Functional Assignment of the 20 S Proteasome from Trypanosoma brucei Using Mass Spectrometry and New Bioinformatics Approaches*

Lan Huang‡, Richard J. Jacob‡, Scott C.-H. Pegg‡, Michael A. Baldwin‡, Ching C. Wang‡, Alma L. Burlingame‡, and Patricia C. Babbitt‡¶

From the ‡Department of Pharmaceutical Chemistry and the §Department of Biopharmaceutical Sciences, University of California, San Francisco, California 94143

As experimental technologies for characterization of proteomes emerge, bioinformatic analysis of the data becomes essential. Separation and identification technologies currently based on two-dimensional gels/mass spectrometry provide the inherent analytical power required. This strategy involves protein spot digestion and accurate mass mapping together with computational interrogation of available data bases for protein functional identification. When either no exact match is found or when the possible matches only partially account for molecular weights actually observed, peptide sequencing by tandem mass spectrometry has emerged as the methodology of choice to provide the basic additional information required. To evaluate the capabilities of bioinformatics methods employed for identifying homologs of a protein of interest, we attempted to identify the major proteins from the 20 S proteasome of Trypanosoma brucei using sequence information determined using mass spectrometry. The results suggest that neither the traditional query engines, BLAST and FASTA, nor specialized software developed for analysis of sequence information obtained by mass spectrometry are able to identify even closely related sequences at statistically significant scores. To address this deficit, new bioinformatics approaches were developed for concomitant use of the multiple fragments of short sequence typically available from methods of tandem mass spectrometry. These approaches rely on the occurrence of congruence across searches of multiple fragments from a single protein. This method resulted in sharply better statistical significance values for correct hits in the data base output relative to that achieved for independent searches using single sequence fragments.

Received for publication, September 12, 2000, and in revised form, March 26, 2001
Published, JBC Papers in Press, April 17, 2001, DOI 10.1074/jbc.M008342200

Fueled by the genome projects, encyclopedic increases in the banking of newly obtained, comprehensive biological data are transforming studies of biology and medicine (1). As the post-genomic era moves into high gear, new “high throughput” technologies are allowing characterization of gene expression profiles, comparisons of genomic complements, and identification of the genetic markers associated with normal, pathological, or environmentally triggered states. Yet information derived from full analysis of genomics alone is clearly inadequate to explain the complexities of cell biology. Recent studies showing differences between the genome and the proteome suggest that the profound understanding we seek will require the complete and direct characterization of the proteome as well (2, 3).

Peptide mass mapping by MALDI-TOF MS (4) or liquid chromatography-electrospray ionization MS (5, 6), combined with interrogation of sequence data bases (7–12), currently is the most widely employed strategy for the identification of expressed proteins. This methodology involves electrophoretic separation of proteins at sub-picomole levels, digestion with trypsin, and measurement of the molecular weights of the resulting peptide mixture by mass spectrometry. This strategy can routinely identify proteins whose actual sequences are present in protein, gene, or EST data bases (13–16, 12, 17, 18). In addition, one would also like to find related proteins even when the actual protein sequence is absent from the data bases. Here, the identity of the protein of interest can frequently be inferred if the similarities to homologs in available data bases are sufficiently high. Such identification by inference cannot be achieved solely by mass mapping, however, because patterns of tryptic digestion are generally not conserved; hence, only nearly identical homologs are likely to be found. Thus, additional experimental information, viz. amino acid sequence, is required to identify proteins whose sequences are not present in the data bases per se or that may differ even slightly from related sequences.

Over the last 5 years, the primary source of sequence information obtained directly from protein has been generated using methods of tandem mass spectrometry, based upon collision-induced dissociation of tryptic fragments in protein digests. During this time, mass spectrometry has emerged as the only technology sufficiently sensitive, rapid, and robust to provide this information on a proteome-wide scale. New instru-

* The mass spectrometry performed at the University of California, San Francisco, Mass Spectrometry Facility, was supported by National Institutes of Health Grant RO1 GM60595 (to A. L. B.), bioinformatics was supported by National Institutes of Health Grant R01 GM60595 (to P. C. B.), and biochemistry was supported by National Institutes of Health Grant R01 AI21786 (to C. C. W.). The costs of publication of this article were defrayed in part by the payment of page charges. This article must therefore be hereby marked “advertisement” in accordance with 18 U.S.C. Section 1734 solely to indicate this fact.

