent” button to submit the data to the ExPASy server. Results will be sent to your e-mail address.

#### 4.2.2. Use of the AACompIdent Tool With Sequence Tags

Protein samples from 2-D gels can be submitted to Edman protein sequencing to create a sequence tag of three or four amino acids, after which the same protein sample can be used for AA composition analysis (*14*). This approach provides protein identification of higher confidence than identification by amino acid composition analysis alone. To use AA composition and sequence tag data together for protein identification, fill in the AACompIdent form as for the basic use described above but do not immediately submit it to ExPASy. Go to the bottom of the form, select the tagging option by clicking in the small box, and enter a protein sequence tag of up to six amino acids in single-AA code into the “Tag” text field. Finally, specify whether the sequence tag is N- or C-terminal, and select the “Run AACompIdent” button to submit the data to the ExPASy server. Results will be sent to your e-mail address.

#### 4.2.3. Interpretation of AACompIdent Results

The output of AACompIdent contains three lists of proteins ranked according to their AA score (**Fig. 5**). The first list is the result of matching the AA composition of the query protein against all proteins from the species of interest that have the specified keyword (if any), but without considering the specified pI and Mw. The second list shows the result of matching the AA composition of the query protein against all proteins from all species in Swiss-Prot that have the specified keyword, again without considering pI and Mw. The third list contains the results of matching the AA composition of the query protein only against the proteins from the species of interest that lie within the specified pI and Mw range (*see Note 15*) and that also have the appropriate keyword. The third list is the most powerful search. In all lists, a score of 0 indicates a perfect match between the query protein and a protein in the database, with larger scores indicating increasing difference.

We have found that a top-ranked protein is likely to represent a correct identification if it meets three conditions (**Fig. 5**). Firstly, the same protein, or type of protein, should appear at the top of the three lists. Secondly, the top-ranked protein in the third list should have a score lower than 30 (indicating a “good fit” of the query protein with that database entry). Finally, the third list should show a large score difference between the top-ranked protein and the second ranked protein (indicating a unique matching of the query protein with the top-ranked database entry). For proteins from *E. coli*, we have shown that a score difference greater than a factor of 2 gives high confidence that the top-ranked protein represents the correct identity (*13*). If the top-ranking protein in the results does not meet these three conditions, the correct identity is often within the list of best-matching proteins (*see Note 22*). In such cases, the use of AACompIdent with a protein sequence tag can provide unambiguous identification due to the high specificity of sequence tag data (*14*). **Figure 5** shows the result of protein identification by AA composition, pI, Mw, species, and sequence tag. Note that when the sequence tag

SEARCH VALUES:  
 Constellation 2  
 Species searched: ESCHERICHIA COLI  
 Keyword searched: ALL  
 Name given to unknown protein: coli147  
 pI: 5.70 Range: (5.20, 6.20)  
 Mw: 34894 Range: (27916, 41872)  
 Calibration protein: OVAL\_CHICK ( P01012 )  
 Tag= MKVA  
 An asterisk (\*) is printed to the left of a protein's rank if it carries the sequence tag.

-----  
 Scan the Swiss-Prot database (139947 entries)  
 -----

The closest Swiss-Prot entries (in terms of AA composition)  
 for the species ESCHERICHIA COLI:

Rank	Score	Protein	(pI	Mw)	Description
* 1	5	MDH_ECOLI	5.61	32337	Malate dehydrogenase (EC 1.1.1.37).
2	30	YHDH_ECOLI	5.63	34724	Protein yhdH.
3	31	K6P2_ECOLI	5.75	32388	6-phosphofructokinase isozyme 2
4	33	ALKH_ECOLI	5.57	22284	KHG/KDPG aldolase
5	35	YEIN_ECOLI	5.37	32910	Hypothetical protein yeiN.

The closest Swiss-Prot entries (in terms of AA composition)  
 for any species:

Rank	Score	Protein	(pI	Mw)	Description
* 1	4	MDH_SALTY	6.02	32451	Malate dehydrogenase (EC 1.1.1.37).
* 2	5	MDH_ECOLI	5.61	32542	Malate dehydrogenase (EC 1.1.1.37).
* 3	10	MDH_PHOPR	5.15	32391	Malate dehydrogenase (EC 1.1.1.37).
* 4	12	MDH_HAEIN	5.86	32542	Malate dehydrogenase (EC 1.1.1.37).
5	13	PUR2_CHICK	7.51	106543	Trifunctional purine biosynthetic

The closest Swiss-Prot entries (in terms of AA composition)  
 and having pI and Mw values in the specified range  
 for the species ESCHERICHIA COLI:

Rank	Score	Protein	(pI	Mw)	Description
* 1	5	MDH_ECOLI	5.61	32337	Malate dehydrogenase (EC 1.1.1.37).
2	30	YHDH_ECOLI	5.63	34724	Protein yhdH.
3	31	K6P2_ECOLI	5.75	32388	6-phosphofructokinase isozyme 2
4	35	YEIN_ECOLI	5.37	32910	Hypothetical protein yeiN.
5	36	SUCC_ECOLI	5.37	41392	Succinyl-CoA synthetase beta chain

Fig. 5. Output from the AACompIdent tool where a protein from an *Escherichia coli* 2-D gel has been correctly identified by matching its amino acid composition, sequence tag, estimated pI, and Mw against database entries for *E. coli*. The correct protein identity, malate dehydrogenase, was the top-ranked protein in the three lists, and showed a large score difference between the top- and second-ranked proteins in the third list (where pI and Mw windows are applied), and also in the first list. See text for more details about the significance of score patterns for identification confidence. In addition, the sequence tag MKVA has been found only for that protein (as shown by the asterisk). In the second list, where the amino acid composition of the query protein was matched against all entries in the Swiss-Prot database without considering protein pI and Mw, malate dehydrogenase from four different species was ranked in the top four positions. This illustrates that protein amino acid composition is well-conserved across species boundaries. The AACompIdent output has been shortened for this figure.

option is selected, the AACompIdent output will show either 40 amino acids of each protein's predicted N- or C-terminal sequence or its description, and show an asterisk to the left of a protein's rank if the protein carries the sequence tag anywhere in its sequence. If the tag is found in the displayed N- or C-terminal sequence, it will be shown in lowercase letters. We are confident that a protein from Swiss-Prot represents a correct identification if the query protein's empirically determined sequence tag of three amino acids or more is present at the expected N- or C-terminal position, and that this protein is ranked within the first 10 or so closest entries by amino acid composition.

### 4.3. AACompSim Tool

The AACompSim tool (<http://www.expasy.org/tools/aacsim/>) allows the theoretical AA composition of one protein in the Swiss-Prot database to be compared to proteins from one or all species in the database (*see Note 23*). This serves two main purposes: first, to allow the simulation of matching undertaken for identification purposes with AACompIdent (*see Subheading 4.2.*); second, to allow the detection of weak similarities between proteins by comparison of their compositions rather than sequences, as explored by Hobohm and Sander (*15*).

To use AACompSim, first select the constellation of amino acids you wish to work with (*see Subheading 4.2.1.*). If you wish to simulate matching undertaken with empirical data, you should specify constellation 2. To match against the database for detecting protein similarities, you should use all 20 amino acids in constellation 0. Then specify an e-mail address where the results can be sent, the Swiss-Prot identification name (e.g., IPIA\_TOBAC) or accession number (e.g., Q03198) of the protein you would like to compare against the database, and the Swiss-Prot abbreviation for the species to match against (e.g., SALTY for *Salmonella typhimurium*). A document containing a full list of all Swiss-Prot species and their organism codes can be found at <http://www.expasy.org/cgi-bin/speclist>. If desired, matching can be done against all species in the database by specifying "ALL." Finally, select the "Search" button to submit the query to the ExPASy server. Results will be sent to your e-mail address. AACompSim will return three lists of proteins, similar to those from AACompIdent (*see Notes 23, 24*).

### 4.4. Multident Tool

Proteins can be identified by virtue of their peptide masses alone, but frequently other data are needed to provide high-confidence identification. The same is true for protein identification with AA composition. Following our earlier observations that high-confidence protein identification can be achieved with a combination of peptide mass and AA composition data (*16,17*), we have developed the protein identification tool MultiIdent (<http://www.expasy.org/tools/multiident/>). This tool uses parameters of protein species, estimated pI and Mw, keyword, AA composition, sequence tag, and PMF data to achieve protein identification (*18*). Currently, the program works by first generating a set of proteins in the database with AA compositions close to the unknown protein, as for AACompIdent (*see Subheading 4.2.*). Theoretical peptide masses from the proteins in this set are then matched with the peptide masses of the unknown protein to find the number of peptides in common (number of "hits"). Three types of lists are produced in the results: first, a list where proteins from the database are ranked

according to their AA composition score (*see Subheading 4.2.*); second, a list where proteins are ranked according to the number of peptide hits they showed with the unknown protein; and thirdly, a list that shows only proteins that were present in the both the above lists, where these proteins are ranked according to an integrated AA and peptide hit score. In all these lists, protein pI, Mw, species of origin, and Swiss-Prot keyword can be used as in AACompIdent to increase the specificity of searches.

