Laboratory for Computational Proteomics

www.FenyoLab.org

E-mail: Info@FenyoLab.org

Facebook: *NYUMC Computational Proteomics Laboratory*

Twitter: *@CompProteomics*

Helen I. Field[1]
David Fenyö[2]
Ronald C. Beavis[3]

[1]The Rockefeller University, New York, NY, USA
[2]ProteoMetrics, New York, NY, USA
[3]ProteoMetrics Canada, Winnipeg, Canada, and Professor of Biochemistry, School of Medicine, University of Manitoba, Canada

RADARS, a bioinformatics solution that automates proteome mass spectral analysis, optimises protein identification, and archives data in a relational database

RADARS, a rapid, automated, data archiving and retrieval software system for high-throughput proteomic mass spectral data processing and storage, is described. The majority of mass spectrometer data files are compatible with RADARS, for consistent processing. The system automatically takes unprocessed data files, identifies proteins *via in silico* database searching, then stores the processed data and search results in a relational database suitable for customized reporting. The system is robust, used in 24/7 operation, accessible to multiple users of an intranet through a web browser, may be monitored by Virtual Private Network, and is secure. RADARS is scalable for use on one or many computers, and is suited to multiple processor systems. It can incorporate any local database in FASTA format, and can search protein and DNA databases online. A key feature is a suite of visualisation tools (many available *gratis*), allowing facile manipulation of spectra, by hand annotation, reanalysis, and access to all procedures. We also described the use of Sonar MS/MS, a novel, rapid search engine requiring ~40 MB RAM *per* process for searches against a genomic or EST database translated in all six reading frames. RADARS reduces the cost of analysis by its efficient algorithms: Sonar MS/MS can identifiy proteins without accurate knowledge of the parent ion mass and without protein tags. Statistical scoring methods provide close-to-expert accuracy and brings robust data analysis to the non-expert user.

Keywords: Protein identification / Matrix-assisted laser desorption/ionization mass spectrometry / Tandem mass spectrometry / Liquid chromatography-mass spectrometry / Software / Bioinformatics / Database
PRO 0145

The proteome is the state of all proteins in one cell at one time. Proteomics methodologies access the proteome, defining many proteins in a biological sample [1, 2]. Mass spectrometry can identify nanomoles of a protein [3], facilitating high throughput (HT) proteomics, *e.g.* primary characterization by two-dimensional gel electrophoresis (2-DE) [4] or peptide analysis [5]; comparison of cellular states by 2-DE [6], liquid chromatography [7] or silicon chips [8, 9]; definition of organellar proteins in subfractionated organelles [10]; and protein interaction networks, by "pull-out" experiments using antibodies or affinity tags [11, 12]. The method of choice for identifying proteins is mass spectrometry (MS) [1–12]. A bottleneck for HT proteomics studies is bioinformatics, from MS output and protein identification, to data storage and interpretation. This paper describes a software system that goes some way towards meeting this need.

Proteins are extracted from SDS-PAGE gels (if necessary) [13]: commercial, automated systems are available for HT preparation. Partially purified proteins are often proteolysed (*e.g.* with trypsin) before MS or tandem MS (MS/MS) analysis. In MALDI, protein extracts are dried onto a plate with an organic matrix material, ionised with a laser, then analysed using a TOF analyzer [14, 15]. The arrival time of an ion at the detector depends on the mass, charge and kinetic energy of the ion. To generate sequence-specific information, peptides can be further fragmented in the gas phase. Where MALDI is the ionisation process, spontaneous post-source decay (PSD) fragmentation occurs after the ions have left the ion source region, and can be measured by TOF. In LC-MS, liquid chromatography is used to separate peptides, and the eluate is fed into an electrospray mass spectrometer. Electrospray ionisation (ESI) generates molecular ions directly from solution, by creating a fine spray of highly charged droplets in the presence of a strong electric field. Ions are

Correspondence: Dr. Helen I. Field, ProteoMetrics Inc., PO Box 32323, London SW17 8JZ, UK
E-mail: field@proteometrics.com
Fax: +44-20-8672-3550

Abbreviations: HT, high throughput; **HTTP**, hyper text transfer protocol; **PAWS**, Protein Analysis Worksheet application program; **RADARS**, RApid Data Archival and Retrieval System

1615-9853/02/0101–36 $17.50+.50/0

introduced into a cell where fragmentation occurs by controlled collision with a gas [16]. The *m/z* of ions are measured by a quadrupole or ion trap mass analyzer. These methods allow analysis of fragments from each peptide parent ion. Output from MS is a spectrum of ions with defined *m/z*. Intensity is plotted against *m/z*. A software engine takes the MS output, calls significant peaks and identifies the protein from genomic, protein or expressed sequence tag (EST) databases: to identify proteins, experimental mass is compared with theoretical masses generated from hypothetically translated, digested (and optionally fragmented) proteins [17–21].

MS and MS/MS protein data are being generated at ever increasing rates. Automated MS has created a new rate limiting step in bioinformatics – protein identification. Overcoming this required an automated protein identification system that stores and links relevant data, making data retrieval trivial. Requirements for a HT protein identification and data retrieval system were: secure, networked and compatible with intranet protocols; scalable, for growth; used without special computer hardware requirements; able to fit into any laboratory system workflow or laboratory information management system (LIMS); providing indelible, flexible, organised data storage for data mining, and customized reporting; incapable of dropping a sample if a computer is temporarily disabled; robust for 24 h, 7 d operation; consistent for use with any mass spectrometer; accurate protein identification; fast protein identification especially from large DNA (genomic) databases; statistically valid, objective scoring; eliminating an absolute requirement for an expert human user. A software suite named RApid Data Archiving and Retrieval System (RADARS) was created to fulfil these principles.

MoverZ is a software module for mass spectral display and contains tools for peak calling, mass labelling, and identification of post-translational modifications by mass shift (Fig. 1). MoverZ currently takes MS files from: Applied Biosystems Perkin Elmer, Bruker, Finnegan, Micromass