The nucleotide sequence(s) reported in this paper has been submitted to the GenBank®/EBI Data Bank with accession number(s)AF198386, AF148125, AF198387, AF198388, AF140353, AJ131148, AF169651, AF169652, AF198389, AF198388, AF169650, AF226673, AF226674, AF148124, AF226674, AF148124, AF290945.

¶ To whom correspondence should be addressed. Tel.: 415-476-3784; Fax: 415-476-0688; E-mail: babbitt@cg1.ucsf.edu.

The abbreviations used are: MALDI-TOF, matrix-assisted laser desorption/ionization-time-of-flight; ACN, acetonitrile; CID, collision-induced dissociation; EST, expressed sequence tags; HPLC, high performance liquid chromatography; MS, mass spectrometry; PSD, post-source decay.

This paper is available on line at http://www.jbc.org
mentation is currently being developed to accommodate the high throughputs required (19). Sequence information, which is much better conserved among related proteins than are tryptic digestion patterns, can then be used for identification of proteins with close or sometimes even remote homologs in the data bases. Searching tools currently in use with such partial sequence data include simple string searches (20) and algorithms such as BLAST (21), Gapped BLAST (22), and FASTA (23). When closely related homologs exist in the data bases or when the available sequence fragments represent highly conserved regions of a homolog, these approaches can be quite successful. However, when only more remote homologs are present or when homologs exist only as partial sequences, as in EST data bases, sequence information from mass spectrometric methods may be insufficient to identify the protein of interest due to two major problems. First, the statistical significance of matches in data base searches is length-dependent (24), whereas segments of sequence obtained by mass spectrometry of tryptic peptides are generally short, often 10 amino acids or less. As a consequence, searching algorithms often fail to find homologs at statistically significant scores. Second, the location and coverage of peptides relative to the parent protein is highly variable and may represent regions of low or no similarity to a homolog in the data base. Thus, a score of poor statistical significance could signify an absence of homologs in the data base or that the particular peptides sequenced may represent poorly conserved regions of even fairly close homologs. The inability to distinguish between these two possibilities could result in misleading interpretation of the novelty of the sequence of interest. In a high throughput experiment, the consequences could also be substantial, resulting in a decision to divert resources elsewhere or to over-commit resources toward further analysis of the protein or protein system in question.

Here we describe new strategies for more effective identification of both close and remote homologs from partial sequence information using the identification and preliminary structural analysis of the 20 S proteasome from Trypanosoma brucei (25) as an example. Proteasomes are multicatalytic proteolytic complexes found in the cells of most organisms and play important roles in protein turnover in both cytosol and nucleus. These roles include the removal of abnormal, misfolded, or improperly assembled proteins associated with a wide range of biological functions in stress response, cell differentiation, metabolic adaptation, and cellular immune responses (26, 27). The proteasome provides a good model system for this study because it is a macromolecular assembly whose component subunits cannot be directly inferred from genomic information alone, thus requiring detailed characterization of its protein components. Furthermore, its relative complexity provides an appropriate challenge for initial development of high throughput strategies for characterization of an important protein machine (28).

A combination of modern mass spectrometric techniques and newly developed informatics approaches was used to identify and characterize subunits of the proteasome. This is the first time that such a complex system has been investigated in this manner and offers evidence for the value of the approach. Further development of these informatics strategies could extend substantially the usefulness of mass spectrometric sequence information for identifying proteins in species whose genomes have not been sequenced.

**EXPERIMENTAL PROCEDURES**

*Protein Separation by Two-dimensional Gel Electrophoresis and in Gel Digestion*

Preparation and electrophoretic separation of 20 S proteasomes from *T. brucei* was performed as described previously (25, 29). After silver staining, protein Spots were excised from the gel. Minced gel pieces were washed with 25 mM NH₄HCO₃, 50% ACN, dried in a Speedvac, and then rehydrated in 25 mM NH₄HCO₃ solution containing trypsin. After overnight digestion at 37 °C, peptides were extracted, once with HPLC grade water and then 3 times with 50% ACN, 5% trifluoroacetic acid. The combined supernatants were dried by Speedvac and then re-dissolved in 50% ACN, 5% trifluoroacetic acid. Aliquots were analyzed directly without separation. The remainder was separated by reverse phase HPLC on a Vydac microbore C18 column (1.0 mm × 15 cm), fractions from which were collected and concentrated.