#### 4.4.1. Use of the Multident Tool

After selecting MultiIdent from the Tools page, you must first choose the constellation of amino acids you wish to work with. Then provide information including your e-mail address, details about the unknown protein (name, pI and Mw estimations, amino acid composition, sequence tag data if available), species of interest for matching (*see Notes 16, 22*), and Swiss-Prot keyword (if any). This should be done in essentially the same manner as for AACompIdent (*see Subheading 4.2.*). To include peptide masses for protein identification, first specify the size of the list to be created with the query protein's AA composition (e.g., 500). Then click in the checkbox next to the "Peptide Mass Fingerprinting" title in the program to enable this option, enter the list of peptide masses into the text box with an accuracy of at least one decimal place, and specify whether masses are monoisotopic or average. Then specify the enzyme used to create the peptides (e.g., trypsin), whether the protein was reduced and alkylated with any reagent (e.g., iodoacetamide, iodoacetic acid, 4-vinylpyridene) before cleavage, whether artifactual protein modifications such as oxidation of methionine or acrylamide adducts to cysteine are expected, and the mass tolerance to be used in matching with peptides. The mass tolerance should reflect the known accuracy of your mass spectrometer. Finally, specify which results lists you would like to see in the MultiIdent output, and select the "Perform" button to submit the match to ExPASy. The results will be sent to you by e-mail.

#### 4.4.2. Interpretation of Results

The list of closest Swiss-Prot entries in terms of protein AA composition is the same as for the AACompIdent output, and thus results and scores for proteins in this list can be interpreted as in **Subheading 4.2.3**. If sequence tags are used as part of a search strategy, the list of closest proteins in terms of protein AA composition will show the predicted protein N- or C-terminal sequence, and any sequence tags present will be highlighted in lowercase letters. Asterisks are also shown to the left of protein rank numbers to indicate that the sequence tag is present in the corresponding protein. These asterisks are used in the list of best matches by AA composition, as well as the lists of proteins generated by PMF and the integrated score.

The list of closest Swiss-Prot entries in terms of peptide hits is simply the list of proteins that have the most peptides in common with the query protein. The "hits" are the number of peptides that match with a database entry, and the peptide masses shown in the output are those from the database entry that match with those from the query protein. The top-ranked protein in this list will be the most likely identification of the protein; however, this may not be so if matching has been done with very large Mw windows. In any case, use of sequence tags of even three or four amino acids with peptide mass data can greatly increase the confidence of a database entry representing a correct identification. Note that for this purpose, sequence tags generated by tandem



mass spectrometry (MS/MS) or by postsource decay matrix-assisted laser desorption/ionization (MALDI)-time-of-flight (TOF) techniques can be used in MultiIdent as well as tag data generated at protein N- or C-termini.

The list of proteins with best integrated scores represents the most powerful form of matching (*see Note 25*). It can simultaneously consider the protein parameters of pI, Mw, AA composition, sequence tag, and peptide masses in order to rank proteins from the database for the species of interest (*see Note 22*) with a given keyword. The integrated score is a measurement of difference between the query protein and a database entry, and is derived by dividing the AA analysis score by the number of peptide hits that were found for that protein. Accordingly, an integrated score of 0 represents a perfect match for a query protein, with larger scores representing increasing differences. We find that the integrated score is useful for defining confidence limits if it is not immediately apparent whether a protein has been correctly identified.

## 4.5. Aldente Tool

### 4.5.1. Description

Aldente (Advanced Large-scale iDENTification Engine, <http://www.expasy.org/tools/aldente/>) is a tool that allows the identification of proteins using peptide-mass fingerprinting data.

Experimentally measured, user-specified peptide masses are compared with the theoretical peptides calculated for all proteins in the Swiss-Prot/TrEMBL databases. Isoelectric point, molecular weight, and a species (or group of species) can be specified in order to restrict the number of candidate proteins and reduce false-positive matches.

The main features of Aldente are:

- Use of a robust method (the Hough transform) to determine the deviation function of the mass spectrometer and to resolve peptide match ambiguities. In particular, the method is relatively insensitive to noise (*see Note 26*).
- Tuneable score parametrization: the user can choose the parameters he or she wants to take into consideration in the score and in which proportion.
- Extensive use of the annotations (protein mature form, posttranslational modifications, alternative splicing) in Swiss-Prot/TrEMBL, offering a degree of protein characterization as part of the identification procedure (*see Note 27*).
- Consideration of user-defined chemical amino acid modifications (oxidation of methionine, acrylamide adducts on cysteine residues, alkylation products on cysteine residues), and the possibility to define their contribution to the score.

### 4.5.2. Use of Aldente

After selecting the Aldente tool from the ExPASy Tools page, you should enter the sample information, with a query name, pI, and Mw estimations if known (*see Note 28*). A list of experimental peaks (*see Notes 10, 38*) can be entered directly into the text area (peptide masses with or without peak intensities), or by uploading a text file (*see Note 29*). You should then select the database to search (Swiss-Prot and/or TrEMBL), and the species (or group of species) if you want to restrict your query (*see Note 30*).

You have to specify the enzyme that was used to generate the peptides. In order to take into account partial cleavages, you can specify 0 or 1 missed cleavage sites to be allowed. You must define a minimum number of peptide-mass hits required for a matching protein to be included in the result list. The default value is 4.



Mass tolerance has to be set on protein and peptide levels. On protein level, upper and lower mass limits can be specified, which serve to filter the results but are not taken into account in the scoring (in contrast to the  $M_w$  estimation mentioned above). Peptide mass tolerance corresponds to the estimated internal precision of the mass spectrometer: the instrument's accuracy can be specified, either with an absolute value in Daltons or with a relative value in ppm (parts per million), or with both (*see Note 31*). Less accurate peptide mass data will require a larger mass tolerance and will result in a lower accuracy of your search (*see Note 32*).

Then specify the chemical modifications occurring on the unknown protein before cleavage, and the way to take them into account in the score (*see Note 33*).

You may also choose which types of PTMs annotated in Swiss-Prot you want to take into account (e.g., only experimentally proven or also computationally predicted ones) and the way to consider them in the score (*see Notes 34 and 35*).

Finally specify the maximum number of proteins you want to be displayed in the Aldente output, and select the "Submit" button to send your query to ExPASy. Depending whether you provide your e-mail address or not, the results will be sent to you by e-mail or displayed directly in your browser.

#### 4.5.3. Aldente Output

The top part of the output result provides the date of the query, the database release number and current number of entries, and some statistics about the search. The protein statistics give the total number of proteins in the selected database and taxa, the number of proteins in the protein mass range, the number of proteins with enough peptides in the peptide mass tolerance, the number of proteins with the minimum number of hits after alignment, and the number of displayed proteins (*see Note 36*). Then the peptide statistics give the number of peptides generated in the mass range of your sample, the number of peptides matching a peak of the sample, and the average number of theoretical peptides per protein in the mass range.

Then follows a summary of the best-matching proteins from the database (**Fig. 6**), with a "quick jump" link to detailed peptide information provided further down in the same page.

After that, for each matching protein, detailed information concerning matching peptides is given (**Fig. 7**), with individual score, difference between the experimental and calculated masses, information regarding PTM or chemical modification if any, peptide position, and sequence. Finally the protein sequence is visualized with identified peptides in blue and uppercase, where trypsin loci (K, R) are shown in red.

Aldente results are displayed online or sent by e-mail, in the form of an html table, or in XML or text format for easier parsing. The html result contains direct links to FindMod (*see Subheading 4.6.*), GlycoMod (*see Subheading 4.7.*), and FindPept (*see Subheading 4.8.*) to further characterize matching proteins by predicting potential protein post-translational modifications, finding potential single-amino acid substitutions and potential unspecific cleavage, to PeptideMass (*see Subheading 3.3.*), and to BioGraph (*see Subheading 4.9.*), for the graphical representation of the theoretical spectrum. Relevant input data and/or information about the matching database entry are automatically transferred to those programs.