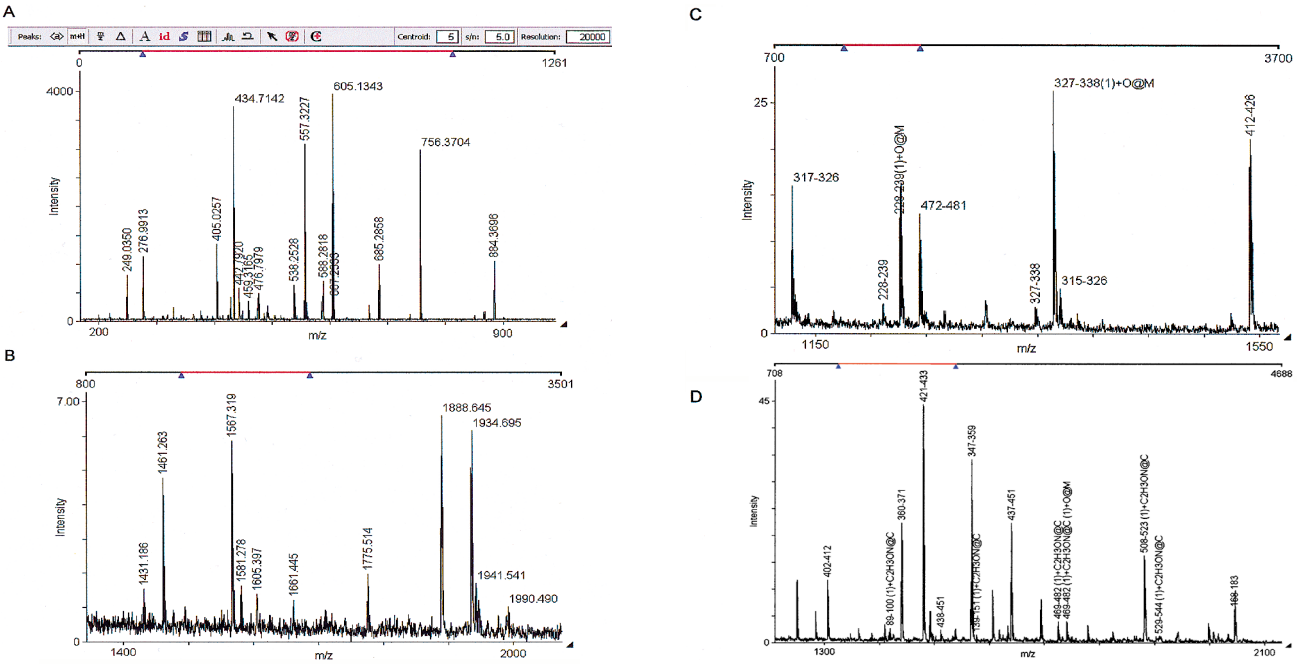


Figure 1. MoverZ output functions. MoverZ displays: inputs are raw data files from Finnegan LCQ (A) or MALDI files from Voyager DE (B, C) and Bruker (D) mass spectrometers. (A), Toolbar for MoverZ (from left to right): [<a>] annotate manually; [m+H] add a 1Da mass unit to peak masses; [m/z] annotate peaks by hand; [Δ] assign difference masses relative to one peak (user chooses reference peak, mass differences are called for other peaks, and modifications assigned); [A] autoannotate all peaks (peaks are derived from raw data and masses called); [id] identify proteins (sends peak list to ProFound for analysis); [S] *ditto* for Sonar MS/MS; [table] shows list of masses, mode (monoisotopic or average mass assignment), peak intensity and S/N for each annotated peak; [⇐] go back one move; [↖] permits the user to remove single labels upon selection of a labeled peak; [m∅z] removes annotations; [C+] calibrates (user assigns calibration masses and chooses peaks). Centroid width, S/N (for peak detection) and resolution may be user defined. Zoom, copy and print tools are available. The bar over the spectrum indicates which portion of that spectrum is being viewed (red), and is a navigation tool. (B–D), MALDI spectra annotated by mass before protein identification (B), annotated with peptide residues and modifications after protein identification by RADARS (C). (D), high resolution fragment of a MALDI spectrum (Bruker), showing monoisotopic peak calling at higher mass with masses labelling each peak (label orientation is user defined). Intensity of a peak is the sum of the intensities across the selected number of centroid units. Intensity measurements are in arbitrary units as supplied by the mass spectrometer.

and Sciex instruments. There is no standard for MS file formats, which are regularly altered by instrument manufacturers, so MoverZ is constantly being updated in response to reports of incompatibility. Compatibility is achieved by providing a translation layer between the files and a standard set of software objects used by all modules accessing MS information. The objects were designed using the Pioneer object scheme (http://canada.proteometrics.com/Pioneer/index.html), and use a simplified XML (MASSML) to store and transmit spectra and associated information. Another XML (BioML) is used for detailed biopolymer sequence information (http://www.bioml.com). Compatibility with some MS/MS outputs is achieved by using ASCII peak lists (DTA and PKL files) supplied, with user chosen S/N superimposed. Other than these data formats, MoverZ calls peaks as follows.

Peak detection in peptide MS is complicated by the existence of a distribution of peaks for each peptide, caused by the presence of ^{13}C atoms in any population of molecules ($\sim$ 1.1% of natural carbon is ^{13}C). The most important peak for protein identification is the lowest mass peak in the distribution (the A_0, or monoisotopic peak), which contains only ^{12}C, ^{14}N, ^{16}O, ^{1}H and ^{32}S; if it corresponds to an unmodified peptide its mass will precisely match the theoretical peptide mass. Accurate selection of this peak in a noisy signal is critical: selection of the wrong peak gives a mass error of at least one Dalton, regardless of the accuracy of the mass measurement. MoverZ finds A_0 as follows: first it detects isotope peak clusters of appropriate width for a peptide with a chemical average mass that would correspond to a particular cluster. The cluster is tested to see if its integrated root-mean-squared (RMS) intensity is higher than that of the local background, calculating the S/N ratio of the cluster RMS intensity to the background RMS intensity (the user sets an appropriate minimum S/N for their data). The cluster should have an average mass and an appropriate width which correlates with a predicted Poisson distribution of isotope peak intensities for that mass. If the cluster has appropriate width and S/N, the position of the peak that should correspond to the A_0 peak is approximated, based on the predicted Poisson distribution of isotope peak intensities for an average mass. The peak closest to the predicted mass for the A_0 peak is identified and its intensity compared to the other peaks in the distribution. If the intensity is within 2 standard deviations of that predicted by the Poisson distribution, the mass of this peak is assigned as the A_0 mass representing the cluster (Fig. 1C).

MoverZ was part of the client-server software system PROWL [22]. It has filtering and smoothing functions [23]. Autoannotation calls peaks from MALDI-MS data, choosing A_0 peaks where possible. If not, or if average mass calling is selected, MoverZ uses a published algorithm to calculate the mass shift [24]. This, importantly, yields better protein identification scores in protein identification for MALDI using the search engine ProFound [17]. ProFound ranks hits by a Bayesian algorithm having protein data as input (p*I*, molecular mass, species) and pattern recognition algorithms in the output, and was developed semi-empirically [17].

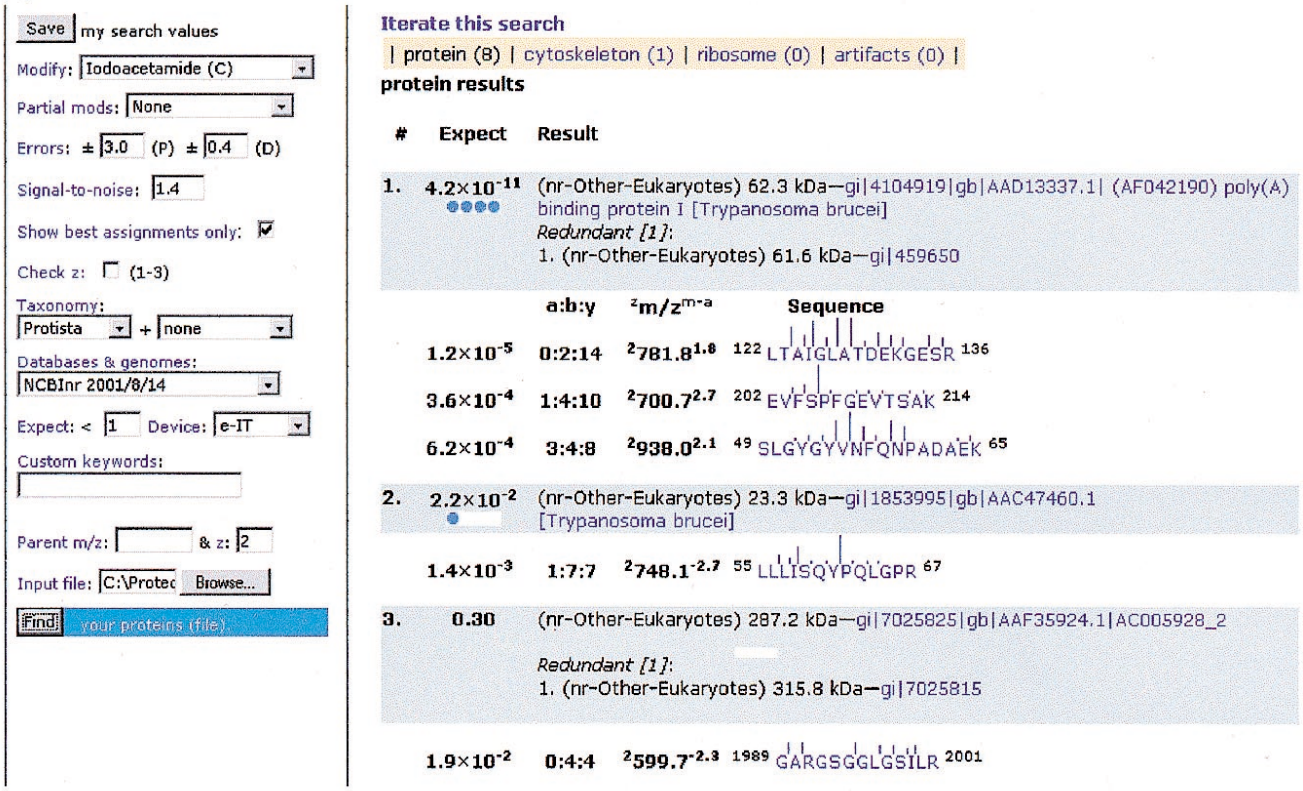
For MS/MS a novel search engine was created: Sonar MS/MS. Sonar MS/MS looks for b and y ions, and returns as many proteins as the user requests. A results panel is returned for each protein, containing an analysis of each peptide from that protein (Fig. 2A). The fragmentogram was designed as a rapid visual quality check of the experiment. Performance was refined by identifying proteins from $>10^5$ experimental spectra (discussion of the algorithms used to create Sonar MS/MS are beyond the scope of this paper and will be published elsewhere). Sonar MS/MS searches against genomic data (as well as protein data). This is particularly important for unfinished genomes, where proteins are incompletely annotated (and therefore missed). Sonar MS/MS provided novel peptide identifications in a genomic search, not found by searching against a protein database (Fig. 2A, B). Correct identification is independent of parent (peptide) ion mass: a window of $\pm$2 Da or $\pm$1000 Da given to the parent ion gives the same top hit (Fig. 2C, D). Thus, biologically modified proteins can be identified by Sonar MS/MS *via* unmodified peptides. Sonar MS/MS is rapid (Table 1). Thus Sonar MS/MS may be used for HT processing on standard computer equipment (in this case a Pentium 750 MHz dual processor server).

