David Fenyö
Jun Qin*
Brian T. Chait

The Rockefeller University,
New York, NY, USA

Protein indentification using mass spectrometric information

In an effort to gain an understanding of the value of the information in different mass spectrometric measurements for protein identification, the genome of *Saccharomyces cerevisiae* was studied *in silico*. We calculate how constraining the knowledge of the mass of a proteolytic peptide is as a function of mass and mass accuracy. We also assess the value for protein identification of additional information concerning a proteolytic peptide, including the presence or absence of a given amino acid, the number of exchangeable hydrogens, the *N*-terminal sequence, and the masses of mass spectrometrically produced fragment ions. Knowledge of the relative value of these different constraints is useful in the design of efficient protein identification experiments. Finally, we describe a software tool, PepFrag, for searching protein and DNA sequence databases that can use different types of mass spectrometric information to restrict the search.

1 Introduction

The constantly increasing stream of high quality DNA sequence data from the various genome projects has made mass spectrometry (MS) a preferred method for identifying unknown proteins. Mass spectrometric protein identification has been applied to the study of biological problems, such as apoptosis [1], human cancer [2–4], and also to elucidate the components of several multi-protein complexes [5–7] as well as for large-scale identification of the proteins in organisms with fully sequenced genomes [8–10]. Usually, the proteins of interest are isolated and then separated from each other by gel electrophoresis [11]. The separated proteins are digested by a protease with high specificity (usually trypsin) either directly in the gel or on a membrane subsequent to electroblotting. After digestion, the peptides are extracted from the gel or membrane, and the resulting peptide mixture is analyzed to obtain an MS peptide map using either matrix-assisted laser desorption/ionization mass spectrometry or electrospray ionization mass spectrometry (the latter with or without prior chromatographic separation). Protein identification is achieved by: (i) calculating the masses of all the possible enzymatic cleavage products of all proteins with known sequence, (ii) comparing the measured masses to these calculated masses, and (iii) selecting the protein that gives the best agreement [12–27]. If an unambiguous identification cannot be made, it is necessary to constrain the search with additional information, *e.g.*, an MS peptide map from a second proteolytic digest and information about the intact protein (*e.g.*, apparent mass on an SDS-gel, isoelectric point, *N*-terminal sequence and amino acid composition [28–29]). Useful mass spectrometric information about the proteolytic peptides includes knowledge of the presence or absence of particular amino acids, the number of exchangeable hydrogens [19], the *N*-terminal sequences [23], and the masses of MS fragment ions [30–40]. In this paper we discuss the value of the different constraints that can be obtained with MS information and also describe Pepfrag, a software tool for searching protein and DNA sequence databases using different types of MS information to restrict the search.

2 Material and methods

The publicly available *S. cerevisiae* genome containing 6129 predicted open reading frames was used for most of the calculations [41–42]. For *H. sapiens* the calculations were performed on the collection of human proteins in SWISS-PROT release 34 [43] with an extrapolation which assumed that the human genome contains 100 000 proteins and the assumption that the distribution of amino acids is the same for proteins not yet sequenced as for those already sequenced. The code was written in C and the calculations were performed on Silicon graphics Indigo II (R4400), Origin 200 (2 × RA 10 000), and Dell Dimension XPS (200 MHz Pentium Pro) computers. The PepFrag searches were performed with SWISS-PROT release 34, GenPept release 101, and dbEST release 101 [44]. The mass spectra were collected with a matrix-assisted laser desorption/ionization-ion trap-mass spectrometrometer (MALDI-IT-MS) developed at the Rockefeller University [45] and an electrospray ionization ion trap mass spectrometer (LCQ) manufactured by Finnigan MAT Thermoquest (San Jose, CA).

3 Results and discussion

Protein identification using MS peptide mapping usually requires the masses of several proteolytic peptides (sometimes together with additional constraints) to unambiguously identify a single protein. In an effort to evaluate the relative utility of different constraints, we studied the number of predicted open reading frames

Correspondence: Dr. David Fenyö, The Rockefeller University, 1230 York Avenue, New York, NY 10021, USA (Tel: +212–327–8848; Fax: +212–327–7547; E-mail: fenyod@rockvax.rockefeller.edu)

Abbreviations: **EST,** expressed sequence tag; **MALDI-IT-MS,** matrix assisted laser decorption/ionization-ion trap-mass spectrometry; **ORF,** open reading frame

Keywords: Mass spectrometry / Protein identification / Peptide mapping / *Saccharomyces cerevisiae*

* Current address: National Institute of Heart, Lung, and Blood, NIH, Bethesda, MD 20892

0173-0835/98/0606-0998 $17.50+.50/0

(ORF, which we will here refer to as proteins) in *S. cervisiae* (out of a total of 6129) that match different constraints.

3.1 Information content in one tryptic peptide

Figure 1 shows the distribution of the number of proteins as a function of tryptic peptide mass for a few representative organisms. The tryptic peptides were assumed to result either from exhaustive cleavages of the protein or cleavage that leave one possible trypsin cleavage site intact. Inspection of Fig. 1A shows, for example, that there are, respectively, 270 and 108 proteins in *S. cervisiae* that have tryptic peptides with masses of 1000.0 ±0.5 Da and 2000.0±0.5 Da – *i.e.*, the mass of a single tryptic peptide (in the mass range 1000–2000 Da with accuracy ±0.5 Da) reduces the number of matching proteins by a factor of 20–60. Assuming that the information from different tryptic peptides is independent, an approximation in the overal reduction of the number of matching proteins for the peptide map can be obtained by multiplying the reductions for the individual peptides. This means that, in an ideal case, the masses of a few peptides are sufficient for unambiguous identification of a protein. The number of matching proteins decreases rapidly as a function of increasing peptide mass, showing that the higher the mass of the peptide, the better its value as a constraint for protein identification. For other enzymes that cut more sparsely than trypsin, the information content in the mass of each proteolytic peptide is greater although fewer peptides will be produced (data not shown). As expected, smaller bacterial genomes contain fewer proteins that match a single peptide (Fig. 1A) and less information is necessary for unambiguous protein identification.