A new Aldente search can be launched directly from the result output. This allows the user to submit a second search with slightly modified parameters, i.e., with modi-

Aldente version Beta 15/01/2004 feedback is welcome [Documentation](#) [Input summary](#) [Printable page](#)

Date 22/01/2004 10:34:04 UTC

Release Swiss-Prot Release 42.7 of 15-Dec-2003: 141681 entries

Proteins Scanned 163999 / In mass range 155784 / Enough hits before alignment 53888 / Enough hits after alignment 33067 / Displayed 20

Peptides Generated 8507732 / Matching 473381 / Average of 54 peptides per protein

Rank	Z-score	Hits	Taxon	AC	ID	Mass	pl	Shift	Slope	DE
<a href="#">1</a>	25.61	20	Homo sapiens	<a href="#">P17844</a>	DDX5_HUMAN	69147	9.06	-0.016	37	Probable RNA-dependent helicase p68 (DEAD-box protein p68) (DEAD-box)
<a href="#">2</a>	20.62	13	Homo sapiens	<a href="#">Q9Y2X3</a>	NOP5_HUMAN	59577	9.03	-0.027	47	Nucleolar protein NOP5 (Nucleolar protein 5) (NOP58) (HSPC120).
<a href="#">3</a>	17.90	17	Mus musculus	<a href="#">Q61656</a>	DDX5_MOUSE	69319	9.06	-0.035	50	Probable RNA-dependent helicase p68 (DEAD-box protein p68) (DEAD-box)
<a href="#">4</a>	10.96	9	Schizosaccharomyces pombe	<a href="#">O14254</a>	IDHP_SCHPO	47293	8.86	0.104	-37	Probable isocitrate dehydrogenase [NADP], mitochondrial precursor
<a href="#">5</a>	10.86	9	Rattus norvegicus	<a href="#">Q9QZ86</a>	NOP5_RAT	60070	8.7	-0.008	27	Nucleolar protein NOP5 (Nucleolar protein 5) (Nopp140 associated)
<a href="#">6</a>	10.75	4	Escherichia coli	<a href="#">P09168</a>	OGT_ECOLI	19179	8.48	0.152	-50	Methylated-DNA--protein-cysteine methyltransferase (EC 2.1.1.63) (6-O-
<a href="#">7</a>	10.75	4	other Bacteria	<a href="#">P09168</a>	OGT_ECOLI	19179	8.48	0.152	-50	Methylated-DNA--protein-cysteine methyltransferase (EC 2.1.1.63) (6-O-
<a href="#">8</a>	10.61	3	Escherichia coli	<a href="#">P10396</a>	REP1_ECOLI	8941	10.92	0.096	-10	Replication initiation protein (Fragment).
<a href="#">9</a>	10.58	8	other Bacteria	<a href="#">P51962</a>	RIBB_PHOPO	40064	5.3	-0.069	70	3,4-dihydroxy-2-butanone 4-phosphate synthase (DHBP synthase).
<a href="#">10</a>	10.04	6	Viruses	<a href="#">Q65163</a>	9GL_ASFB7	14378	8.52	-0.013	40	Late protein 9GL.
<a href="#">11</a>	9.91	6	Viruses	<a href="#">P03346</a>	GAG_HTLV2_C3	24386	8.4	-0.059	80	CORE PROTEIN P12 (P15).
<a href="#">12</a>	9.80	5	Chlorophyta	<a href="#">Q9TKZ6</a>	RR7_NEPOL	17812	10.78	-0.080	63	Chloroplast 30S ribosomal protein S7.
<a href="#">13</a>	9.28	5	Archaea	<a href="#">Q8PTU1</a>	PSMA_METMA	27353	5.14	0.264	-87	Proteasome alpha subunit (EC 3.4.25.1) (Multicatalytic endopeptidase)
<a href="#">14</a>	9.16	4	other Bacteria	<a href="#">Q8DC31</a>	RL19_VIBVU	13187	10.43	0.328	-233	50S ribosomal protein L19.
<a href="#">15</a>	8.94	7	Chlorophyta	<a href="#">P54214</a>	SFAS_DUNBI	31412	5.42	0.296	-173	SF-assembly.
<a href="#">16</a>	8.84	5	other Bacteria	<a href="#">Q87C45</a>	PRMA_XYLFT	33029	4.43	0.144	-33	Ribosomal protein L11 methyltransferase (EC 2.1.1.-) (L11 Mtase).
<a href="#">17</a>	8.62	5	Homo sapiens	<a href="#">Q15649</a>	TR13_HUMAN	17318	5.52	0.203	-120	Thyroid receptor interacting protein 3 (TRIP-3) (Fragment).
<a href="#">18</a>	8.20	4	other Bacteria	<a href="#">Q9KYP1</a>	PAAD_STRCO	22614	8.89	0.155	-103	Probable aromatic acid decarboxylase (EC 4.1.1.-).
<a href="#">19</a>	8.11	6	other Streptophyta	<a href="#">Q00058</a>	MI25_ORYSA	22087	9.5	0.245	-210	Mitochondrial 22 kDa protein (ORF 25).
<a href="#">20</a>	7.96	5	other Bacteria	<a href="#">Q9PJL6</a>	RL23_CHLMU	12134	9.98	-0.011	7	50S ribosomal protein L23.

[Resubmit](#)

Graphical visualisation of the results : [BioGraph](#)

Fig. 6. First part of the Aldente output, showing the summary of the best-matching proteins from the database. The top of the page summarizes some statistics about processed proteins and peptides, followed by the list of best-matching proteins, with related information. Access to documentation, submitted parameters, printable page, graphical visualization of the results with BioGraph (see **Subheading 4.9.** and **Fig. 8.**), and “resubmit” function are provided.

fied molecular weight or pI ranges, number of missed cleavages, taxonomic range; or to resubmit an archived query at a later stage, for later database releases. This is particularly useful if the initial identification was unsuccessful or ambiguous.

#### 4.6. FindMod Tool

The FindMod, GlycoMod (see **Subheading 4.7.**), and FindPept (see **Subheading 4.8.**) tools are used to identify the origin of peptide masses obtained by PMF that are not matched by protein identification tools such as Aldente. They also take into account posttranslational modifications annotated in Swiss-Prot or supplied by the user, and chemical modifications of peptides. It is quite common for PMF tools not to be able to find matching theoretical peptides for a few of the less intense peaks that were detected and submitted to the identification process.

FindMod (<http://www.expasy.org/tools/findmod/> [19]) is a program for *de novo* discovery of protein PTM or single-amino acid substitutions. It examines PMF results of known proteins for the presence of more than 20 types of PTMs of discrete mass, such

2) [Q9Y2X3](#) NOP5\_HUMAN Swiss-Prot: Homo sapiens Nucleolar protein NOP5 (Nucleolar protein 5) (NOP58) (HSPC120). [Up](#)

Z-score : 20.62 Hits : 13 Mw : 59577 pl : 9.03 Coverage : 28% Shift : -0.026667 dalton Slope : 47 ppm

Exp	Theo	Intensity			Delta		Dev	Cont	MC	CAM	MSO	PTM	Position		Sequence
		Da	Da	UI	%	rank							Da	ppm	
975.502	975.504004	1795	13	51	-0.00	-2	-21	-	1	1/1	-	-	207	- 214	CLQKVGDR
* 1327.712	1327.664064	13417	100	1	0.05	36	9	-	-	-	-	-	269	- 278	TQLYEYLQNR
1398.754	1398.704546	8105	60	3	0.05	35	7	-	-	-	0/1	-	121	- 133	SQMDGLIPGVEPR
1414.74	1414.69946	2545	18	28	0.04	28	0	-	-	-	1/1	-	121	- 133	SQMDGLIPGVEPR
1584.875	1584.836664	2450	18	33	0.04	24	-5	-	-	-	-	-	222	- 235	LSELLPEEVEAEVK
* 1733.942	1733.922062	1453	10	67	0.02	11	-20	-	-	-	-	-	338	- 353	YGLIYHASLVGQTSPK
1834.819	1834.754792	4564	34	8	0.06	34	2	-	-	-	0/1	-	372	- 388	YDAFGEDSSSAMGVENR
1850.814	1850.749706	1481	11	65	0.06	34	2	-	-	-	1/1	-	372	- 388	YDAFGEDSSSAMGVENR
1882.02	1881.967308	4014	29	12	0.05	27	-4	-	-	1/1	0/1	-	102	- 117	LNLSCHISPPVNLMLR
1959.07	1959.02238	2133	15	40	0.05	24	-9	-	-	-	0/3	-	279	- 297	MMAIAPNVTVMVGVGELVGAR
1978.001	1977.98036	1220	9	83	0.02	10	-23	-	1	-	-	-	22	- 37	LQEVDSLWKEFETPEK
2122.043	2121.955714	842	6	100	0.09	41	6	-	1	-	-	-	468	- 485	VEEEEEKVAEEEEESVK
2139.182	2139.104858	896	6	97	0.08	36	1	-	1	1/1	0/1	-	100	- 117	EKLNLSCHISPPVNLMLR

#### Details of the alignment

```

1  mlvlfetsvg yaifkvlnek kLQEVDSLWK EFETPEKank ivklkhfef qdtaealaaf talmegkink qlkvvkkliv
81 keaheplava daklggwikE KLNLSCHISP VVNLMLRgir SQMDGLIPGV EPRemaamcl glahslsryr lkfsadvdt
161 mivqaislld dldkelnnyi mcrcreygwh fpelgkiisd nltycKCLQK VGDknyasa kLSELLPEEV EAEVkaaaei
241 smgtevseed icnllhctq vieiseyTQ LYEYLQNRMM AIAPNVTVMV GELVGARlia hagellnlak haastvqilg
321 aekalfraik srrdtpkYGL IYHASLVGQT SPKkkgkisr mlaaktvlai rYDAFGEDSS SAMGVENRak learlrtdled
401 rgirkisgtg kalaktekye hksevktyp sgdstlptcs kkrkiegvdK edeitekkak kakikvVEE EEEBVAEEE
481 ETSVKKkkkr gkkkhikeep lseeeptct aiaspekkkk kkkkrened

```

[GlycoMod](#) [FindMod](#) [FindPept](#) [PeptideMass](#)

Fig. 7. Second part of the Aldente output, showing details for one of the matching proteins. The output shows information on the matched peptides, including individual score, difference between the experimental and calculated masses, information regarding posttranslational modification (PTM) or chemical modification if any, peptide position, and sequence. The protein sequence is displayed with identified peptides in blue and upper case, and trypsin loci (K, R) (not visible in this figure). Links to ExPASy characterization tools are also provided.

as acetylation, amidation, biotin, C-mannosylation, deamidation, *N*-acyl diglyceride cysteine (tripalmitate), FAD, farnesylation, formylation, geranyl-geranyl,  $\gamma$ -carboxy-glutamic acid, *O*-GlcNAc, hydroxylation, lipoyl, methylation, myristoylation, palmitoylation, phosphorylation, pyridoxal phosphate, pyrrolidone carboxylic acid, and sulfation.