**Mass Spectrometric Analysis of Tryptic Peptides**

Molecular masses of tryptic peptides were determined by analyzing 1 μg of each unseparated digest and 1/10th of each HPLC fraction by MALDI-TOF MS (Voyager DESTR, Perseptive Biosystems, Framingham, MA). Peptides were cocrySTALLized with equal volumes of a saturated solution of a-cyano-4-hydroxycinnamic acid in 50% ACN, 1% trifluoroacetic acid. MALDI spectra were internally calibrated (<20 ppm) using trypsin autolysis products. Peptide masses were submitted for data base searching using the MS-Fit program (12, 14). Selected peptides were subjected to PSD analysis for partial sequence determination. High resolution-high sensitivity CID tandem MS was performed on a prototype QqTOF mass spectrometer (Sciex, Toronto, Canada) (30) equipped with a nanoelectrospray ion source (Protana A/S, Odense, Denmark), using 2–3 μl of each HPLC fraction. A conventional mass spectrum was first obtained to identify peptide masses by multiple charging, and then CID spectra were collected. Mass resolution of ~30,000 allowed unambiguous charge state determination for ions at least up to m/z 5000. Mass accuracy was typically 20 ppm or better with external calibration and was usually sufficient to distinguish between glutamine and lysine, phenylalanine and oxidized methionine, and between several combinations of two amino acid residues with the same nominal masses.

In determining peptide sequences, all spectra were interpreted manually using known rules and established principles (31, 32). Potential ambiguities exist for isomeric Ile/Leu and isobaric Gln/Lys, and frequently in the assignments of terminal residues. Here PSD was the primary technique, in which the reflectron voltage in a MALDI-TOF mass spectrometer was scanned to select fragment ions (33), supplemented by high quality tandem MS from the QqTOF mass spectrometer (34). PSD data is generally inferior to the mass calibration is variable, whereas newer instruments such as the QqTOF give high quality tandem data with sufficiently accurate mass measurements to increase the success of data base searching (34) and to assist manual interpretation by removing the ambiguities arising from isobaric ions. The MALDI-TOF/TOF tandem mass spectrometer is another new alternative that has the additional advantage of high energy CID giving more comprehensive fragmentation, allowing Ile and Leu to be differentiated (19, 35).

**Data Base Searching and Analysis of Sequence Information**

A static data base, non-redundant Static, was generated by downloading the non-redundant National Center for Biotechnology Information protein data base (as of 3/11/99) and modified by removing sequences matching 3 of our query sets that had been submitted to the data base following their identification early in this project (29). All searches used either the BLOSUM62MS scoring matrix, a version of BLOSUM62 (36), modified to account for Ile/Leu and Gln/Lys ambiguities associated with determination of protein sequences using mass spectrometry (37) or the PAM30MS matrix, modified in an analogous manner from the PAM30 matrix (38).

**String Searches Using MS-Pattern**

MS-Pattern is a program within the ProteinProspector suite (prospector.ucsf.edu) (20) that searches the data bases and retrieves entries associated with a text string comprised of a keyword, accession number, or sequence.