Table 1. Sonar MS/MS search times (as of August 2001) on a Dell PowerEdge 2500 computer with 512 MB RAM, two 1GHz Pentium III processors, and with a 10 000 RPM SCSI hard drive configured in RAID 5.

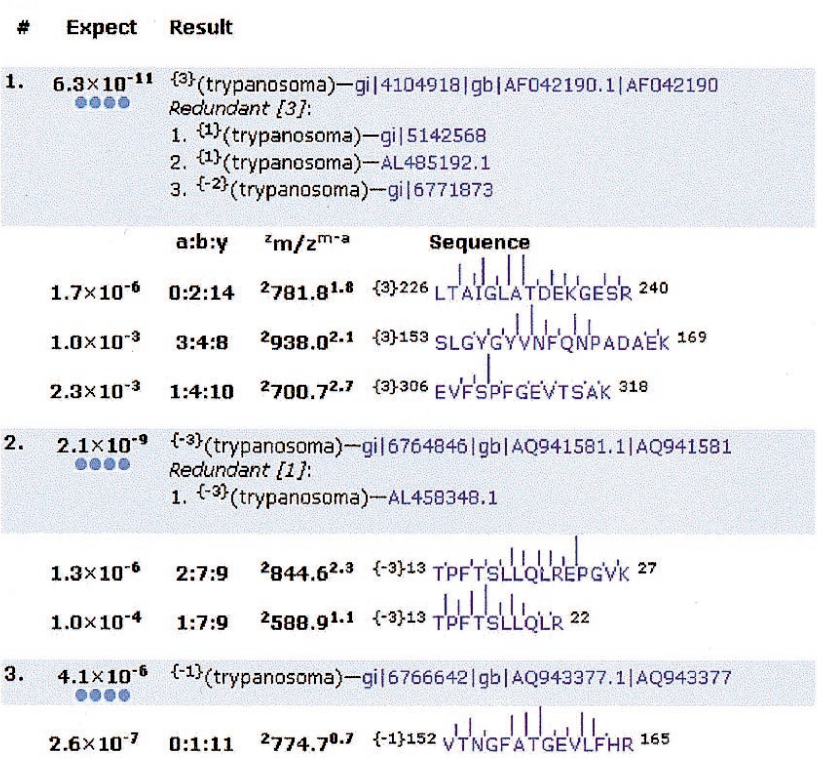
Spectral run	database	details	Time/ spectrum
244 MS/MS from LC/MS	nr-human		0.107 s
244 MS/MS from LC/MS	dbEst-human	6 reading frames	5.1 s
244 MS/MS from LC/MS	human genome	6 reading frames	7.2 s

The development with the greatest impact on HT work was the statistical quality control scoring, developed for ProFound and Sonar MS/MS. For ProFound, protein hits were obtained from pseudo-spectra, which consisted of groups of peptides taken randomly from different proteins

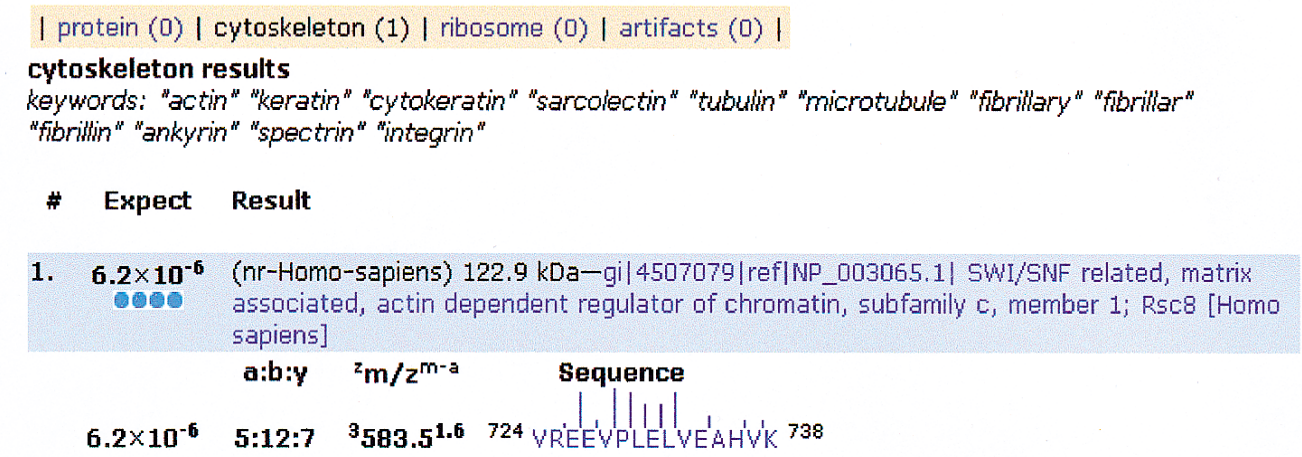
A


Save my search values
Modify: Iodoacetamide (C)
Partial mods: None
Errors: ± 3.0 (P) ± 0.4 (D)
Signal-to-noise: 1.4
Show best assignments only:
Check z: (1-3)
Taxonomy:
Protista + none
Databases & genomes:
NCBInr 2001/8/14
Expect: < 1 Device: e-IT
Custom keywords:
Parent m/z: & z: 2
Input file: C:\Protec Browse...
Find your proteins (file)
Iterate this search
| protein (8) | cytoskeleton (1) | ribosome (0) | artifacts (0) |
protein results
Expect Result
1. 4.2×10-11 (nr-Other-Eukaryotes) 62.3 kDa—gi|4104919|gb|AAD13337.1| (AF042190) poly(A) binding protein I [Trypanosoma brucei]
Redundant [1]:
1. (nr-Other-Eukaryotes) 61.6 kDa—gi|459650
a:b:y z m/z m-a Sequence
1.2×10-5 0:2:14 2 781.8 1.8 122 LTAIGLATDEKGESR 136
3.6×10-4 1:4:10 2 700.7 2.7 202 EVFSPFGEVTSAK 214
6.2×10-4 3:4:8 2 938.0 2.1 49 SLGYGYVNFQNPADAEK 65
2. 2.2×10-2 (nr-Other-Eukaryotes) 23.3 kDa—gi|1853995|gb|AAC47460.1 [Trypanosoma brucei]
1.4×10-3 1:7:7 2 748.1 -2.7 55 LLLISQYPQLGPR 67
3. 0.30 (nr-Other-Eukaryotes) 287.2 kDa—gi|7025825|gb|AAF35924.1|AC005928_2
Redundant [1]:
1. (nr-Other-Eukaryotes) 315.8 kDa—gi|7025815
1.9×10-2 0:4:4 2 599.7 -2.3 1989 GARGSGGLGSILR 2001

B


Expect Result
1. 6.3×10-11 {3}(trypanosoma)—gi|4104918|gb|AF042190.1|AF042190
Redundant [3]:
1. {1}(trypanosoma)—gi|5142568
2. {1}(trypanosoma)—AL485192.1
3. {-2}(trypanosoma)—gi|6771873
a:b:y z m/z m-a Sequence
1.7×10-6 0:2:14 2 781.8 1.8 {3}-226 LTAIGLATDEKGESR 240
1.0×10-3 3:4:8 2 938.0 2.1 {3}-153 SLGYGYVNFQNPADAEK 169
2.3×10-3 1:4:10 2 700.7 2.7 {3}-306 EVFSPFGEVTSAK 318
2. 2.1×10-9 {-3}(trypanosoma)—gi|6764846|gb|AQ941581.1|AQ941581
Redundant [1]:
1. {-3}(trypanosoma)—AL458348.1
1.3×10-6 2:7:9 2 844.6 2.3 {-3}13 TPFTSLLQLREPGVK 27
1.0×10-4 1:7:9 2 588.9 1.1 {-3}13 TPFTSLLQLR 22
3. 4.1×10-6 {-1}(trypanosoma)—gi|6766642|gb|AQ943377.1|AQ943377
2.6×10-7 0:1:11 2 774.7 0.7 {-1}152 VTNGFATGEVLFHR 165