This is done by looking at mass differences between experimentally determined peptide masses and theoretical peptide masses calculated from a specified protein sequence. If a mass difference corresponds to a known PTM not already annotated in Swiss-Prot, rules are applied that examine the sequence of the peptide of interest and make predictions as to what amino acid in the peptide is likely to carry the modification. The same method is applied when predicting potential amino acid substitutions.

#### 4.6.1. Input Parameters

FindMod is usually launched after a PMF identification run, for the most likely protein suggested by an identification program such as Aldente. The output of Aldente contains a link to the FindMod submission form with most parameters already filled in. If you wish to launch FindMod directly, you should specify the sequence of the protein you would like to characterize and for which you have determined a set of peptide masses. If this protein is known in Swiss-Prot/TrEMBL, enter the Swiss-Prot ID code or the protein accession number. Otherwise, you can enter the sequence of your protein of interest, in single-letter amino acid code, in either upper or lower case (*see Note 37*).

Protein sequences from other sources (e.g., word-processor programs or other Web pages) can be copied and pasted directly into this field. If there are spaces in your sequence, these will be ignored.

The characters *B*, *X*, or *Z* are accepted, but no masses are computed for peptides containing one or more of these characters.

A set of experimental masses must also be provided (*see* **Notes 29, 38**). The experimental peptide masses will first be compared to theoretical unmodified peptides and to peptides modified as documented in Swiss-Prot or by chemical modifications. The user can choose whether all peptide masses or only those that have not been attributed a theoretical peptide in this process should be examined for potential PTMs and/or single-amino acid substitutions.

If you wish to take into account other posttranslational modifications than those already predictable by FindMod, you can enter, for each of these modifications, its name, its atomic composition, and the amino acids on which this modification can be observed.

Further parameters are isotopic resolution (average or monoisotopic masses), chemical treatment of cysteine (*see* **Note 39**), oxidation state of methionine (*see* **Note 40**), mass tolerance (in ppm or in Daltons), digestion agent, and number of missed cleavages (up to three). You can enter the masses of your peptides as  $[M]$  or as  $[M+H]^+$  (*see* **Note 10**).

#### 4.6.2. FindMod Output

The results from FindMod are divided into a header and up to three tables.

The header contains information about the submitted protein: a link to the Swiss-Prot/TrEMBL entry and the description line (if the protein is in Swiss-Prot/TrEMBL), pI, and molecular weight. Then the input parameters are listed, followed by an active link to PeptideMass. This allows the user to perform a theoretical cleavage of the protein of interest.

The tables report the peptides whose experimental masses match unmodified or modified theoretical digest products of the protein of interest.

The first table reports matches to theoretical digest products as unmodified, modified with the annotations in Swiss-Prot, and chemically modified as specified in the input form.

The second table reports those user masses that differ from a theoretical database mass by a mass value corresponding to one of the considered PTMs. These peptides are further examined, and FindMod checks whether the peptide sequences contain amino acids likely to carry the modification in question. This is done by applying a set of prediction rules that have been defined using information in the PROSITE database (**20**), examining all the PTM annotations in Swiss-Prot and information in the literature. The program first lists the matches conforming to these rules, highlighting potentially modified residues in color. Potential PTMs detected by mass difference but not confirmed by the rules are included in a second list.

The third table shows potential single-AA substitutions detected by mass difference. The following particularities are worth pointing out:

1. A BLOSUM62 score (**21**) is given for each suggested single-AA substitution. This provides information about the probability of substitution: Lowest score: -4 (low probability of substitution); highest score: 11 (high probability of substitution).

2. Potential single-amino acid substitutions are not displayed if they occur on the cleavage site and substitute the AA for an AA after which the digestion enzyme does not cleave.
3. If the suggested AA substitution corresponds to a sequence variant or conflict as annotated in the Swiss-Prot feature table, this substitution is highlighted in color, and a hypertext link is provided to the corresponding annotated variant or conflict.

At the end of the output page, the user will find a list of those entered matches that did not match in any of the previous tables (if any).

#### **4.7. GlycoMod and GlycanMass Tools**

Protein glycosylation is one of the most common and most complex post-translational modifications. Although the problem of predicting glycosylation from peptide-mass fingerprinting data is in principle the same as the one addressed by FindMod, the complexity and heterogeneity (the high number of possible combinations of monosaccharides forming glycan structures) made it necessary to conceive a separate tool, specializing in glycan structures and glycopeptides, GlycoMod.

GlycoMod (<http://www.expasy.org/tools/glycomod/> [22, 23]) finds all possible compositions of a glycan structure from its experimentally determined mass. It may be used to calculate the possible compositions of free or derivatized glycan structures, or compositions of glycans attached to glycoproteins and glycopeptides. The motivation and use of the tool are quite similar to FindMod. As there has been a recent book chapter devoted entirely to the use of GlycoMod (23), we will not detail its use here.

GlycanMass (<http://www.expasy.org/tools/glycomod/glycanmass.html>) allows the user to calculate the mass of a glycan from its monosaccharide composition. Available elements to build the oligosaccharide are hexose, HexNAc, deoxyhexose, NeuAc, NeuGc, pentose, sulfate, phosphate, KDN, and HexA. The user has the possibility to specify whether the monosaccharide residues are underivatized, permethylated, or peracetylated, and whether to use average or monoisotopic mass values.

#### **4.8. FindPept Tool**

##### **4.8.1. Description**

FindPept (<http://www.expasy.org/tools/findpept/>, [24]) is designed to predict peptides resulting from the following causes: unspecific proteolytic cleavage, missed cleavage, protease autolysis, and keratin contaminants (*see* **Notes 41, 42**).

Unspecific cleavage is the process by which peptides whose termini do not correspond to the cleavage specificity rules implemented in computer programs are produced by proteolysis. These rules are often simplistic and reflect our incomplete understanding of the specificity of certain enzymes (*see* **Note 43**). Other causes include a contamination with other proteases (e.g., trypsin usually contains traces of chymotrypsin), biological processes such as protein degradation, or a change in enzyme specificity over time (*see* **Note 44**).

##### **4.8.2. Using FindPept**

FindPept is not a part of the identification procedure, as it requires a protein as input. Therefore, a search with a tool like Aldente should be carried out first to match peptides resulting from specific cleavage and report a candidate protein. FindPept can be directly launched from the Aldente output, or can be accessed via its submission form, but also from the results of the FindMod or GlycoMod programs.



A protein sequence should be provided. If it is a Swiss-Prot/TrEMBL accession number, the program will read the annotation relative to posttranslational modifications in Swiss-Prot feature tables, and use these to generate a table of posttranslational modifications (*see* **Notes 45, 46**). If it is a user-entered sequence, expected post-translational modifications may be specified by entering their abbreviations within brackets. User-defined modifications can also be applied to any position or residue of choice. They can be entered by specifying the name and atomic composition of the PTM and the position(s) to which the PTM applies. A position can be supplied as a number (“18” = residue number 18 relative to the Swiss-Prot or user-entered sequence), one or more amino acids (“E D” = all glutamate and aspartate residues), or an anchor (“<” = N-terminus of each peptide). This functionality is especially useful if an atypical chemical reactant has been added during the experiment.

A set of experimental masses must also be provided, in the same format as specified for FindMod (*see* **Subheading 4.6.**). Expected chemical modifications should be supplied (*see* **Notes 39, 47**).

The enzyme or chemical reagent used to generate the peptides can optionally be indicated. In this case, cleavage sites that obey cleavage rules at either end are highlighted in the results with a red slash, and peptides that obey them at both ends are displayed in a separate table. Additionally, a set of human keratins is theoretically digested, and matching masses are reported. A drop-down list adjacent to that used to select the enzyme can be used to specify the source of the enzyme from a list of the most common sources and variants, when the sequence of these enzymes is known and present in Swiss-Prot/TrEMBL. If one is selected, its sequence is submitted to a theoretical self-digestion, and the user masses are checked for autolysis fragments, accepting missed cleavages. PTMs present in the feature table of the Swiss-Prot entry for the enzyme are also taken into account (e.g., phosphorylation of pig pepsin, accession number P00791).

The output page (*see* **Note 48**) is divided into a header and up to seven tables. Each table is displayed only if matching peptides/PTMs have been found in the given category. The tables “Post-Translational/Artefactual Modifications for the protease/the protein” list the PTMs applied to the protease, and to the studied protein. The table “Masses resulting from possible contaminants” lists the masses that correspond to the specific cleavage of a number of human keratins (*see* **Note 49**). The table “Peptides resulting from protease autolysis” lists the peptides obtained by specific self-digestion of the protease that match the user masses. The table “Matching peptides for specific cleavage” lists the peptides obtained by digestion of the studied protein for which both ends are either sites of specific cleavage by the protease, or the extremities of the original peptide. The table “Matching peptides for unspecific cleavage” lists all peptides obtained by allowing cleavage at any position in the sequence of the studied protein that match the user masses, for which at least one cleaved extremity does not match the enzyme cleavage rules. The table “Unmatched masses” lists the masses that could not be identified by the program, i.e., not assigned to one of the tables above. A set of buttons allows you to directly submit those masses to the FindMod and GlycoMod tools, and possibly identify PTMs or glycosylated sites on the protein.