**Gapped BLAST for Single and Catenated Sequences, Automation of Multiple Searches with Multi-BLAST and Congruence Analysis Using MS-Shotgun**

All types of BLAST searches were performed using version 2.0a19 of BLASTP as implemented by Washington University, St. Louis. Further information about this implementation is available (blast.wustl.edu/blast/README.html). Congruence analysis of Gapped BLAST (BLAST 2.0) searches was carried out for all sequence fragments from a given protein. The MultiBLAST Suite was written in Perl

---

2 The original Shotgun program is available by E-mail request. As the research issues described under “Results and Discussion” are solved, additional software will be developed for public use.
Isolate, digest, and fractionate the 20S proteasome

Determine molecular weights of tryptic peptides by mass spectrometry

Derive sequences of peptide fragments from major spots by tandem mass spectrometry

Identify candidate sequences possibly homologous to peptide fragment sequences

Map sequences of peptide fragments onto a multiple alignment of the “best” candidate sequences identified from database searches

Evaluate relationships from multiple alignment information

A

B

Individual sequence files from a single spot

All permutations and catenations of the individual sequences from a single spot

n files

Single BLAST searches (n)

Shotgun congruence results

Multiple BLAST searches (n!)

n! files

Shotgun congruence results

Final result files

Evaluation of Functional Assignments Using Multiple Alignments—As required, functional assignments inferred from the data base search outputs were further evaluated using multiple alignments of top scoring protein hits from each MS-Shotgun with catenation search. Percent identities were calculated for these hits to identify a divergent set of sequences from available species when possible. These were aligned with the individual peptide sequences using either Pileup (41) or ClustalW (42). To evaluate the correctness of the linear ordering, we subsequently compared the ordering of each catenated sequence to the ordering of those same peptides within the sequences of the full-length proteasome protein subunits that were obtained independently from DNA sequencing after the completion of this study (29). While the data base search scores for these full-length sequences were useful for comparison with scores obtained for sequence fragments obtained from sequencing by mass spectrometry, these full-length sequences were not used in the bioinformatic analysis of the sequences obtained from mass spectrometry.

RESULTS AND DISCUSSION

The overall strategy used for identifying unknowns is illustrated in the flow chart in Fig. 1A. The protocols for congruence and Python to facilitate the generation and submission of multiple permutations and catenations of peptide sequences to Gapped BLAST, input as text files containing all sequence fragments for a single Spot. Default options for all Gapped BLAST searches were \( E = 1000, B = 200, \) and \( V = 200. \) MS-Shotgun, a modification of Shotgun (39), was written for congruence analysis of MS data. For simple analysis, searches used individual peptide sequences as queries and then determined congruence across these searches by determining which hits were found in common by more than one peptide sequence fragment. The number of query sequences that found each hit is designated the “Shotgun Score.” For MS-Shotgun with catenation analysis, files containing the total number of peptide sequences from a single Spot (\( n \)) were generated, permuted, and catenated to give all possible sequence orders. Gapped BLAST searched each of the \( n! \) catenations, and then congruence analysis was performed across all output files from either the set of single peptides or the \( n! \) catenations from the peptide sequences from a given Spot. Hits were sorted either by Shotgun Score or by \( p \) values.

During preparation of this manuscript, other approaches using congruence analysis came to our attention. These include extensions to the CIDentify program developed by Johnson and Taylor (40), i.e. the CID Results Compiler, and the FASTS program, a modified version of the FASTA search engine. For comparison with our methods, data base searches were also performed with the FASTS program. Like MS-Shotgun, FASTS exploits the advantages of using all of the available sequence information from a single Spot to enhance the sensitivity and significance of the searches. The modifications reflected in FASTS relative to the parent FASTA search engine include allowing submission of multiple short sequences from one protein as a single query to the search engine and removal of the penalty for gaps between query sequences. Pairwise similarity scores for all the peptides that show matches to a given hit are included in the calculation of statistical significance. Pairwise alignments for all the hits are also reported in the FASTS output, showing which query sequences were used in the match and where they align. Because neither a description nor an evaluation of the FASTS algorithm has yet been published by its authors, we did not consider it appropriate to present extensive comparisons between our MS-Shotgun with catenation analysis and FASTS. For this reason, data from our FASTS analysis of the proteasome data are not provided in our figures or tables. To give the reader some idea of the relative performance of MS-Shotgun with catenation and FASTS on our data, important points of similarity and difference between the two approaches are provided under “Results and Discussion,” however.

The parent search engine, FASTA, was also tested using all individual peptide sequences as queries to confirm that there was no absolute advantage for this study in using FASTA over BLAST. This test was performed using both the BLOSUM62MS and the PAM30MS scoring matrices.