The human genome, on the other hand, is much larger than that of *S. cervisiae* so that a single tryptic peptide is considerably less constraining (Fig. 1B). At present, the situation for human protein identification is even more challenging because the human genome has not yet been fully sequenced and a high quality protein sequence is only available for a few percent of the human genes. However, partial single pass cDNA sequences (so-called expressed sequence tags (ESTs) [46]) are available in public databases for many human genes [44]. EST databases can be used for protein identification, but require higher-quality mass spectrometric data than that required for searching protein sequence databases. The EST sequences contain many errors and only cover a part of the gene, making it necessary to use high-quality mass spectrometric fragmentation information from single peptides in the search [30, 31] and to perform searches with information about several different peptides until a hit is registered. Usually, the reading frame cannot be determined with certainty for these partial gene sequences, making it necessary to translate the EST sequence into all six reading frames, resulting in a further increase in the level of noise.

3.2 Mass accuracy

Improving the mass accuracy is one way of increasing the information content in a measurement [22]. Figure 2 shows an example of how the number of *S. cerevisiae* proteins containing a tryptic peptide of 2359.29±ΔDa

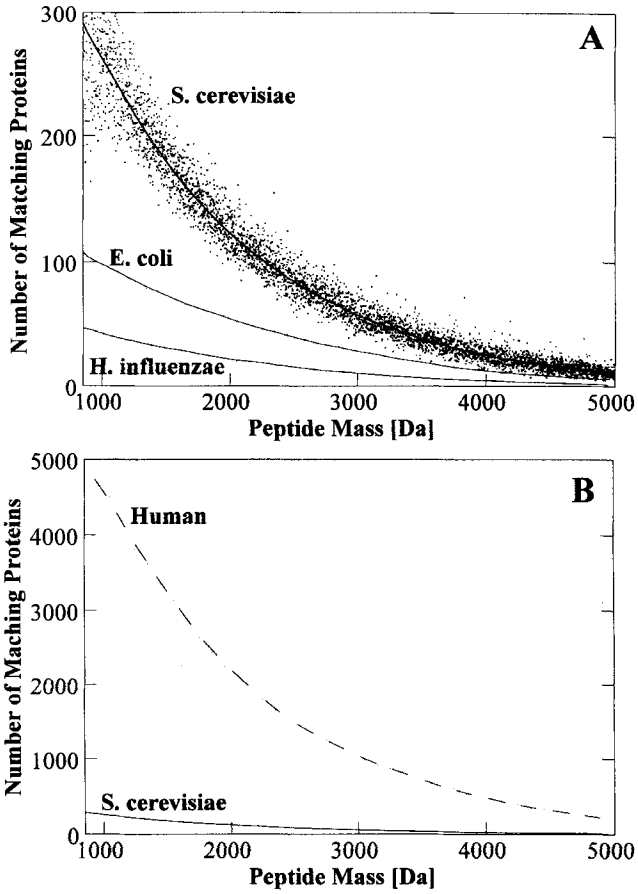

Figure 1. (a) The distribution of the number of proteins (strictly, the number of predicted ORFs) in *S. cerevisiae, E. coli* and *H. influenzae* as a function of tryptic peptides mass (mass accuracy ±0.5 Da). The tryptic peptides were assumed to result from complete cleavage or having one possible trypsin cleavage site intact. For *S. cerevisiae* the number of proteins at every mass unit is shown together with a smooth curve fitted to the data. For *E. coli* and *H. influenzae* only the smooth fits are shown for clarity. (B) The estimated distribution of the number of human proteins as a function of tryptic peptide mass (mass accuracy ±0.5 Da).

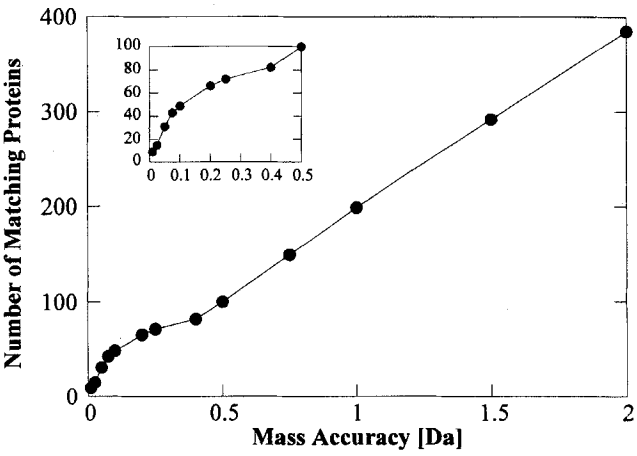

Figure 2. The number of proteins from *S. cerevisiae* that contain one tryptic peptide with an isotopically averaged mass of 2359.29 Da as a function of mass accuracy. The tryptic peptides were assumed to result from complete cleavage or having one possible trypsin cleavage site intact. The total number of ORFs, predicated from the *S. cerevisiae* genome, is 6129.

varies as a function of the accuracy of the mass determination, Δ. For example, changing the mass accuracy from 2 Da to 0.5 Da gives a fourfold improvement for a single peptide. For $\Delta>0.4$ Da the number of matches varies linearly with the mass accuracy, while for $0.1<\Delta<0.4$ the curve has a plateau caused by the uneven distribution of peptide masses due to the fact that amino acids are composed of only a few atom types all having near-integer masses. For $\Delta<0.1$ Da the number of matches decreases rapidly.