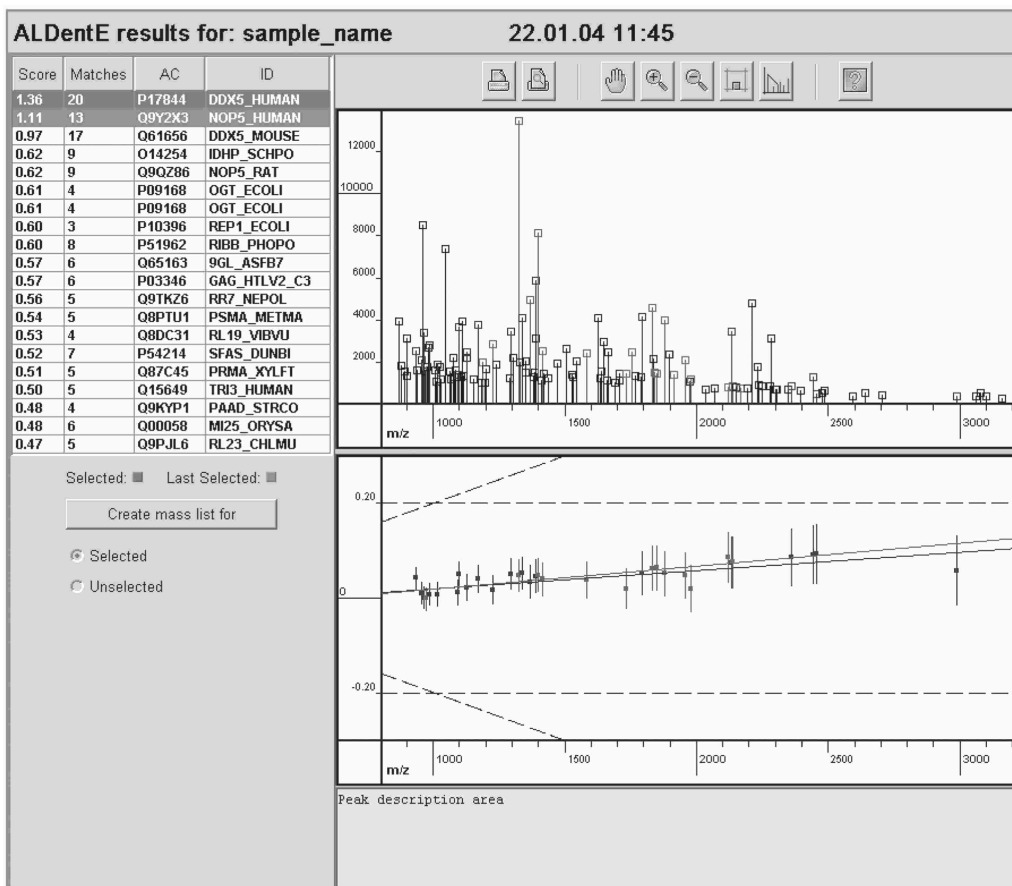


Fig. 8. The BioGraph applet, used to visualize an Aldente identification result. The lower part of the spectrum analysis panel shows the  $m/Z$  ratio on the horizontal axis and the arithmetic difference between theoretical and experimental mass (in a user-defined range) on the vertical axis. The two best-matching proteins, P17844 and Q9Y2X3, are selected in the score list, which causes the straight lines corresponding to their spectra to appear in this panel. As both lines have very similar slopes, this validates the assumption of their co-occurrence in the sample.

## 4.9. BioGraph Tool

### 4.9.1. Description

BioGraph (<http://www.expasy.org/tools/BiographApplet/>) is a Java applet that aims at providing ExPASy users with an interactive interface to visualize results of some proteomics tools. BioGraph is therefore accessible from Aldente, FindMod, or FindPept results by clicking on the “BioGraph” button.

This viewer is composed of three main components, or panels (*see Fig. 8*, which shows an Aldente identification result, visualized with BioGraph): first, the “Title panel,” intended to give general information about the source program; then the “Tool

results panel,” to summarize the source program results and interact with the spectrum; and lastly, the “Spectrum manipulation panel,” to interactively visualize the user-entered spectrum.

#### 4.9.2. General Features

The “Title panel” provides three information items:

- The source tool name (either Aldente, FindMod or FindPept),
- The user-entered protein name,
- The date and time at which source program has been run.

The “Spectrum manipulation panel” is composed of three basic components: the toolbar, the spectrum, and the peak information panel. The toolbar consists of eight buttons to:

- Print the content of the spectrum area,
- Preview what will be printed,
- Move the spectrum on the horizontal axis,
- Zoom in to the spectrum on the horizontal axis,
- Zoom out of the spectrum on the horizontal axis,
- Select a region and zoom it on the horizontal axis,
- Compare two peaks on the spectrum.

The spectrum summarizes user-entered data, i.e., the  $m/Z$  ratio on the horizontal axis and the intensities on the vertical axis.

In the case where Biograph is called from Aldente, another graph is displayed, which is described in **Subheading 4.9.3**.

The peak information panel displays, if the mouse is moved over a peak, this peak’s properties, i.e., mass and intensity values on the one hand and data about the matched proteins on the other hand, including:

- Swiss-Prot or TrEMBL accession code and ID,
- matched peptide mass,
- difference between current mass and user-entered mass, number of missed cleavages,
- a symbol to indicate whether or not the matched sequence contains modifications,
- “from” and “to” positions of the match
- corresponding sequence.

#### 4.9.3. Source Program-Specific Features

These features, which depend on the program from which BioGraph has been called, are summarized in the “Tool results panel.”

In the case where BioGraph was called from Aldente, the tool results panel displays a table of matched proteins whose rows can be selected to highlight the corresponding peaks on the spectrum.

Another Aldente-specific feature is a graph that aims at providing the user with a new visualization of data. This graph shows the  $m/Z$  ratio on the horizontal axis and the arithmetical difference between theoretical and experimental mass (in a user-defined range) on the vertical axis. The goal of such a representation is first to evaluate the spectrometer-intrinsic user-defined error rates in order to extract noise from signal, and then to validate true positive matches for the current run (as their corresponding points follow the same straight line). This is illustrated in **Fig. 8**.

When BioGraph is called from FindMod or FindPept, the tool results panel consists of four checkboxes, whose selection leads to the highlighting of peaks corresponding to peptides with specified properties (e.g., matching or potentially modified peptides).

The option “Mass list creation” is available for both types of tool results panel: the user can export peaks of interest by using the “Create mass list” button.

## 5. Integration of the Tools With Each Other and With Swiss-Prot/TrEMBL

The ExPASy protein identification and characterization tools, in particular Aldente, FindMod, GlycoMod, and FindPept, are closely integrated and hyperlinked with Swiss-Prot and TrEMBL entries on ExPASy, and among each other. Navigation between database entries, data submission forms, and program results is made easy (in both directions), and the number of copy/paste and mouse click operations is minimized.

For example, in addition to its submission form, GlycoMod can be used, after the identification of a potential candidate protein from peptide-mass fingerprinting with Aldente, to further characterize this protein: GlycoMod may explain some of the unmatched, peaks which correspond to glycopeptides, i.e., the linkage of an oligosaccharide. FindMod, which allows the user to predict discrete posttranslational modifications or amino acid substitution, or FindPept, which detects unspecific cleavage or peaks due to contaminants or enzyme autolysis, may also be used before or after GlycoMod. Independently of the order in which you choose to apply these tools, direct submission forms are available with all the relevant data already filled in, and it is not necessary to go back to the original submission forms.

All tool submission pages (just like all other html pages on ExPASy) contain a search menu, which allows for easy keyword searches in the Swiss-Prot/TrEMBL as well as any other ExPASy database.

The NiceProt view of any Swiss-Prot/TrEMBL entry on ExPASy contains direct links to the results or, if additional parameters are required, to the submission forms of several protein-characterization tools. In the latter case, the submission forms already have the sequence information filled in, which again minimizes the number of copy/paste operations.

Sometimes it may be of interest to perform sequence analysis or prediction on a subsequence of the precursor molecule annotated in Swiss-Prot. This option is also supported on ExPASy: The positions (ranges) of certain regions of interest annotated in the Swiss-Prot feature tables (FT) are hyperlinked in the NiceProt view, giving access to a page that highlights the region in color, and that contains links allowing the user to submit just that region to the same analysis tools as those available for the complete sequence via NiceProt.

## 6. Notes

1. Protein pI is calculated using pK values of amino acids described in Bjellqvist et al. (25,26), which were defined by examining polypeptide migration between pH 4.5 and 7.3 in an immobilized pH gradient gel environment with 9.2 M and 9.8 M urea at 15°C or 25°C. Prediction of protein pI for highly basic proteins is yet to be studied, and it is possible that current Compute pI/Mw predictions may not be adequate for this purpose. The buffer capacity of a protein will affect the accuracy of its predicted pI, with poor buffer capacity leading to greater error in prediction (25,26). Because of this, pI predictions for small proteins can be problematic.