Evaluation of Functional Assignments Using Multiple Alignments—As required, functional assignments inferred from the data base search outputs were further evaluated using multiple alignments of top scoring protein hits from each MS-Shotgun with catenation search. Percent identities were calculated for these hits to identify a divergent set of sequences from available species when possible. These were aligned with the individual peptide sequences using either Pileup (41) or ClustalW (42). To evaluate the correctness of the linear ordering, we subsequently compared the ordering of each catenated sequence to the ordering of those same peptides within the sequences of the full-length proteasome protein subunits that were obtained independently from DNA sequencing after the completion of this study (29). While the data base search scores for these full-length sequences were useful for comparison with scores obtained for sequence fragments obtained from sequencing by mass spectrometry, these full-length sequences were not used in the bioinformatic analysis of the sequences obtained from mass spectrometry.

RESULTS AND DISCUSSION

The overall strategy used for identifying unknowns is illustrated in the flow chart in Fig. 1A. The protocols for congruence

3 W. Pearson, unpublished information (available online, fasta.bioch.virginia.edu/).
4 C. C. Wang, unpublished information.
analyses using the MS-Shotgun program are illustrated graphically in Fig. 1B.

**Separation, Digestion, and Analysis of Tryptic Peptides**

The 20 S proteasome of *T. brucei* separated by two-dimensional gel electrophoresis and stained with silver yielded 26 protein Spots (Fig. 2), the pattern of which was highly reproducible from batch to batch although minor Spot intensities varied across different preparations. Compared with two-dimensional gel analysis of the 20 S proteasome from rat (18), the two-dimensional fingerprint showed far fewer Spots, indicating that the trypanosome proteasome is likely to be less complex. Fifteen major Spots were excised from the gel, digested overnight with trypsin, and analyzed by MALDI-TOF. The peak list for each protein digest was searched against the calculated peptide mass values for proteins in the data bases, using the known specificity for trypsin to cleave at arginine and lysine.

When this experiment was performed, no cDNA sequences of known specificity for trypsin to cleave at arginine and lysine were present in public databases, so mass mapping was limited to searches for homologous proteins that might show conserved trypsic peptides. Not surprisingly, no proteasome entries from other species were found; therefore, after separation of digests by capillary HPLC, peptide sequencing by tandem mass spectrometry was undertaken on all abundant peaks from each Spot (2–8 sequences were obtained per Spot). The length of sequences ranged from 5 to 18 residues, averaging ~10 amino acids (Table I). The variability in the number and length of the sequence fragments from each Spot was typical of an analysis of this type.

**Identification of Possible Homologs from Searches**

By using new informatics approaches developed here, we could identify the functions of the proteasome proteins represented by 9 of the 15 major Spots marked with arrows in Fig. 2 as follows: Spots 1, 2, 4, 6, 7, 11, 12, 15, and 17. Of the remainder, functions for 3, 5, 8, 9, and 13 could not be predicted from sequence information (see “Spots That Required Additional Biological Information for Provisional Identification” below). Spot 10 appears to contain identical sequences to those in both Spots 11 and 12, suggesting either that this protein in *T. brucei* has a different origin from other known proteasomes or that artifacts, such as contamination with other proteins, were introduced during protein isolation and separation. Although such complexities will likely be a difficult and ubiquitous problem for high throughput proteomics, analysis of these problems is beyond the scope of this study. Therefore, data are not presented for Spot 10.

Our new informatics approaches are based on the hypothesis that examination of congruence across searches for a set of peptides from the same protein is useful for distinguishing real homologs from unrelated hits, even though scores for individual data base searches using each of these peptide sequences may fall below the level of statistical significance. This idea has been applied successfully using the original implementation of the Shotgun program to the problem of deducing function of proteins when only remote homologs of those proteins are present in the data bases (39). In those studies (43–45), however, Shotgun analysis used full-length proteins as queries rather than short sequence fragments such as described in this study.

Data base search engines such as BLAST and FASTA are designed to perform pairwise comparisons between a query sequence and each sequence in a data base. As described below, the searches performed in this study could be divided into the following three types with respect to the nature of the query sequences and processing of data base search output: 1) searches using single sequences of tryptic peptides as queries, performed one at a time; 2) congruence analysis across the search outputs of all of these single searches from the same Spot; or 3) congruence analysis across the outputs of all possible permutations/catenations of the peptide sequences from the same Spot (MS-Shotgun with catenation).