C



D

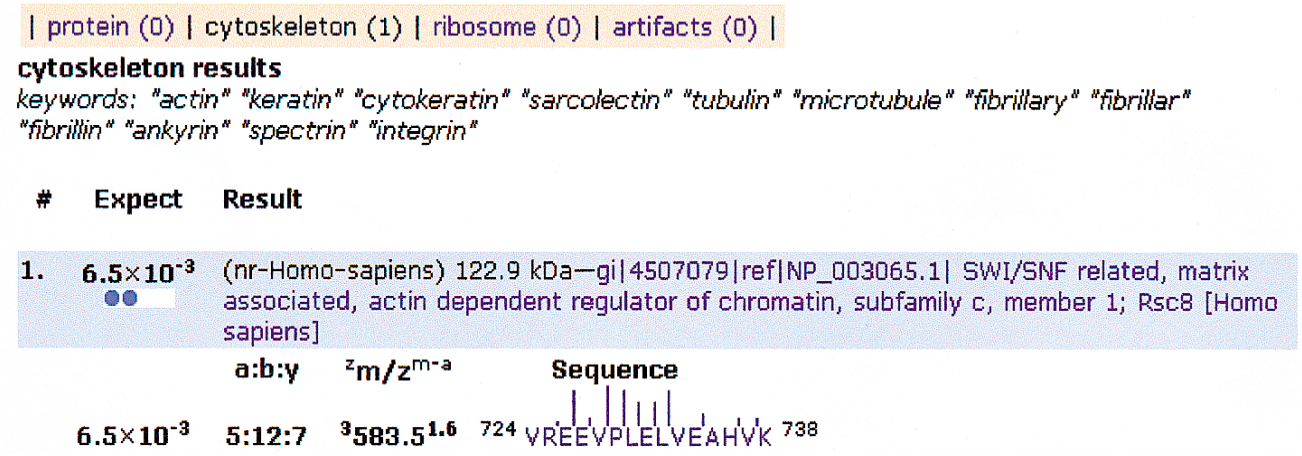


Figure 2. Sonar MS/MS searches showing performance with protein *versus* DNA data, and lack of influence of parent ion mass. (A), Sonar MS/MS interface. (Left) input parameter entry: *Save* button: stores parameter set; *Modify*: choose preparative and *Partial* modifications; error (Da) for Parent (*P*) and Daughter (*D*) ions; S/N for peak detection; *Show best assignments only*: check to show top hits, uncheck for all hits; *Check z*: select this to check all data against parent ion charges of 1, 2 and 3; *Taxonomy*: select taxa for search; *Databases . . .*: choose; *Expect* <: enter number of proteins in the sample (1–4 allowed); *Device*: mass spectrometer type (e-IT is ESI ion trap); *Custom keywords* to sort protein results; *Parent m/z*: enter if known; *z*; enter parent charge, *Input file*: type or search local disc (*Browse . . .* button). In RADARS, data entry is automatic. *Find* button: initiates search. (Right) results panel: (top) Iterate this search takes you to a new interface summarising results and permitting additional assignments; summary of proteins identified, by keyword (default keywords shown). 'Protein results' indicates that a protein (not DNA) sequence database was used for search. Table of peptide hits from the protein: #, rank of protein in list; 'Expect', Expectation Value (*e.g.* 4.2×10^{-11}, see Fig. 3); 'Result' includes: database searched, size of returned protein, GenBank accession numbers for the protein and identified homologues, species. a:b:y ratio of those fragmentation ions; $^{z}m/z^{m-a}$, ^{charge}mass/charge$^{measured\ minus\ calculated\ mass}$; 'Sequence': fragmentogram with peptide sequence, position in the protein (numbers of start and finish amino acids); vertical bar between amino acid pairs indicates ion intensity: one ion (the most intense) is represented (no bar indicates no ion); click-through fragmentogram to detailed lists of the fragmentation ions. (A) shows results from a test MS/MS spectrum, and (B) shows improved search of same data using DNA database, leading to additional, unannotated sequence data being identified as significant. (C, D) Influence of parent ion mass. Sonar MS/MS results on one data spectrum, where parent ion mass is set differently. The same protein is returned whether the parent ion mass window is set to ± 2Da (C), or ± 1000 Da (D).

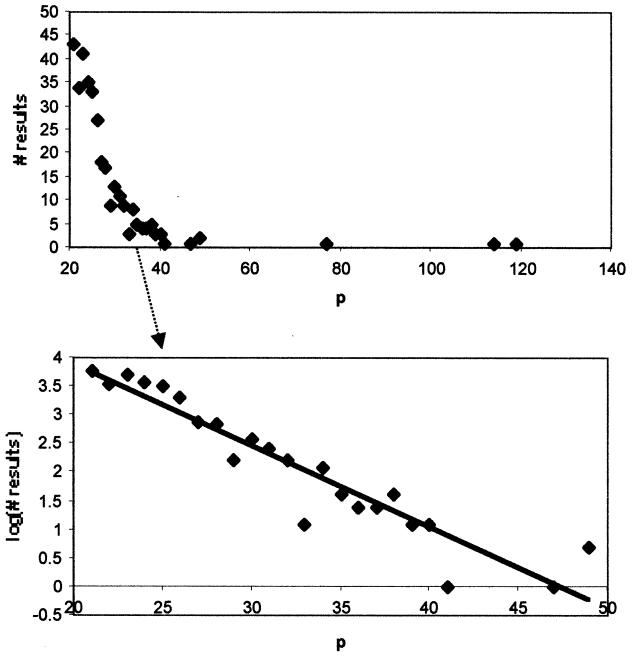


Figure 3. Expectation Value scoring, as a function of distance from the line drawn through the high end of the plot of frequency of random bits. P is the score obtained from a given mass fragment, plotted against the number of experiments that give that score. Every score obtained from a set of MS/MS data are plotted in this way (top). At the high scoring end of the essentially random (insignificant) distribution of scores, the curve matches a logarithmic curve (boxed). These high end random scores are plotted logarithmically (bottom). A best fit line is used as the base line, where the Expectation Value = 1, *i.e.* the probablility of that score being random is 100%. Expectation Values of outlying, high scores (greater than 50 in this case) can be calculated as the probablility that they would be random. Expectation Values then get smaller as the probability of a nonrandom hit increases (*e.g.* 10^{-2} is a 1 in 100 chance of being obtained at random, 10^{-3} is a 1 in 1000 chance, a smaller score is better). Scores below 10^{-10} are routinely returned.

in the database. Scores returned by ProFound were plotted against the frequency at which they occurred. A distribution of scores was obtained [25]. Similarly, for Sonar MS/MS, scores were generated for every peak analyzed, and plotted against the number of experiments giving each score. For both ProFound and Sonar, the envelope of the plot (at the high, significant, scoring end) was a logarithmic curve (Fig. 3). This envelope represents an expectation value of 1 (unit probability that the hit is random). Expectation values for high, outlying scores are back calculated, relative to the equation for the slope of the line when plotted logarithmically (Fig. 3 and legend). In both Sonar MS/MS and ProFound, high scoring outliers have a greater theoretical probability of being a true hit. Expectation values get smaller (10^{-3} to 10^{-11}) for more significant hits. Sonar MS/MS scores each peptide, and calculates the Expectation value for a protein from the scores of all peptides identified from that protein: probabilities (<1) are multiplied together then multiplied by the square root of the number of spectra searched.