2. Protein Mw is calculated by the addition of average isotopic masses of amino acids in the protein and the average isotopic mass of one water molecule. This program does not account for the effects of posttranslational modifications; thus, modified proteins on a 2-D gel may migrate to a position quite different from that predicted. Protein glycosylation in particular can affect protein migration in both pI and Mw dimensions. Note, however, that the “GET REGION ON 2D PAGE” function in SWISS-2DPAGE (27) (accessed by selecting a “GET REGION ON 2D PAGE” hypertext link from a Swiss-Prot entry) uses the Compute pI/Mw algorithm to highlight the region on a 2-D gel to where an unmodified protein should run, and suggests a region where the modified protein might be found if it has modifications documented in the Swiss-Prot database.
3. Signal sequences or transit peptides of unknown length, however, become increasingly rare (currently 444 Swiss-Prot protein sequences out of 139,947): whenever signal sequences (and their length) are not experimentally determined, the manual annotation process includes the use of prediction programs (e.g., SignalP [28]), which results in annotation of potential signal sequences.
4. It is not possible to specify posttranslational modification for your protein, nor will ProtParam know whether your mature protein forms dimers or multimers. If you do know that your protein forms a dimer, you may just duplicate your sequence (i.e., append a second copy of the sequence to the first), as all computations performed by ProtParam are based on either compositional data, or on the N-terminal amino acid.
5. ProtParam sums the contributions of the different amino acids as if they were independent, not taking into account secondary or tertiary structure. Exact coefficients need to be measured experimentally.
6. The extinction coefficient is calculated using the equation:

$$E(\text{Prot}) = \text{Numb}(\text{Tyr}) \times \text{Ext}(\text{Tyr}) + \text{Numb}(\text{Trp}) \times \text{Ext}(\text{Trp}) + \text{Numb}(\text{Cystine}) \times \text{Ext}(\text{Cystine})$$

The absorbance (optical density) can be calculated using the following formula:

$$\text{Absorb}(\text{Prot}) = E(\text{Prot})/\text{Molecular\_weight}$$

The conditions at which these equations are valid are: pH 6.5, 6.0 M guanidium hydrochloride, 0.02 M phosphate buffer.

7. It has been shown (29–31) that the identity of the N-terminal residue of a protein plays an important role in determining its stability in vivo. It seems that the N-terminal residue plays a major role in the process of ubiquitin-mediated proteolytic degradation (for a review see ref. 32). The authors have, by site-directed mutagenesis, created beta-galactosidase proteins with different N-terminal amino acids. The  $\beta$ -gal proteins thus designed have strikingly different half-lives in vivo, from more than 100 h to less than 2 min, depending on the nature of the amino acid at the amino terminus and on the experimental model (yeast in vivo; mammalian reticulocytes in vitro, *E. coli* in vivo). The set of individual amino acids can thus be ordered with respect to the half-lives that they confer when present at the amino terminus of a protein (this is called the “N-end rule”).
8. Statistical analysis of 12 unstable and 32 stable proteins has revealed (33) that there are certain dipeptides, the occurrence of which is significantly different in the unstable proteins compared with those in the stable ones. The authors of this method have assigned a weight value of instability to each of the 400 different dipeptides (DIWV). Using these weight values, it is possible to compute an instability index (II), which is defined as:

$$i = L-1$$

$$II = (10/L) \times \text{Sum DIWV}(x[i]x[i+1])$$

$$i=1$$

where: L is the length of sequence

DIWV(x[i]x[i+1]) is the instability weight value for the dipeptide starting in position i.

9. The aliphatic index of a protein is calculated according to the following formula (34):

$$\text{Aliphatic index} = X(\text{Ala}) + a \times X(\text{Val}) + b \times [X(\text{Ile}) + X(\text{Leu})]$$

where X(Ala), X(Val), X(Ile), and X(Leu) are mole percent (100 × mole fraction) of alanine, valine, isoleucine, and leucine. The coefficients a and b are the relative volume of valine side chain (a = 2.9) and of Leu/Ile side chains (b = 3.9) to the side chain of alanine.

10. The “Monoisotopic Mass” option is useful in the mass prediction of small peptides (<3000 Da) that can often be isotopically resolved on mass spectrometers. The (M+H)<sup>+</sup> option will calculate all peptide masses with an extra hydrogen atom, to give values for singly charged peptides as found in electrospray and MALDI-TOF mass spectrometers.
11. The program does take into account most types of N- or O-linked glycosylation or other complex modifications like glycan phosphatidyl-inositol anchors because of their unpredictable heterogeneity. However, the discrete O-GlcNAc and C-mannosylation modifications are considered.
12. The PeptideMass program does not *predict* new potential posttranslational modifications in user-entered sequences, and thus does not consider these in mass calculations. However, one can use the PROSITE database (<http://www.expasy.org/prosite/>), e.g., by using the ScanProsite tool (<http://www.expasy.org/tools/scanprosite/>, [35]) to predict the presence of posttranslational modifications in a sequence. A list of modifications documented in the PROSITE database can be found at: <http://www.expasy.org/prosite/browse/>. The ExPASy tools page further contains a section of prediction tools for sequence-based posttranslational modifications (e.g., for prediction of signal sequences, tyrosine sulfation, glycosylation, and so on). If PMF data is available for a protein, FindMod or GlycoMod can be used to predict potential posttranslational modifications.
13. PeptideCutter does not take into consideration any kind of modification, neither of the protein sequence (post-translational) nor of modifications evoked by the cleavage (e.g., conversion of Met into homoserine lactone upon cleavage with CNBr). If the user is interested in a more detailed analysis of the resulting peptide, we recommend using the PeptideMass program (see **Subheading 3.3**).
14. An amino acid scale is defined by a numerical value assigned to each type of amino acid. The most frequently used scales are hydrophobicity scales, most of which were derived from experimental studies on partitioning of peptides in apolar and polar solvents, with the goal of predicting membrane-spanning segments that are highly hydrophobic, and secondary structure conformational parameter scales. In addition, many other scales exist, based on different chemical and physical properties of the amino acids.
15. Protein pI and Mw in TagIdent/AACompIdent are calculated as described for Compute pI/Mw (see **Subheading 3.1**).
16. Care must be taken in the use of pI and Mw estimates from 2-D gels as part of protein-identification strategies. Windows around these estimates that are too narrow can exclude the correct identification from the list of candidate identifications. As a general rule, we use windows of pI ± 0.5 units for proteins from bacteria and yeast, and pI ± 1.0 units for mammalian proteins. We generally use a Mw window of ± 20%, but for proteins larger than 60,000 Da, a window of ± 10% is sufficient, because of the more accurate estimations (in percentage terms) that can be made with higher mass proteins on gels. If proteins are thought to be highly posttranslationally modified, very large pI and/or Mw windows may be needed.
17. TagIdent is extremely useful for searching proteins in the database for the presence of sequence tags, as it can search in a species-specific manner and with pI and Mw param-



eters. This is a valuable alternative to Basic Local Alignment Search Tool (BLAST) (36), which, although it can now be used with sequences as short as four amino acids (by increasing the E value), does not allow the restriction of a search with pI and Mw parameters.

18. Although TagIdent is certain to find all sequences present in Swiss-Prot and TrEMBL, it is still possible that the result misses some existing proteins if their coding sequences (CDS) have not been annotated in the nucleotide sequence database by the submitters. This would prevent a protein entry from being included in TrEMBL, and the sequence will only be integrated in Swiss-Prot if an annotator detects this previously unannotated CDS during the manual annotation process. The same is true if the original EMBL entry has an incorrectly predicted initiation site that has not yet been corrected by a Swiss-Prot annotator.
19. If you specify parameters that generate an extremely large TagIdent output (>1 megabyte), only the first 1000 lines will be sent by e-mail. This is to avoid problems that can arise when large messages arrive at some e-mail sites.
20. If AA analyses yield unreliable data for one or more amino acids, such values can be ignored and matching undertaken only with "good" amino acids, using the AACompIdent free constellation (37). The free constellation also allows the user to modify the bias and weight for each AA, if desired.
21. When calibration proteins are used, AACompIdent compares the experimental composition of the protein against the theoretical composition in the Swiss-Prot database to create a factor set. This factor set is then applied to the experimental composition of the unknown protein before it is matched against the Swiss-Prot database. Use of calibration proteins can increase identification efficiency dramatically, and is advised wherever possible. Note, however, that calibration proteins should be electrophoretically prepared in the same manner as unknown proteins, and subjected to AA analysis in parallel with unknown proteins. It is also essential that the complete sequence of calibration proteins be in the Swiss-Prot database, as calibration cannot be done if only a fragment of the calibration protein sequence is available.
22. Protein AA composition and Mw are highly conserved across species boundaries and serve as useful parameters for cross-species protein identification (15,16). Protein pI is, however, poorly conserved between species. Cross-species protein identification in AACompIdent can be done by specifying "ALL" for the species of interest, or specifying the Swiss-Prot species code of a well-defined organism that is closely related to the species under study. It must be noted that high-confidence cross-species protein identification usually requires peptide mass data or sequence as well as AA composition (see **Subheading 4.4.**).
23. AACompSim automatically uses the theoretical pI and Mw of the specified protein in the matching procedure. The pI and Mw are calculated as in Compute pI/Mw (see **Subheading 3.1.**).
24. Default windows of  $pI \pm 0.25$  and  $Mw \pm 20\%$  are used by AACompSim in matching; however, matches undertaken without restriction to these windows are also included in the program output.
25. While peptide masses for any protein type are not as well conserved across species boundaries as other parameters (38), they can be used for cross-species protein identification in conjunction with, for example, amino acid composition (16,17).
26. The algorithm has no major difficulties when working with very crowded spectra (i.e., with a high number of input masses, >100) and with a large number of theoretical masses. Consequently, increasing the number of possible peptides by taking into account combinations of missed cleavages, posttranslational modifications, alternative splicing, and chemical modifications, is conceivable. However in addition to increasing the number of

true-positive peptide matches, there is also a risk of increasing the number of false-positive hits.