3.3 Presence or absence of particular amino acids

In addition to the mass of the proteolytic peptide, it is valuable to find out if an amino acid is either present or absent in the peptide. This information can be obtained, for example, from immonium ions produced in gas-phase fragmentation or by chemical modification of the protein (*e.g.*, oxidation of methionines or alkylation of cysteines). Figure 3A shows the distribution of the number of proteins as a function of tryptic peptide mass when we know that the peptide contains a certain amino acid. For example, knowing that a peptide of mass 2000±0.5 Da contains a cysteine reduces the number of matching *S. cerevisiae* proteins by a factor of five. The improvement ratio (Fig. 3B) is defined as the inverse of the reduction in the number of matching proteins – *i.e.* it is the ratio between the number of matching proteins satisfying an additional constraint divided by the number of matching proteins when the additional constraint is not considered. Although the information that an amino acid is either present or absent in the peptide does not give a dramatic improvement for one peptide, it can be constraining when applied to many peptides in the peptide map. For example, assume that there are 10 tryptic peptides in the mass range 2000–4000 Da and, to simplify the calculation, take the average improvement ratio over the mass range (0.3 for Cys, 0.4 for Met, 0.6 for Tyr, and 0.9 for Ser; Fig. 3B). The reduction in the number of matching proteins, when the only information that is utilized is that a specific amino acid is present, is estimated* to be a factor of 40, 40, 20, and 3 for Cys, Met, Tyr, and Ser, respectively. When information about both the presence and absence of an amino acid is used, the reduction in the number of matching proteins is estimated* to be a factor of 400, 800, 800, and 30 for Cys, Met, Tyr, and Ser, respectively. In this mass range, the highest reduction in the number of matching proteins is obtained for the less abundant amino acids. By using chemical modifications (*e.g.*, oxidation and alkylation), it may be possible (in favorable cases) to count the number of a particular amino acid in a peptide – providing an even stronger constraint. In this case, the reduction in the number of matching proteins will be larger if it proves possible to count a more abundant amino acid.

* For a constant improvement ratio (p) and a large number of proteolytic peptides (N), the number of peptides containing a particular amino acid can be approximated by Np. When only the information that the peptide contains a particular amino acid is considered, the reduction in the number of matching proteins can be written as p^{-Np}. The largest reduction is obtained for $p=1/e$ (dashed line in Fig. 3B). When the information about both the presence and absence of an amino acid is used, the reduction in the number of matching proteins can be written as $p^{-Np}(1-p)^{-N(1-p)}$. In this case, the largest reduction is obtained for $p=1/2$ (solid line in Fig. 3B).

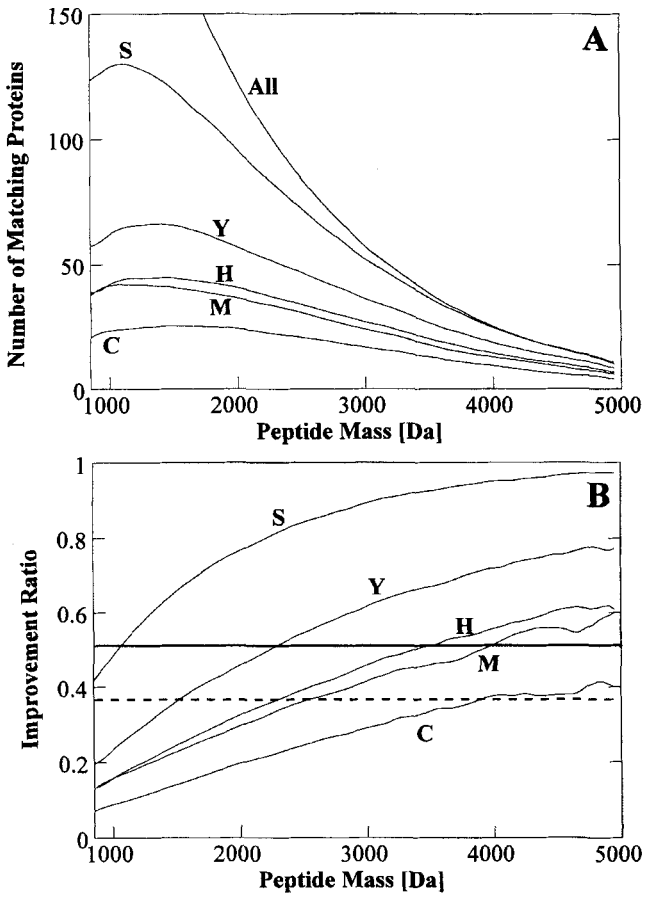

Figure 3. (A) The distribution of the number of *S. cerevisiae* proteins as a function of tryptic peptide mass (All). Also, the number of matching proteins satisfying the additional constraint that the tryptic peptide contains cysteine (C), methionine (M), serine (S), tyrosine (Y), and histidine (H) is shown as a function of tryptic peptide mass. (B) The distribution of the improvement ratio as a function of tryptic peptide mass. The improvement ratio is defined as the number of matching proteins satisfying an additional constraint (*e.g.* that the tryptic peptide contains Cys) divided by the number of matching proteins when the additional constraint is not considered. The dashed line shows the improvement ratio giving the largest reduction in the number of matching proteins for a peptide map when only the information that peptides contain a particular amino acid is considered. The solid line represents the improvement ratio giving the largest reduction in the number of matching proteins for a peptide map when both the presence and absence of an amino acid is used.

3.4 *N*-terminal amino acid sequence

It is feasible to perform one or more steps of Edman chemistry on a peptide mixture, mass-analyze the mixture, and obtain information concerning the *N*-terminal sequence of the more abundant peptides [47]. This information can be used to constrain a database search [23]. Figure 4A shows the distribution of the number of proteins as a function of tryptic peptide mass when we know the *N*-terminal amino acid of the peptide for a few representative amino acids. Knowledge of one *N*-terminal amino acid reduces the number of matching proteins by a factor 6–80, depending on the amino acid (Fig. 4B). On average, the reduction is a factor of 13 for one peptide. This information provides a strong constraint since it may be possible to determine the