RADARS was built using components that are freely available: MoverZ, ProFound, Sonar MS/MS, Protein Analysis Worksheet (PAWS), amino acid font, BioBrowser, BioML (http://www.proteometrics.com). RADARS system architecture permits addition of computers, mass spectrometers, any search engine and any database, at any time. Intranet databases speed searches and add security. RADARS can be installed on one computer, or a distributed system of networked computers and servers (Fig. 4A). Each processor increases capacity and speed. For practical reasons, communications between client and servers are achieved using common gateway interface (CGI) and hyper text transfer protocol (HTTP). Windows 2000 (NT) running a server is used for RADARS Admin, while peripheral servers (computers that do jobs for the client computer) may run UNIX. RADARS is being ported to UNIX. An Oracle database, whose interface is invisible, stores the state of the system (spectra, methods, results, and the status of each spectrum, indicating which items are currently being processed). Data is protected, so that samples cannot be dropped or lost. If a computer or server is disconnected or the HTTP link broken, upon resumption of services interrupted tasks are recommenced and continue to completion. Data organisation permits iterative analysis (Fig. 2A, 4B).

User(s) interact with RADARS over an intranet *via* standard web browser (Fig. 4A). A navigation panel permits the user to access different functions of RADARS at any time (Fig. 5A, left). Briefly, the user prompts RADARS to import spectra from the mass spectral file server (instrument data files are stored on a separate file server for back-up and disposal according to normal laboratory practice). A search tool is included to recover mass spectra and linked data/results (Fig. 58). The user sets up methods for analysis of the raw data files (by MoverZ), protein chemistry (Comparison) and sample preparation, and database search (DB Search, Fig. 5C–E). Batches of one or more spectra are selected, and queued for analysis or search (Fig. 5F). RADARS farms out jobs to application servers (Fig. 4A). After MoverZ analysis, a list of peaks representing the spectrum is stored in the Oracle database, reducing storage requirements.

Additional software tools accompany RADARS, facilitating interpretation. MoverZ, post analysis, annotates identified peaks with the corresponding amino acids and modifications (*e.g.* annotations like GlcNAc, *O*, Na,

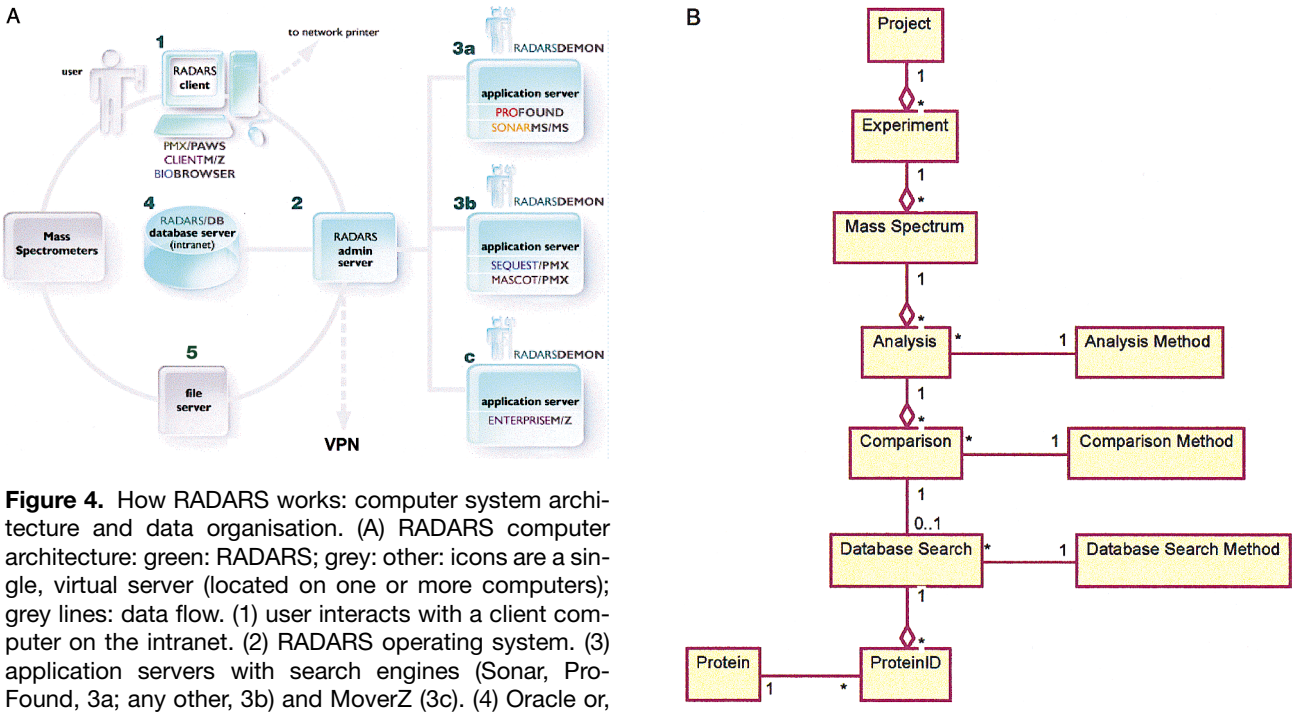


Figure 4. How RADARS works: computer system architecture and data organisation. (A) RADARS computer architecture: green: RADARS; grey: other: icons are a single, virtual server (located on one or more computers); grey lines: data flow. (1) user interacts with a client computer on the intranet. (2) RADARS operating system. (3) application servers with search engines (Sonar, ProFound, 3a; any other, 3b) and MoverZ (3c). (4) Oracle or, other relational database. For multi-server systems, each server has a daemon which contacts the Administrator to request jobs. (5) File server for mass spectral data from mass spectrometers. VPN: optional Virtual Private Network connection for remote maintenance, protected with 2-way encryption. (B) Data organisation in the Oracle database (1. . . .* indicates 1 to many relationships). A Project folder contains Experiment folders, which contain spectra. A spectrum may be analysed by MoverZ many times (Analysis, each analysis associated with a method); each Analysis may be searched multiple times against the databases: each search will contain one Comparison method, a protein identification Database Search with associated parameters. Protein identifications manually confirmed by the user remain linked with all hierarchical stages in this tree.

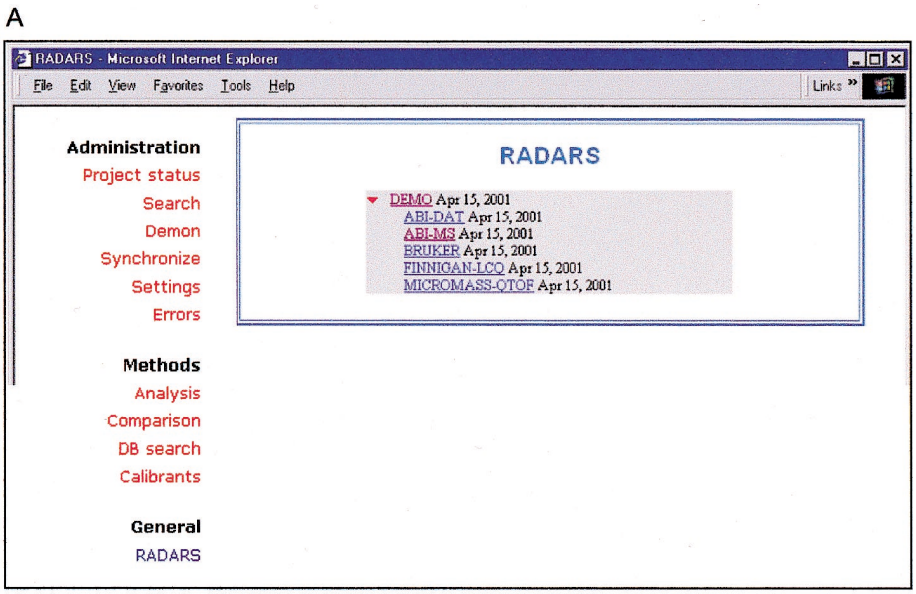
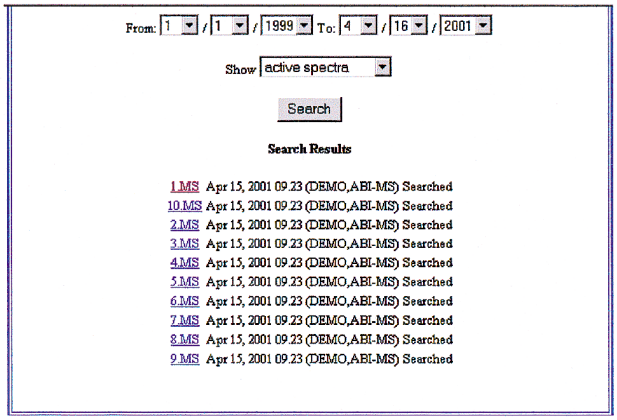
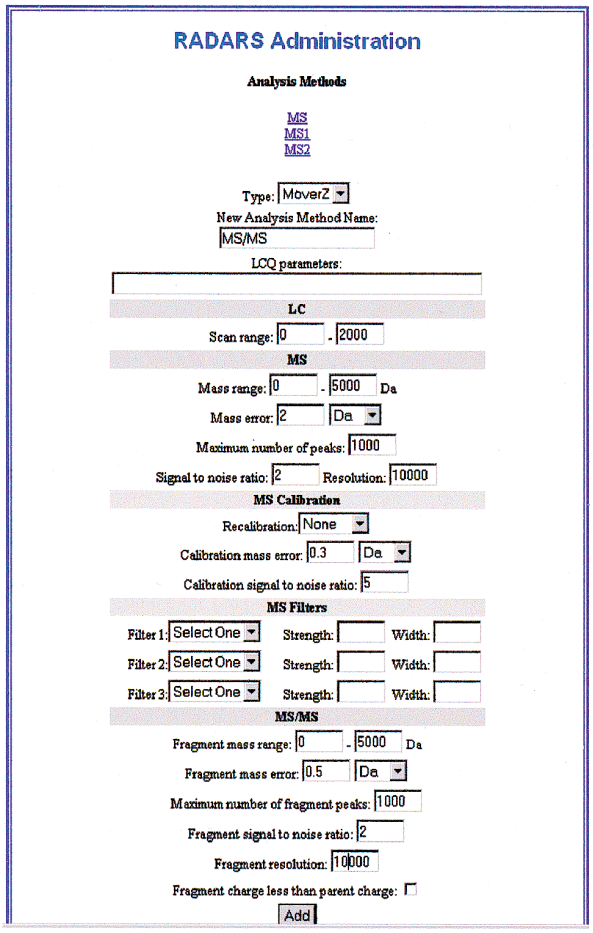