27. Aldente does not do any *de novo* prediction of posttranslational modifications on proteins. All modified peptides shown in the results will be the verification of an event documented in Swiss-Prot. However, Aldente can match peptides whose modifications are documented in Swiss-Prot as “potential” or “by similarity,” and thus allows predicted post-translational modifications to be validated.
28. If no number is specified for pI and/or Mw, the parameter will not be used in the Aldente score.
29. The peptide masses should be specified with a high precision and can be supplied in the form (one per line) or uploaded in a file in plain-text format or in the .pkm (GRAMS) or .dta (SEQUEST) or .pkt (Data Explorer) formats used by peak identification software. These formats are described in detail in the online documentation on ExPASy.
30. Multiple selection is possible by holding down the “Ctrl” key. We define “single species matching” where you, for example, have proteins from *E. coli* that you then match against only the *E. coli* proteins in the database. This is a good approach to use when the organism you are working with is molecularly well defined, or ideally, the subject of a genome project. If the source of your proteins is not molecularly well defined, it is best to do “cross-species matching.” Thus, for example, if you are working with proteins from *Candida albicans*, you may wish to either match your proteins against all proteins from fungi or against the fully sequenced yeast *Saccharomyces cerevisiae*. Note, however, that when cross-species matching, protein pI is frequently poorly conserved, but protein mass is generally very well conserved (38). You should take this into consideration when setting your pI and Mw values. On the contrary, peptide masses are not well conserved across species boundaries. The poor conservation of peptide mass data is expected, as a single-amino acid substitution in any peptide can drastically change its mass.
31. Mass spectrometers typically have a mass-dependent error associated with mass measurements, which cannot be uniformly expressed in Daltons. The use of ppm can therefore be more accurate. If both  $\Delta$ Da and  $\Delta$ ppm have been specified, the program will combine them with a logical OR.
32. Both MALDI and electrospray machines are now capable of achieving single decimal point mass resolution; however, this may depend on the care that has been taken in machine calibration and use of internal standards. We recommend the use of a tolerance of 0.2 Da or 200 ppm or better whenever it is possible. Electrospray ionization (ESI)-TOF mass spectrometers or MALDI-TOF apparatus equipped with delayed extraction and ion reflectors are ideal for this, since most can deliver monoisotopic masses below  $\pm 40$  ppm, when two-point internal calibration is used.
32. Both MALDI and ES machines are now capable of achieving single decimal point mass resolution; however, this may depend on the care that has been taken in machine calibration and use of internal standards. We recommend the use of a tolerance of 0.2 Da or 200 ppm or better whenever it is possible. ESI-TOF mass spectrometers or MALDI-TOF apparatus equipped with delayed extraction and ion reflectors are ideal for this, because most can deliver monoisotopic masses below  $\pm 40$  ppm, when two-point internal calibration is used.
33. Aldente supports any user-defined chemical modifications to amino acids, given the locus where the modification should appear and the chemical formula of the product to add/remove on this locus. Two types of modifications can be applied. Use fixed modification whenever amino acids should be modified, and specify in the tolerance field the number of exceptions allowed. For example, for carboxymethylation on cysteine (CAM) with fixed option and a tolerance of 1, the program will generate the theoretical peptide with all cysteines modified, and peptides with all but one cysteine modified. On the contrary, use

variable modification when residues may or may not be modified, and specify in the tolerance field the maximum number of modified residues expected. For example, for methionine oxidation (MSO) with variable option and a tolerance of 2, the program will generate the theoretical peptides with 0, 1, or 2 methionines modified. The program supports every combination of possible modifications (or PTMs, *see Note 11*) occurring on the same locus. In an advanced mode, the user can specify for each modification a factor to be applied on the score to penalize peptides with (un)modified locus.

34. For all types of PTMs annotated in Swiss-Prot, two modes of peptide modification are available: fixed or variable. In fixed mode, the program will generate the theoretical peptide with all modified loci. In variable mode, the program will generate theoretical peptides with all combinations of modified or unmodified loci. If several PTMs (or chemical modifications; *see Note 33*) are possible at the same sequence position, the program will generate the theoretical peptides corresponding to every possible combination.
35. Swiss-Prot annotation distinguishes between experimentally proven and computationally predicted posttranslational modifications, as well as those inferred by similarity (5). The Swiss-Prot document “How is biochemical information assigned to sequence entries” (<http://www.expasy.org/txt/annbioch.txt>) describes how these nonexperimental qualifiers are used.
36. The program performs two runs. First, it keeps the *n* best proteins within the user protein and peptide mass tolerance. In a second pass, the program finds the best line fitting the maximum of hits (matching peptide masses) only for those *n* best proteins. Because of a higher precision in the second step, it is possible that some hits from the first step are removed, which means that the number of hits in the output result can become smaller than the number of hits you requested to see.
37. In the case of a manually entered sequence, the user is required to specify the biological source of the query protein. This information is used to determine whether certain PTMs are likely to occur in the sequence.
38. Users should avoid using peptide masses known to be from autodigestion of an enzyme (e.g., trypsin!), or other artifactual peaks (e.g., matrix peaks). If you are not sure whether your set of masses contains such peaks, you may use FindPept (*see Subheading 4.8.*) to detect them.
39. Cysteine residues in proteins are usually subjected to reduction and then alkylation with different reagents before they are used to generate peptides. Such a reactant can be specified here as one of iodoacetamide, iodacetic acid, or 4-vinyl pyridene. If no reactant has been used and the protein has undergone polyacrylamide gel separation, acrylamide adducts are to be expected on part of the free cysteines, so the reactant “acrylamide” should be chosen.
40. You can request for all methionines in theoretical peptides to be oxidized. If this option is selected, the program will modify the theoretical masses of Met-containing peptides accordingly and consider both peptides with unmodified methionines and peptides with modified methionines. Note that proteins prepared by gel electrophoresis often show this modification.
41. Other possible causes for nonmatched masses include signal autosuppression on MALDI or ESI and modifications on peptides (natural or artifactual; FindMod and GlycoMod can identify some of the former).
42. A set of 20 different proteases of different sources are available, and their sequence and cleavage rules are used to detect specific cleavage, autolysis peaks, and also to digest frequent contaminants.
43. For example, the cleavage rule for trypsin that is widely used in identification programs is “cleavage after Lys or Arg, except if followed by Pro.” Experimental data (9) show, how-

ever, a much more complex specificity pattern—in particular, the negative influence of charged residues immediately adjacent to the targeted Lys or Arg. This rule has been used to refine the cleavage prediction and is available in ExPASy tools as “Trypsin, higher specificity.”

44. Protease specificity may be adversely affected by chemical alterations in the enzyme. Solution conditions and temperature can affect specificity. Longer incubation times increase the yield of unspecific cleavage. Specificity can be increased by the addition of inhibitors, short incubation times, and an appropriate ratio of protease to substrate, also to limit protease autolysis (39).
45. The program can take into account a notable number of posttranslational modifications of discrete mass (it shares the database used by the FindMod and Aldente tools), plus several chemical modifications resulting from the experimental treatment of the protein. Most of these modifications cannot be expected to be applied quantitatively, except some chemical treatments such as carbamidomethylation (by iodoacetamide) or methanolic esterification. Therefore, FindPept applies them combinatorially on each peptide, assuming that no two modifications can simultaneously occur on a single amino acid.
46. The combined effect of unspecific cleavage and posttranslational and artifactual modifications gives rise to a huge number of possible peptides that may match the input masses. Since unrelated peptides may have very similar masses, false attributions can be limited only if the measurements are done with a high precision. The experimental mass error should not exceed 20–25 ppm. Even then, several attributions are sometimes made for a single mass. In these cases, the most likely peptides are those that have a specific cleavage site at one of their ends (shown in the output with a red slash). Additional evidence should usually be sought by experimental methods to obtain a conclusive determination.
47. Side-chain and C-terminal carboxylic acid groups can be esterified to methyl esters. Additionally, the box “N-Acetylation and N-formylation” can be used to submit the N-terminal residue of the protein to possible acetylation or formylation. Both the modified and unmodified versions of N-terminal peptides are examined.
48. An intermediate page may appear instead of the output page if you have submitted the accession number of a Swiss-Prot/TrEMBL entry that contains several chains or mature peptides and/or a cleaved initiator methionine at the beginning of the sequence. You should select the sequence to be analyzed: either a single chain, the uncleaved precursor, or the sequence with the initiator methionine added. If you suspect that the characterized protein has not matured as expected *in vivo*, you should consider the precursor sequence for the analysis. The numbering of the residues will be the same as in the original Swiss-Prot/TrEMBL entry. If an initiator methionine is added before the beginning of the sequence, it is assigned the position 0.
49. The considered contaminant keratins are human cytokeratins 1, 2, 9, and 10 (Swiss-Prot accessions P04264, P35908, P35527, and P13645), which have been determined to be abundant in skin and dandruff (40) and are often encountered as contaminants in biological samples handled in the laboratory.