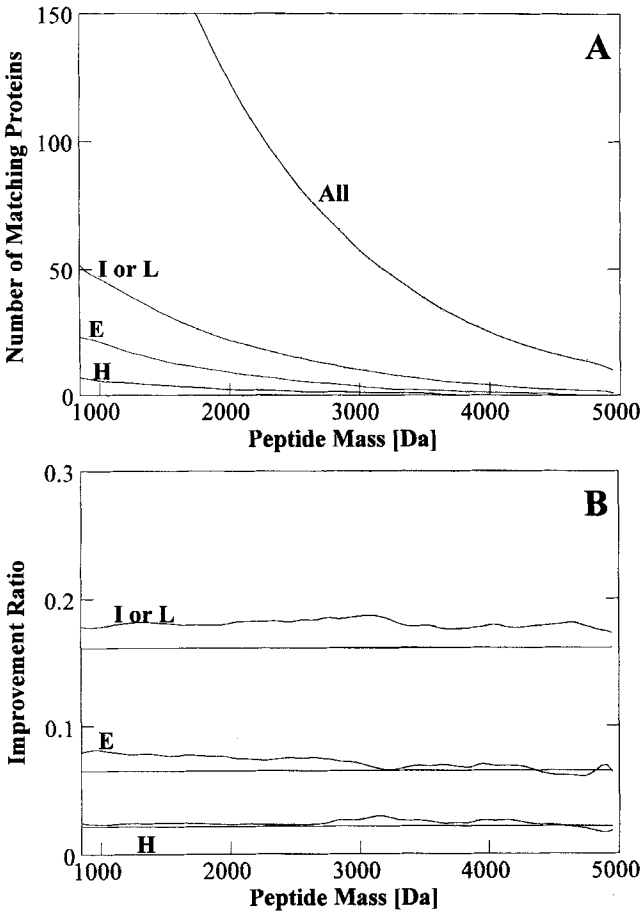

Figure 4. (A) The distribution of the number of *S. cerevisiae* proteins with the *N*-terminal residue of the tryptic peptide being unknown (All), isoleucine or leucine (I or L), glutamic acid (E), and histidine (H) as a function of tryptic peptide mass. (B) The distribution of the improvement ratio as a function of tryptic peptide mass. The improvement ratios that would be obtained for random amino acid distribution are shown as straight lines.

N-terminal amino acid for several of the tryptic peptides in a peptide map at the same time [23]. Similar information can be obtained by exopeptidase digestion of the *C*-terminal of the tryptic peptides. However, the information content in the first *C*-teminal amino acid is more limited than the *N*-terminal amino acid because the *C*-terminal residue for tryptic peptides can only be arginine or lysine (with the single exception of the *C*-terminal of the protein). On the other hand, for enzymes cleaving on the *N*-terminal side of particular amino acids (*e.g.*, endopeptidase Asp-N), it may be preferable to sequence the *C*-terminal of the resulting proteolytic peptides.

3.5 Hydrogen/deuterium exchange

The number of exchangeable hydrogens in proteolytic peptides can be determined by first obtaining a mass spectrum of the peptide map, exposing the peptide mixture to D_2O, and subsequently collecting another mass spectrum [19]. Figure 5 shows the distribution of the number of *S. cerevisiae* proteins with mass 1000.0 ±0.5 Da (solid line) and 2000.0±0.5 Da (dashed line) as a function of the number of exchangeable hydrogens. For example, a tryptic peptide of mass 1000.0±0.5 Da matches 270 *S. cerevisiae* proteins while only 53 of these

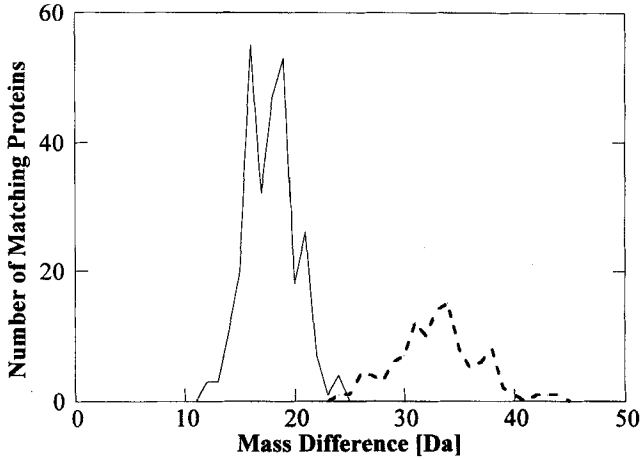

Figure 5. The distribution of the number of *S. cerevisiae* proteins with mass 1000.0±0.5 Da (solid line) and 2000.0±0.5 Da (dashed line) as a function of the difference between the tryptic peptide mass before and after exhaustive H/D exchange. The total number of *S. cerevisiae* proteins that contain a tryptic peptide with mass 1000.0±0.5 Da and 2000.0±0.5 Da, are 270 and 108, respectively.

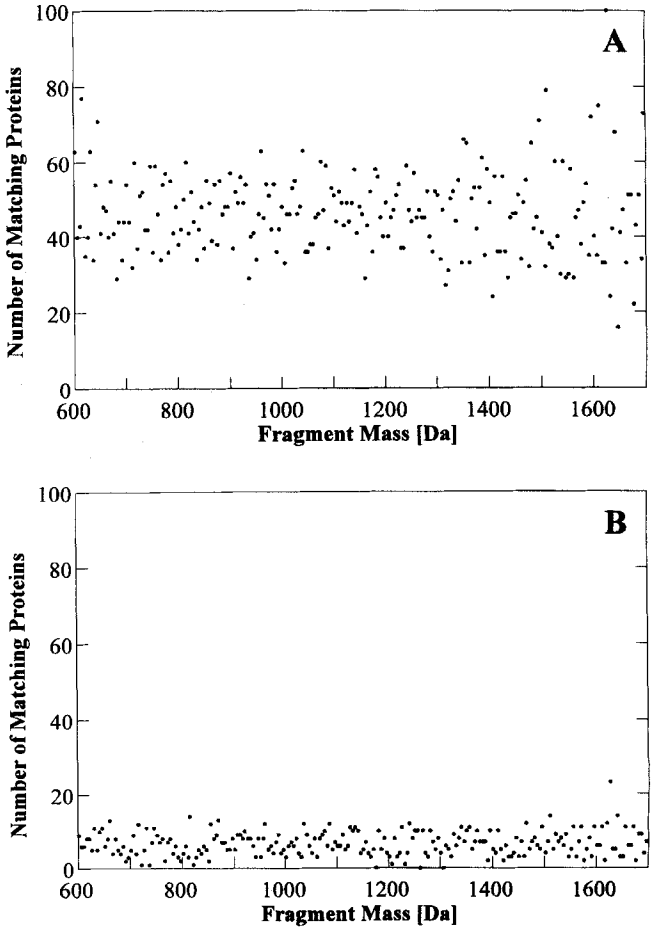

Figure 6. The distribution of the number of *S. cerevisiae* proteins that contain one tryptic peptide with mass 2000±2 Da (total of 486 proteins) as a function of fragment ion mass for mass accuracy ±2 Da (only *b* and *y* ions are considered here). (A) Fragmentation at any amino acid. (B) Fragmentation at aspartic (D) or glutamic (E) acid.