Figure 5. RADARS user interfaces in Administrator mode. (A) RADARS web browser interface. (Left) navigation, (right) user interface showing top level of access to data files (grey box), with *Project* file 'DEMO' open (arrowhead), displaying enclosed *Experiment* folders ('ABI-DAT' *etc.*) containing spectra: underlined items are hypertext linked. (Left) *Project status* (selected) takes you to the top of the folders of projects, *Search* gives the search interface (see B); *Demon* allows access to the computer daemon, to see number of jobs being processed, or stop and start the daemon. *Synchronize* adds mass spectral files to the RADARS system. *Settings* allows you to set up the daemon. *Errors* lists any errors incurred. 'Methods' interfaces link to summaries of method parameters (C-E). 'General': *RADARS* toggles between admin (shown) and user modes (user mode has no access to set up methods or schedule analyses). (B) Search for spectra in the database: (top) search parameter entry by name, date, MS type; (bottom) list of retrieved spectra. 1 spectrum

B

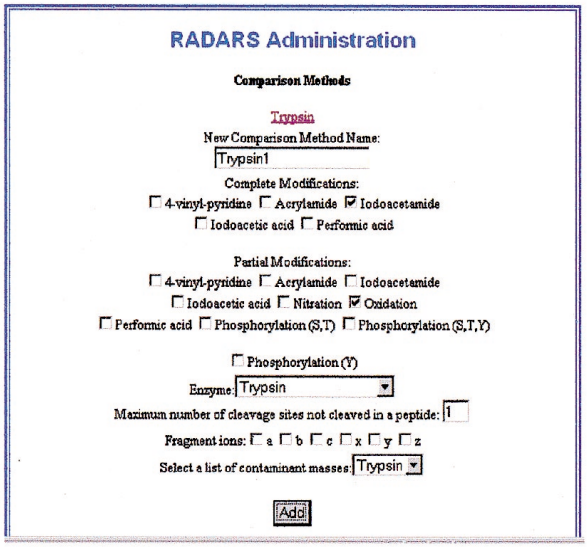


C

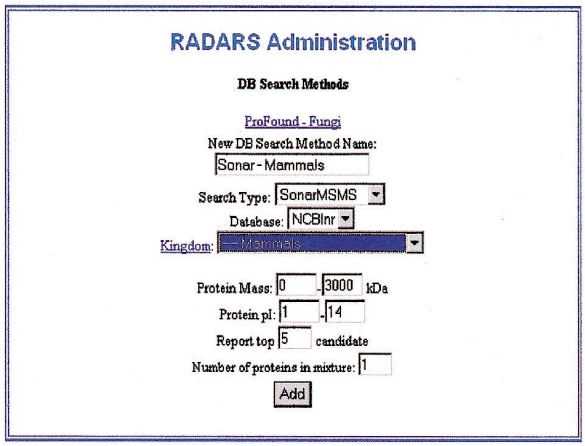


per line, annotated with: name (click through to results), date, time of last analysis; parent folder names (Project, Experiment); status 'Searched', (C) Parameter input for Analysis by MoverZ: (top) list of current methods (click-through to summary); *Type*: for instrument analysis rather than MoverZ; *New . . . Name*: your file name for method; *LCQ parameters*: for Finnegan command line entry; 'LC': *Scan range* in Da; 'MS': *Mass range*, *Mass error*, *Maximum number of peaks*, *S/N*, instrument *Resolution*, 'MS Calibration': *Recalibration* (select a stored list of