## Acknowledgments

This work was supported by the National Institutes of Health (NIH) grant 1 U01 HG02712-01 and by the Swiss Federal Government through the Federal Office of Education and Science.

## References

1. Boeckmann, B., Bairoch, A., Apweiler, R., et al. (2003) The Swiss-Prot protein knowledge-base and its supplement TrEMBL in 2003. *Nucleic Acids Res.* **31**, 354–370.

2. Gasteiger, E., Gattiker, A., Hoogland, C., Ivanyi, I., Appel, R. D. and Bairoch, A. (2003). ExPASy—the proteomics server for in-depth protein knowledge and analysis. *Nucleic Acids Res.* **31**, 3784–3788.
3. Apweiler, R., Bairoch, A., Wu, C. H., et al. (2004) UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res.* **432**, D115–D119.
4. Jung, E., Gasteiger, E., Veuthey, A.-L., and Bairoch A. (2001) Annotation of glycoproteins in the SWISS-PROT database. *Proteomics* **1**, 262–268.
5. Farriol-Mathis, N., Garavelli, J. S., Boeckmann, B., et al. (2004), Annotation of post-translational modifications in the Swiss-Prot knowledgebase. *Proteomics*, in press.
6. Gill, S. C. von Hippel, P. H. (1989) Calculation of protein extinction coefficients from amino acid sequence data. *Anal. Biochem.* **182**, 319–326.
7. Kyte, J., and Doolittle, R. F. (1982) A simple method for displaying the hydropathic character of a protein. *J. Mol. Biol.* **157**, 105–132.
8. Wilkins, M. R., Lindskog, I., Gasteiger, E., et al. (1997) Detailed peptide characterization using PEPTIDEMASS—a World-Wide-Web-accessible tool. *Electrophoresis* **18**, 403–408.
9. Keil, B. (1992) *Specificity of proteolysis*. Springer-Verlag Berlin-Heidelberg-New York, p. 335.
10. Wilkins, M. R., Gasteiger, E., Sanchez, J.-C., Appel, R. D., and Hochstrasser, D. F. (1996) Protein identification with sequence tags. *Curr. Biol.* **6**, 1543–1544.
11. Wilkins, M. R., Gasteiger, E., Tonella, L., et al. (1998) Protein identification with N- and C-terminal sequence tags in proteome projects. *J. Mol. Biol.* **278**, 599–608.
12. Ashburner, M., Ball, C. A., Blake, J. A., et al. (2000), Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat. Genet.* **25**, 25–29.
13. Wilkins, M. R., Pasquali, C., Appel, R. D., et al. (1996) From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis. *Bio/Technology* **14**, 61–65.
14. Wilkins, M. R., Ou, K., Appel, R. D., et al. (1996) Rapid protein identification using N-terminal “sequence tag” and amino acid analysis. *Biochem. Biophys. Res. Commun.* **221**, 609–613.
15. Hobohm, U. and Sander, C. (1995) A sequence property approach to searching protein databases. *J. Mol. Biol.* **251**, 390–399.
16. Cordwell, S. J., Wilkins, M. R., Cerpa-Poljak, A., et al. (1995) Cross-species identification of proteins separated by two-dimensional gel electrophoresis using matrix-assisted laser desorption time of flight mass spectrometry and amino acid composition. *Electrophoresis* **16**, 438–443.
17. Wheeler, C. H., Berry, S. L., Wilkins, M. R., et al. (1996) Characterisation of proteins from 2-D gels by matrix-assisted laser desorption mass spectrometry and amino acid compositional analysis. *Electrophoresis* **17**, 580–587.
18. Wilkins, M. R., Gasteiger, E., Wheeler, C., et al. (1998) Multiple parameter cross-species protein identification using MultiIdent—a world wide web accessible tool. *Electrophoresis* **19**, 3199–3206.
19. Wilkins, M. R., Gasteiger E., Gooley, A. A., et al. (1999) High-throughput mass spectrometric discovery of protein post-translational modifications. *J. Mol. Biol.* **289**, 645–657.
20. Hulo, N., Sigrist, C. J., Le Saux, V., et al. (2004) Recent improvements to the PROSITE database. *Nucleic Acids Res.* **32**, D134–D137.
21. Henikoff, S., and Henikoff, J. G. (1993) Performance evaluation of amino acid substitution matrices. *Proteins* **17**, 49–61.
22. Cooper, C. A., Gasteiger, E., and Packer, N. (2001) GlycoMod—a software tool for determining glycosylation compositions from mass spectrometric data. *Proteomics* **1**, 340–349.



23. Cooper, C. A., Gasteiger, E., and Packer, N. (2003) Predicting glycan composition from experimental mass using GlycoMod. In: (Conn, P.M., ed.) *Handbook of Proteomic Methods*, (Humana, Totowa, NJ: pp. 225–231.
24. Gattiker, A., Bienvenut, W. V., Bairoch, A., and Gasteiger, E. (2002) FindPept, a tool to identify unmatched masses in peptide mass fingerprinting protein identification. *Proteomics* **2**, 1435–1444.
25. Bjellqvist, B., Hughes, G., Pasquali, C., et al. (1993) The focusing positions of polypeptides in immobilized pH gradients can be predicted from their amino acid sequences. *Electrophoresis* **14**, 1023–1031.
26. Bjellqvist, B., Basse, B., Olsen, E., and Celis, J. E. (1994) Reference points for comparisons of two-dimensional maps of proteins from different human cell types defined in a pH scale where isoelectric points correlate with polypeptide compositions. *Electrophoresis* **15**, 529–539.
27. Hoogland, C., Sanchez, J.-C., Tonella, L., et al. (2000) The 1999 SWISS-2DPAGE database update. *Nucleic Acids Res.* **28**, 286–288.
28. Nielsen, H., Engelbrecht, J., Brunak, S., and von Heijne, G. (1997) Identification of prokaryotic and eukaryotic signal peptides and prediction of their cleavage sites. *Protein Eng.* **10**, 1–6.
29. Bachmair, A., Finley, D., and Varshavsky, A. (1986) In vivo half-life of a protein is a function of its amino-terminal residue. *Science* **234**, 179–186.
30. Gonda, D. K., Bachmair, A., Wunning, I., Tobias, J. W., Lane, W. S., and Varshavsky, A. J. (1989) Universality and structure of the N-end rule. *J. Biol. Chem.* **264**, 16,700–16,712.
31. Tobias, J. W., Shrader, T. E., Rocap, G., and Varshavsky, A. (1991) The N-end rule in bacteria. *Science* **254**, 1374–1377.
32. Ciechanover, A. and Schwartz, A. L. (1989) How are substrates recognized by the ubiquitin-mediated proteolytic system? *Trends Biochem. Sci.* **14**, 483–488.
33. Guruprasad, K., Reddy, B. V. B., and Pandit, M. W. (1990) Correlation between stability of a protein and its dipeptide composition: a novel approach for predicting in vivo stability of a protein from its primary sequence. *Protein Eng.* **4**, 155–161.
34. Ikai, A. J. (1980) Thermostability and aliphatic index of globular proteins. *J. Biochem.* **88**, 1895–1898.
35. Gattiker, A., Gasteiger, E., and Bairoch, A. (2002) ScanProsite: a reference implementation of a PROSITE scanning tool. *Applied Bioinform.* **1**, 107–108.
36. Altschul, S. F., Gish, W., Miller, W., Myers, E. W., and Lipman, D. J. (1990) Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410.
37. Golaz, O., Wilkins, M. R., Sanchez, J.-C., Appel, R. D., Hochstrasser, D. F., and Williams, K. L. (1996) Identification of proteins by their amino acid composition: an evaluation of the method. *Electrophoresis* **17**, 573–579.
38. Wilkins, M. R. and Williams, K. L. (1997) Cross-species protein identification using amino acid composition, peptide mass fingerprinting, isoelectric point and molecular mass: a theoretical evaluation. *J. Theor. Biol.* **186**, 7–15.
39. Hara, S., Rosenfeld, R., and Lu, H. S. (1996) Preventing the generation of artifacts during peptide map analysis of recombinant human insulin-like growth factor-I. *Anal. Biochem.* **243**, 74–79.
40. Parker, K. C., Garrels, J. I., Hines, W., et al. (1998) Identification of yeast proteins from two-dimensional gels: working out spot cross-contamination. *Electrophoresis* **19**, 1920–1932.