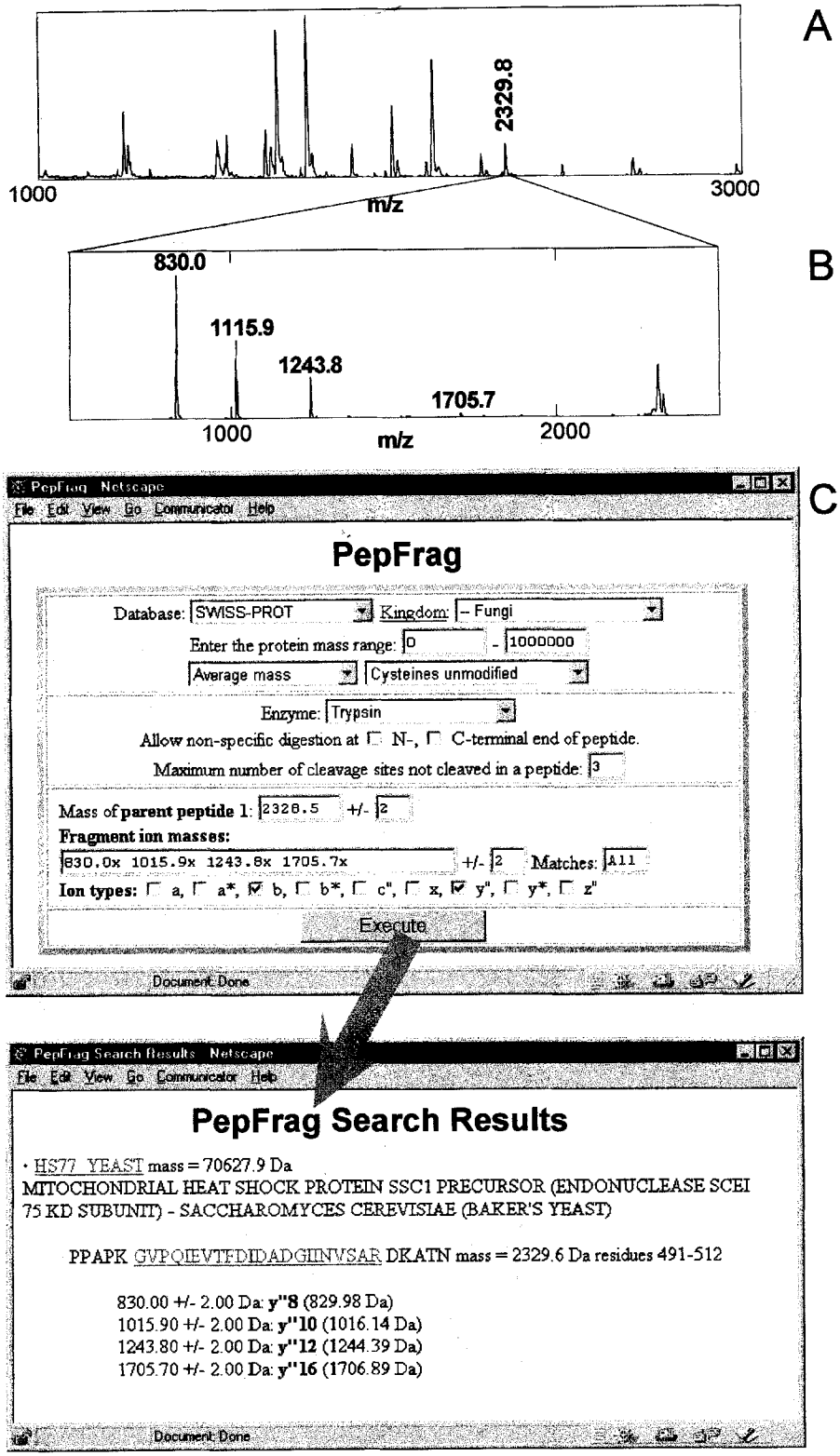

Figure 7. (A) A MALDI-IT-MS spectrum of a tryptic digest of an unknown protein from *S. cerevisiae*. (B) MALDI-IT-MS fragmentation spectrum of ions with m/z=2329.8. (C) The WWW interface to the database search tool, PepFrag. A database is selected and a series of restrictions are chosen (taxonomic division, chemical modification, enzyme specificity, completeness of digestion, proteolytic peptide mass, fragment masses, fragmentation systematics, and partial amino acid composition). In this example, all fungi in SWISS-PROT are searched for tryptic peptides with mass 2328.8±2.0 Da that can yield *b* or *y* fragment ions of mass 830.0, 1115.9, 1243.8, and 1705.7 when fragmenting at acidic amino acids. Only one peptide (GVPQIEVTFDIDADGIINVSAR) from a heat shock protein satisfies this constraint.

proteins match the additional constraint of having 19 exchangeable hydrogens – giving a fivefold reduction in the number of matching proteins. This calculation is based on the assumption that experimentally the hydrogen/deuterium exchange is complete and that no back-exchange occurs. It is probably more realistic to assume that 90%–100% of the hydrogens will be exchanged. Using this latter assumption, 97 proteins match, still yielding a threefold reduction in the number of matching proteins. In this H/D exchange experiment all the peptides in the peptide map will shift in mass, and a comparison can be made for all of the peaks in the two mass spectra (that can be unambiguously associated), allowing for simultaneous counting of the number of

exchangeable hydrogens in a large number of proteolytic peptides.