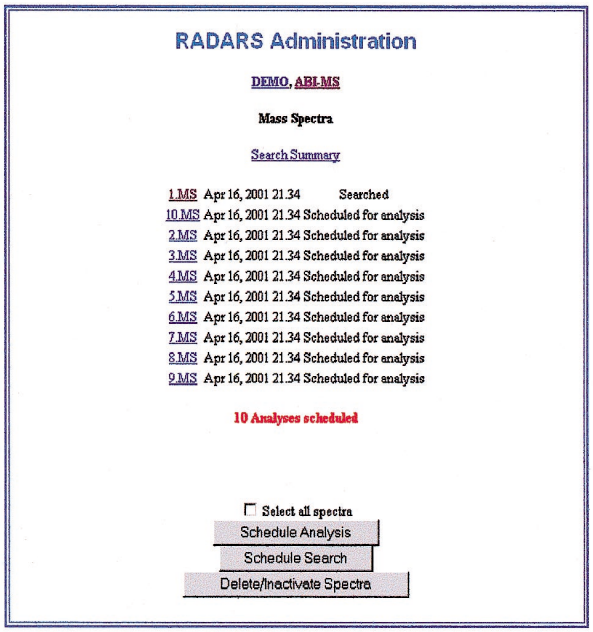
D



E



F

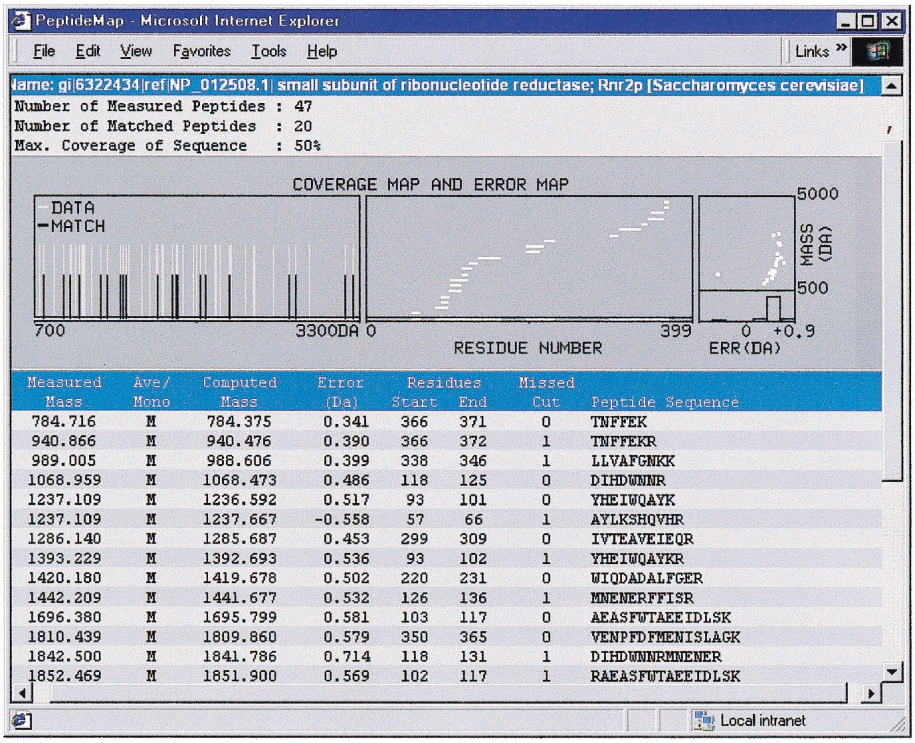


G

Top Protein Candidates

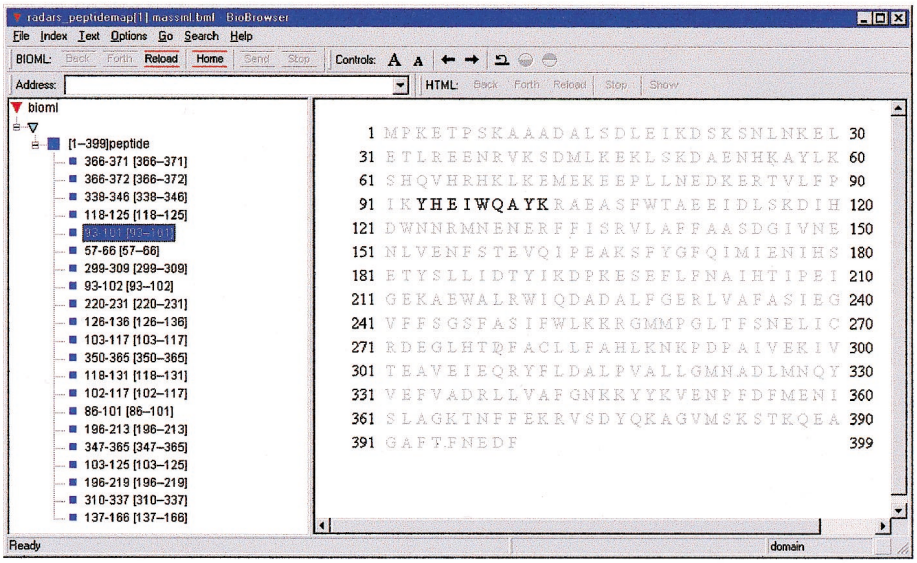
#	Z	Protein Information	%	Mass Spectrum
1	2.32	gi|6322434|ref|NP_012508.1| small subunit of ribonucleotide reductase; Rnr2p [Saccharomyces cerevisiae]	50	1 MS
2	2.32	gi|6322434|ref|NP_012508.1| small subunit of ribonucleotide reductase; Rnr2p [Saccharomyces cerevisiae]	50	10 MS

H



calibrants); *Calibration mass error*, and S/N; 'MS filters': high pass, smoothing, ABC [23]; 'MS/MS': *Fragment mass range*, *error*, *Maximum number of fragment peaks*, *Fragment S/N*, *Fragment resolution, (Fragment charge less than parent charge):* whether the parent ion always has higher charge than its fragments. *Add* button: stores method irreversibly in relational database. Parameters not required are left blank. (D) 'Comparison methods' or Protein chemistry parameter input: (from top) list of existing methods (Trypsin, click-through to summary); New comparison method ... name for new method; *Complete* and *Partial modifications*; *Enzyme*: proteolysis method; *Maximum number* ... of missed cleavage sites (0–4); *Fragment ions*: check preferred. *Select* (prestored) *list of contaminant masses*: remove them from list of protein results. *Add* button: as above. (E) Parameter entry for DB Search Method (protein identification): (from top) list of existing methods (ProFound – Fungi, click through to summary); *New DB Search method* name; *Search Type*: choose engine (*e.g.* Sonar MS/MS, ProFound); *Database*: choose from installed FASTA databases (*e.g.* NCBInr, dbEST); *Kingdom*: select taxa; *Protein Mass* (if unknown a wide range may be set); p*I* (*ditto*); *Report top*: n *candidates* (number of results to be returned, redundant proteins included as part of one candidate result); *Number of proteins in mixture* (1–4 allowed). *Add* button as above. (F) Create batch analysis and search jobs. This function is accessed from *Project Status* (navigation bar, A, left) and batches of mass spectra contained in Experiment folders. (From top) Project, Experiment folders, from which files are listed (under Search summary, which clicks-through to summary (G)). Mass spectra are annotated (as B). *Select all spectra* check box to select all spectra in list. To create a job: *Schedule Analysis* button leads to functions for: selection (or deselection) of individual spectra if required, choose Analysis Method, start. *Schedule Search* button (Search follows Analysis) leads to functions: choose individual spectra, choose Comparison and DB Search Methods, start. The first spectrum, annotated 'Searched' has MoverZ Analysis, and Comparison-DB Search complete. Mass spectrum name is hypertext linked to the results summary for that spectrum**. *10 Analyses scheduled* (red, beneath list of spectra) is a RADARS report of jobs in progress, *i.e.* 10 spectra scheduled from this folder for Analysis. (G) Summary of top hits for a folder of two spectra, hypertext linked to associated data. # column: rank of protein candidate; click through 1 to: summary sheet containing *n* results for single mass spectrum (same as**), with Expectation Values. Z: significance score [25], (green underline) >95% probability of a true hit. Protein information: GenBank descriptors, linked to BLAST, PAWS, *etc.*; %: percent coverage of identified peptides; Mass Spectrum: filename (linked to summary file for that spectrum, of all processes performed and results). (H). **PeptideMap**: (from top) protein information, descriptors, name, species; measured and matched peptide numbers, coverage. Diagrams: (left) peptide hits (black), all measured peaks (white); (centre) coverage map; (right) errors plot: peptide mass *versus* mass error (listed under 'Error'). Error dots are tightly clustered about a normal; a loose configuration indicates that a (software) recalibration of the spectra would achieve a more significant result.

I



J

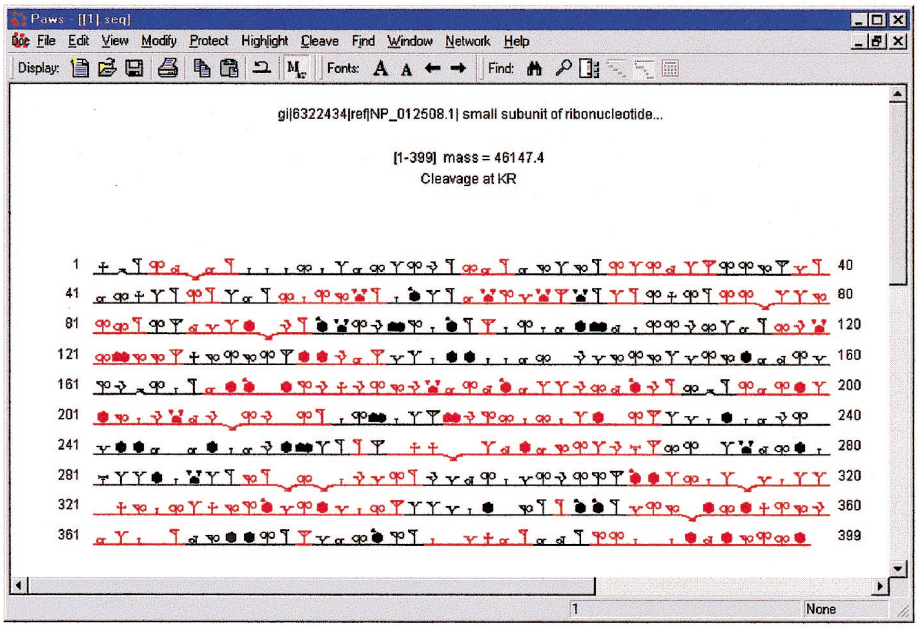


Table (left to right): measured mass; average/monoisotopic mode, mass of theoretical peptide; mass error (difference between measured and theoretical, Da); start and end peptide residue number in protein; number of missed cleavage sites; peptide sequence and modifications. (I) **BioBrowser** RADARS output, with navigation (left) for rapid identification of the location of identified peptides in the protein (right). (J) **Protein Analysis Worksheet (PAWS)** display. (From top): protein reference and name; amino acid residues of protein and mass (Da); cleavage site. (Large display) protein sequence in amino acid font (symbols represent amino acid side chains); colour boundaries indicate cleavage sites. Toolbar (left to right): 'Display': *New* PAWS window; *Open* PAWS file; *Save* PAWS sequence file with added modifications; *Print* current display; *Copy* display to System clipboard; *Paste* contents of clipboard (*e.g.* list of masses generated by MoverZ ready for Find); *Go back* one step; *Use average (or Monoisotopic) masses*; 'Fonts': Bigger, Smaller, widen protein display (increase amino acids *per* line), narrow it; 'Find': *Search for any fragment, Search for cleavage fragment, Search for list of peptides* (generates coverage maps); (greyed out) Display *coverage map plotted by fragment*. Display *coverage map plotted by amino acid length*; Display amino acid text (shown).