3.6 Mass spectrometric fragmentation

An effective way to obtain highly constraining information about a peptide is to cause it to fragment in the mass spectrometer and to measure the masses of the resulting fragment ions [30–40]. Figure 6 shows the number of *S. cerevisiae* proteins containing a tryptic peptide with mass 2000±2 Da as a function of fragment ion mass, assuming that the fragmentation produces exclusively *b*- and *y*-type fragment ions (which is frequently the case for low-energy dissociation of tryptic peptides). The mass of a single fragment ion (mass accuracy ±2 Da) restricts the number of matching proteins approximately tenfold (Fig. 6A). In MALDI-IT-MS it has been observed that tryptic peptides that contain arginine preferentially fragment at the *C*-terminal side of acidic amino acids [48]. If this additional information is used, a further tenfold improvement is obtained (Fig. 6B), *i.e.*, totally, a 100-fold improvement.

3.7 PepFrag

The software tool PepFrag allows for searching protein or nucleotide sequence databases (SWISS-PROT, PIR, GENPEPT, OWL or dbEST) using a combination of different types of information from mass spectra of peptide maps and fragmentation spectra of peptides. It is publicly available over the Internet at URL http://prowl.rockefeller.edu/ as a part of PROWL – an interactive environment on the Word Wide Web for protein mass spectrometry [49] (see also other software tools for protein iden-

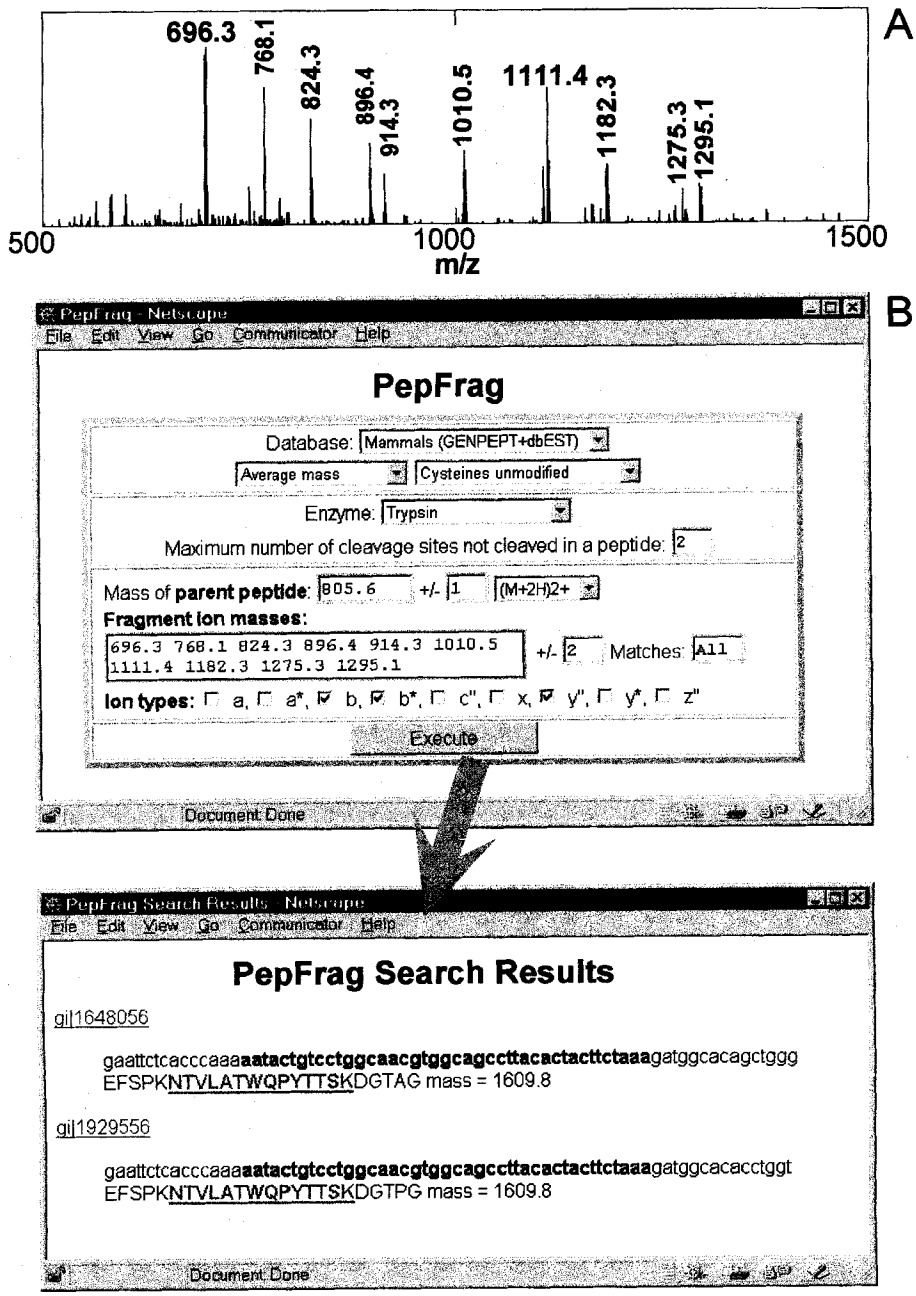

Figure 8. (A) An LCQ MS/MS spectrum of a doubly charged tryptic peptide (m/z = 805.6) from an unknown rabbit protein. (B) PepFrag search of the mammalian sequences in GenPept and dbEST using the fragmentation information. One tryptic peptide (NTVLATWQPYTTSK) from two human EST sequences satisfies the constraints.

tification on the WWW). The databases have been taxonomically divided to allow for faster searches and to minimize the number of unrelated hits. Experimental conditions such as enzyme specificity, approximate protein mass, position in a phylogenetic tree, and modifications of amino acids can be specified in the search (see Figs. 7 and 8). In addition, other search constraints can be specified including fragmentation systematics, masses of other proteolytic peptides, and partial amino acid composition.

We give here an example of the use of Pepfrag for protein identification (Fig. 7). An unknown protein from *S. cerevisiae* was purified with SDS-PAGE, digested with trypsin in the gel and the tryptic peptides were analyzed with MALDI-IT-MS (Fig. 7A). The ions with *m/z* 2329.8 were isolated in the ion trap and fragmented (Fig. 7B). The four fragment ion peaks correspond to fragmentation at the *C*-terminal side of acidic amino acids [48]. In the PepFrag search of the database (SWISS-PROT), the kingdom (fungi), and the enzyme (trypsin) were specified (Fig. 7C). The mass of the parent peptide and the mass of its fragment ions were entered into the form. In addition, it was specified that the fragmentation gives rise to *b*- and/or *y*-type ions and occurs at the *C*-terminal side of acidic amino acids [48]. The result of the search is a list of proteins that each contains a peptide that matches the measured mass of a proteolytic peptide as well as its fragment masses. In the example in Fig. 7, only one peptide GVPQIEVTFDIDADGIINVSAR from a mitochondrial heat shock protein matches the search constraints. The certainty of the identification can be further tested by relaxing the search conditions. For this example no additional proteins match the search constraints when one of the following conditions is allowed: nonspecific digestion, any type of backbone fragmentation, fragmentation at any amino acid, 3 out of 4 fragment ions required to match.

A second example for the use of PepFrag for protein identification is given in Fig. 8. An unknown rabbit protein, believed to be involved in translation, was digested with trypsin and analyzed by electrospray ionization ion trap MS/MS (LCQ). The fragmentation spectrum of a doubly charged tryptic peptide ($m/z = 805.6$) is shown in Fig. 8A. The mammalian sequences in GenPept and dbEST were searched with the fragmentation information. Two human EST sequences, coding for a peptide matching the experimental constraints were found (Fig. 8B), indicating that this protein is highly conserved in mammals. If the specificity of trypsin is ignored in the search, the same peptide is also found in one mouse EST sequence (not shown). An additional five human EST sequences code for a highly similar peptide (the second threonine is serine).

The search time for the examples above on a PC with a 300 MHz Pentium II processor is 9, 33, and 119 s for the entire SWISS-PROT (59021 sequences), GenPept (262153 sequences), and dbEST (11639909 sequences translated into six reading frames) databases, respectively. If only a part of the database is searched, the search time is dramatically decreased. For example, the search time for GenPept is 7, 3, and 3 s for mammals, primates, and fungi, respectively. The search times can be further decreased by keeping the databases in RAM instead of reading them from disks for every search.

* http://prospector.ucsf.edu/ and http://www.mann.embl-heidelberg.de/Services/PeptideSearch/PeptideSearchIntro.html.

4 Concluding remarks

We have investigated the *S. cerevisiae* genome *in silico* in order to compare the information content in different kinds of mass spectrometric measurements. A wide variety of information can be used to increase the confidence level of the identification. A knowledge of the relative value of this information is useful in the design of efficient protein identification experiments.

This work was supported by the National Science Foundation (Grant 9630936) and the National Institute of Health (Grant RR00862). The authors acknowledge many fruitful discussions with Ronald C. Beavis, Júlio C. Padovan, Salvatore Sechi, Wenzhu Zhang and Yingming Zhao.

Received November 6, 1997

5 References

[1] Muzio, M., Chinnaiyan, A. M., Kischkel, F. C., O'Rourke, K., Shevchenko, A., Ni, J., Scaffidi, C., Bretz, J. D., Zhang, M., Gentz, R., Mann, M., Krammer, P. H., Peter, M. E., Dixit, V. M., *Cell* 1996, *85*, 817–827.

[2] Clauser, K. R., Hall, S. C., Smith, D. M., Webb, J. W., Andrews, L. E., Tran, H. M., Epstein, L. B., Burlingame, A. L., *Proc. Nat. Acad. of Sci. USA* 1995, *92*, 5072–5076.

[3] Matsui, N. M., Smith, D. M., Clauser, K. R., Fichmann, J., Andrews, L. E., Sullivan, C. M., Burlingame, A. L., Epstein, L. B., *Electrophoresis* 1997, *18*, 409–417.

[4] Sullivan, C. M., Smith, D. M., Matsui, N. M., Andrews, L. E., Clauser, K. R., Chapeaurouge, A., Burlingame, A. L., Epstein, L. B., *Cancer Res.* 1997, *57*, 1137–1143.

[5] Neubauer, G., Gottschalk, A., Fabrizio, P., Seraphin, B., Luhrmann, R., Mann, M., *Proc. Nat. Acad. Sci. USA* 1997, *94*, 385–390.

[6] Winter, D., Podtelenikov, A. V., Mann, M., Li, R., *Curr. Biol.* 1997, *7*, 519–529.

[7] Qin, J., Fenyö, D., Zhao, Y., Hall, W. W., Chao, D. M., Wilson, C. J., Young, R. A., Chait, B. T., *Anal. Chem.* 1997, *69*, 3995–4001.

[8] Shevchenko, A., Jensen, O. N., Podtelejnikov, A. V., Sagliocco, F., Vorm, O., Mortensen, P., Boucherie, H., Mann, M., *Proc. Nat. Acad. Sci. USA* 1996, *93*, 14440–14445.

[9] Link, A. J., Carmack, E., Yates, J. R., III, *Int. J. Mass Spec. Ion Proc.* 1997, *160*, 303–316.

[10] Fountoulakis, M., Langen, H., Evers, S., Gray, C., Takács, B., *Electrophoresis* 1997, *18*, 1193–1202.

[11] Patterson, S. D., Aebersold, R., *Electrophoresis* 1995, *16*, 1791–1814.

[12] Henzel, W. J., Billeci, T. M., Stultz, J. T., Wong, S. C., Grimley, C., Watanbe, C., *Proc. Natl. Acad. Sci. USA* 1993, *90*, 5011–5015.

[13] Mann, M., Højrup, P., Roepstorff, P., *Curr. Biol.* 1993, *22*, 388–345.