formyl, indicating (respectively) *N*-acetyl glucosamine (a monosaccharide), oxidation, sodium and formic acid groups, Fig. 1C). A novel visualization tool accompanies Sonar MS/MS, the fragmentogram for assessing spectral data quality (Fig. 2). PeptideMap provides a detailed analysis of peptide data compared with the identified protein: coverage maps, expected and measured masses, mass errors; peptide sequences; modifications. Error data indicate whether a search would benefit from recalibration (Fig. 5G). RADARS uses BioBrowser to rapidly identify the location of each peptide in the protein (Fig. 5I), and Protein Analysis Worksheet (PAWS) provides the same function. PAWS can download a protein sequence with one click, and display it in text form (Fig. 5J). It allows customized protein modifications and chemistries to be performed as a virtual experiment, and searches for peptide masses, highlighting the position of each peptide, and generating coverage maps. BioBrowser is capable of rapidly downloading data from various WWW databases for comprehensive review of biomolecular data, with one-click access to literature references [26].

Table 2. Exhibition of salient features of RADARS

RADARS Feature	Performance
System	
Minimum RADARS system	Single 700 MHz processor PC computer (desktop or laptop workstation) with server software and an HTTP connection to the WWW.
Largest RADARS system (supercomputer possible)	Any number of computers and processors, with an appropriate HTTP (intranet) bridge. Largest currently an extensive Linux cluster.
Supports multiple processors	Performance increased (faster)
Scalability	No theoretical limit on computer number or capacity. Growable. Uses an application server administrator architecture.
Storage relational databases	Oracle, SyBase, SQLServer, RDB, DB2, running on Windows NT, Windows 2000, UNIX and VMS operating systems.
Security	2-way encrypted or solely intranet
External maintenance if required	By Virtual Private Network
Functional protein identification	
Mass spectrometer data input	Any major manufacturer
Monoisotopic peak detection	As good as inspection by experienced user, especially for peaks at high mass. Improves search engine results.
MS/MS and MALDI-MS search engines	Sonar MS/MS, ProFound supplied. Bridges exist for SEQUEST, PepSea, Mascot, Protein Prospector, *etc.*
Genomic databases for searching	Can use any database, in-house or public. Database services are supplied for: *E. coli, S. aureus, D. melanogaster, C. elegans, T. brucei, P. falciparum, M. musculis, H. sapiens* (public data as available from NCBI)
Customization of software	Yes. Can fit into any LIMS, deal with minor mass spectrometer manufacturers, *etc.*
Sonar MS/MS advantages	
RAM used in searches	40 Mbytes *per* process
Exon mapping	Sonar MS/MS genomic search
New ORF detection	Sonar MS/MS genomic search
Splice sites	Sonar MS/MS genomic search
Methods for improving hit rate	Use genomic sequence rather than hypothetically translated protein sequence. Use taxonomy specific database.
Parent ion mass cannot be determined because of post-translational modification, *etc.*	Sonar MS/MS is independent of parent ion mass: a window of $\pm$1000 Da gives the same search result.
Cost reduction	By using MS/MS directly, without predetermination of ion mass by MS. Sequencing not required for DNA search.

The biologist might review protein hits and their scores, and use BioML. The protein chemist or mass spectroscopist may reexamine MS data and schedule further work. A bioinformatician may create SQL queries to create customized reports from the relational database. In summary: the development of an automated software system for detecting proteins at HT level has been successful, and has produced improvements and innovations in protein identification software. Proteome MS automation systems like RADARS provide as-good-as-human data processing and quality control. RADARS uses a flexible computer architecture of standard components, and allows consistent, mass spectral data processing. Results are automatically stored in a searchable, flexible relational database for sample tracking and customized reporting. RADARS technology is being used sustainably in 24 h/7 d situations where speed of analysis, qualitative scoring, performance with biologically modified samples and robustness are critical. It is also being used in lower throughput situations, where automated storage and accessibility across the intranet are valuable.

RADARS is useful for: protein/gene identification "factory" lines with 100's of thousands of samples *per* day on hundreds of computers; genome annotation by exon and splice-site mapping; ORF calling from experimental data (*i.e.* protein samples); protein identification on genomes in any stage of construction; protein identification for biologically modified samples. RADARS provides accurate protein identification by a less experienced operator. The major features of RADARS (and Sonar MS/MS) are summarized in Table 2. This automation package has no theoretical capacity limits, and has been scaled to include large computer clusters with multiple storage database instances. RADARS effectively and accurately speeds the process of MS and MS/MS data analysis, providing novel identifications where previously, none were found.

We gratefully acknowledge receipt of NIH SBIRS Phase II grant no. 2R44RR13503-02. The authors would like to thank Drs. Jennifer Krone and Mark Field for comments and suggestions.

Received May 20, 2001

References

[1] Pandey, A., Mann, M., *Nature* 2000, *405*, 837–846.

[2] Andersen, J. S., Mann, M., *FEBS Lett.* 2000, *480*, 25–31.

[3] Shevchenko, A., Wilm, M., Mann, M., *J. Protein Chem.* 1997, *16*, 481–490.

[4] Rosenkrands, I., King, A., Weldingh, K., Moniatte, M. *et al.*, *Electrophoresis* 2000, *21*, 3740–3756.

[5] Schulz-Knappe, P., Zucht, H. D., Heine, G., Jurgens, M., *et al.*, *Comb. Chem. High Throughput Screen* 2001, *4*, 207–217.

[6] Huang, C., Shui, H., Wu, Y., Chu, P., *et al.*, *Brain Res. Mol. Brain Res.* 2001, *92*, 181–192.

[7] Li, L., Masselon, O. D., Anderson, G. A., Pasa-Tolic, L., Lee, S. W., *et al.*, *Anal. Chem.* 2001, *73*, 3312–3322.

[8] Figeys, D., Pinto, D., *Electrophoresis* 2001, *22,* 208–216.

[9] Zhu, H., Bilgin, M., Bangham, R., Hall, D., *Science* 2001, *10*, 1126–1130.

[10] Rout, M. P., Field, M. C., *J. Biol. Chem.* 2001, *10*, 1074–1079.

[11] Rout, M. P., Aitchison, J. D., Suprapto, A., Hjertaas, K., *et al.*, *J. Cell Biol.* 2000, *148*, 635–651.

[12] Rappsilber, J., Siniossoglou, S., Hurt, E. C., Mann, M., *Anal. Chem.* 2000, *72*, 267–275.

[13] Cohen, S. L., Chait, B. T., *Anal. Biochem.* 1997 *247*, 257–267.

[14] Karas, M., Hillenkamp, F., *Anal. Chem.* 1988, *60*, 2299–2301.

[15] Beavis, R. C., Chait, B. T., *Methods Enzymol.* 1996, *270*, 519–551.

[16] Jennings, K. R., Mason, R. S., in: McLafferty, F. W., (Ed.), *Tandem Mass Spectrometry,* Wiley, New York 1983, p. 197.

[17] Zhang, W., Chait, B. T., *Anal. Chem.* 2000, *72*, 2482–2489.

[18] Beavis, R. C., Fenyö, D., *Proteomics: a Trends Guide* 2000, *1*, 22–27.

[19] Perkins, D. N., Pappin, D. J. C., Creasy, D. M., Cottrell, J. S., *Electrophoresis* 1999, *20*, 3551-3567.

[20] Gatlin, C. L., Kleemann, G. R., Hays, L. G., Link, A. J., Yates, J. R., *Anal. Biochem.* 1998, *263*, 93–101.

[22] Fenyö, D., Zhang, W., Chait, B. T., Beavis, R. C., *Anal Chem.* 1996, *68*, 721A–726A.

[23] Carroll, J. A., Beavis, R. C., *Rapid Comm. Mass Spectrom*. 1996, *10*, 1683–1687.

[24] Beavis, R. C., *Anal. Chem.* 1993, *65*, 65–66.

[25] Eriksson, J., Chait, B. T., Fenyö, D., *Anal. Chem.* 2000, *72*, 999–1005.

[26] Fenyö, D., *Bioinformatics* 1999, *15*, 339–340.