[14] Pappin, D. D. J., Højrup, P., Bleasby, A. J., *Curr. Biol.* 1993, *3*, 327–332.

[15] Yates, J. R., III., Speichner, S., Griffin, P. R., Hunkapiller, T., *Anal. Biochem.* 1993, *214*, 397–408.

[16] James, P., Quadroni, M., Carafoli, E., Gonnet, G., *Biochem. Biophys. Res. Commun.* 1993, *195*, 58–64.

[17] Rasmussen, H. H., Mørtz, E., Mann, M., Roepstorff, P., Celis, J. E., *Electrophoresis* 1994, *15*, 406–416.

[18] Mørtz, E., Vorm, O., Mann, M., Roepstorff, P., *Biol. Mass Spectrom.* 1994, *23*, 249–261.

[19] James, P., Quadroni, M., Carafoli, E., Gonnet, G., *Protein Sci.* 1994, *3*, 1347–1350.
[20] Cordwell, S. J., Wilkins, M. R., Cerpa-Poljak, A., Gooley, A. A., Duncan, M., Williams, K. L., Humprey-Smith, I., *Electrophoresis* 1995, *16*, 438–443.
[21] Patterson, S. D., *Electrophoresis* 1995, *16*, 1104–1114.
[22] Jensen, O. N., Podtelenikov, A. V., Mann, M., *Rap. Commun. Mass Spectrom.* 1996, *10*, 1371–1378.
[23] Jensen, O. N., Vorm, O., Mann, M., *Electrophoresis* 1996, *17*, 938–944.
[24] Jensen, O. N., Houthaeve, T., Shevchenko, A., Cudmore, S., Ashford, T., Mann, M., Griffiths, G., Locker, J. K., *J. Virol.* 1996, *70*, 7485–7497.
[25] Patterson, S. D., Thomas, D., Bradshaw, R. A., *Electrophoresis* 1996, *17*, 877–891.
[26] Corchesne, P. L., Luethy, R., Patterson, S. D., *Electrophoresis* 1997, *18*, 369–381.
[27] Zhang, W., Chait, B. T., *Proceedings of the 43rd ASMS Conference on Mass Spectrometry and Allied Topics*, Atlanta, GA 1995.
[28] Wilkins, M. R., Pasquali, C., Appel, R. D., Ou, K., Golaz, O., Sanchez, J.-C., Yan, J. X., Gooley, A. A., Hughes, G., Humphery-Smith, I., Williams, K., Hochstrasser, D. F., *Biotechnology* 1996, *14*, 61–65.
[29] Wilkins, M. R., Ou, K., Appel, R. D., Sanchez, J.-C., Yan, J. X., Golaz, O., Farnsworth, V., Cartier, P., Hochstrasser, D. F., Williams, K. L., Gooley, A. A., *Biochem. Biophys. Res. Commun.* 1996, *221*, 609–613.
[30] Eng, J. K., McCormack, A. L., Yates, J. R., III, *J. Amer. Soc. Mass Spec.* 1994, *5*, 976–987.
[31] Mann, M., Wilm, M., *Anal. Chem.* 1994, *66*, 4390–4399.
[32] Yates, J. R., III, Eng, J. K., McCormack, A. L., Schieltz, D., *Anal. Chem.* 1995, *67*, 1426–1436.
[33] Yates, J. R., III, Eng, J. K., McCormack, A. L., *Anal. Chem.* 1995, *67*, 3202–3210.
[34] Griffin, P. R., MacCoss, M. J., Eng, J. K., Blevins, R. A., Aaronson, J. S., Yates, J. R., III, *Rap. Commun. Mass Spec.* 1995, *9*, 1546–1551.
[35] Yates, J. R., III, Eng, J. K., Clauser, K. R., Burlingame, A. L., *J. Amer. Soc. Mass Spec.* 1996, *7*, 1089–1098.
[36] Mørtz, E., O'Connor, P., Roepstorff, P., Kelleher, N. L., Wood, T. D., McLafferty, F. W., Mann, M., *Proc. Nat. Acad. Sci. USA* 1996, *93*, 8264–8267.
[37] Figeys, D., van Oostveen, I., Ducret, A., Aebersold, R., *Anal. Chem.* 1996, *68*, 1822–1828.
[38] Figeys, D., Ducret, A., Yates, J. R., III, Aebersold, R., *Nature Biotechnol.* 1996, *14*, 1579.
[39] Figeys, D., Aebersold, R., *Electrophoresis* 1997, *18*, 360–368.
[40] McCormack, A. L., Schieltz, D. M., Goode, B., Yang, S., Barnes, G., Drubin, D., Yates, J. R., III, *Anal. Chem.* 1997, *69*, 767–776.
[41] Goffeau, A., Barrell, B. G., Bussey, H., Davis, R. W., Dujon, B., Feldmann, H., Galibert, F., Hoheisel, J. D., Jacq, C., Johnston, M., Louis, E. J., Mewes, H. W., Murakami, Y., Philippsen, P., Tettelin, H., Oliver, S. G., *Science* 1996, *274*, 563–567.
[42] Johnston, M., *Curr. Biol.* 1996, *6*, 500–503.
[43] Bairoch, A., Apweiler, R., *Nucleic Acids Res.* 1997, *25*, 31–36.
[44] Benson, D. A., Boguski, M. S., Lipman, D. J., Ostell, J., *Nucleic Acids Res.* 1997, *25*, 1–6.
[45] Qin, J., Steenvoorden, R. J. J. M., Chait, B. T., *Anal. Chem.* 1996, *68*, 1784–1791.
[46] Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., Moreno, R. F., Kerlavage, A. R., McCombie, W. R., Venter, J. C., *Science* 1991, *252*, 1651–1656.
[47] Chait, B. T., Wang, R., Beavis, R. C., Kent, S. B. H., *Science* 1993, *262*, 89–92.
[48] Qin, J., Chait, B. T., *J. Amer. Chem. Soc.* 1995, *117*, 5411–5412.
[49] Fenyö, D., Zhang, W., Beavis, R. C., Chait, B. T., *Anal. Chem.* 1996, *68*, A721–A726.