The results are summarized in Tables II–IV and discussed below. Table II compares single searches using simple BLAST analysis and congruence analysis using MS-Shotgun with catenation for the 9 Spots identified using those tools. Because we wanted to compare simple BLAST searches with MS-Shotgun with catenation searches using the same conditions, we chose to use the BLOSUM62MS scoring matrix for both types of search because of the following: 1) it is the default matrix of choice for the occasional or non-expert user; 2) no clearly optimal choice of matrix can be predicted from theory for use with the fragmentary sequence information represented by our data (see under “A Note on the Importance of the Scoring Matrix in Data Base Searches Using Peptide Sequences Obtained from Tandem Mass Spectrometry”); and 3) we had seen insufficient differences between the BLOSUM62MS and the theoretically more optimal PAM30MS matrix in searches using individual peptide sequences for many of the Spots we investigated. However, because the searches of individual peptide sequence fragments using the PAM30MS matrix did give slightly improved *p* values on a consistent basis, results using this matrix are also included in Table II.

Functional assignments for all of the Spots are given in Table III and are based on the functions of the proteasome sequences appearing in the top 10 scoring hits from the MS-Shotgun with catenation analysis for each Spot or as described below. As required for evaluation, multiple alignments of the peptide sequences obtained from mass spectrometry and the set of closest likely homologs determined from our analysis were used to confirm the catenation order and mapping of the short peptide sequences onto the homologs. Although none of the functional assignments provided in Table III have been verified unequivocally by experimental characterization of the proteins, recent determinations of full-length cDNA sequences for all these genes allow high credibility assignments to be made from data base searches, independent of this study. The correctness of these assignments is supported by the statistically significant similarities of these full-length sequences to other proteasome sequences (also shown in Table III).

For Spots difficult to identify solely from sequence information, highly provisional assignments of subunit function could
| Spot no. | Peptide sequences obtained by de novo sequencing* | Sequence versions used for database searching | Method(s) used for sequence determination |
|---------|--------------------------------------------------|---------------------------------------------|------------------------------------------|
| 1.0     | GVNTFSPEGR                                       | GVNTFSPEGR                                  | PSD                                      |
| 1.1     | NMTVEEAEN(I/L)TVQ(I/L)LVR                       | NMTVEEAENLTVQLLR                            | PSD                                      |
| 1.2     | YDA(Q/K)IA(I/L)GGGAAEQA/Q/TVFSER               | YDAQALGGGAAEAAQTFSER                       | PSD                                      |
| 2.0     | (I/L)REER                                        | LREER                                       | PSD                                      |
| 2.1     | (I/L)EACAG(I/L)K                                | LGEAQLK                                    | PSD                                      |
| 2.2     | (DTY,K/K/T/TFQ/K/K)                             | TDQTQFFAQ                                  | PSD                                      |
| 2.3     | (S/I/L)SGYEEVA(I/L)TVK                        | DLSGGEVEAVLAVK                             | Quadropole orthogonal acceleration time of flight (QqoaTOF) and PSD |
| 2.4     | (FAA/A/UL/YVSM-oX(I/L)YEAR                   | AFAALVSSMVYEAR                             | Quadropole orthogonal acceleration time of flight (QqoaTOF) |
| 3.0     | ARGEAENSR                                       | ARGEAENSR                                  | PSD                                      |
| 3.1     | Acetyl-SGSGHDQ/K/STLD/LSADGR                   | SGSGHDSQSTLFSADGR                          | PSD                                      |
| 3.2     | D(I/L)FSTP(I/L)(I/L)GS(I/L)(I/L)ASR            | DLFSPTPLGSLLASR                            | Quadropole orthogonal acceleration time of flight (QqoaTOF) and PSD |
| 3.3     | (I/L)(L/H)AQVHV                                 | LHLAQHV                                   | PSD                                      |
| 4.0     | (Q/K)(I/L)DGFK                                  | QYLDGFK                                    | PSD                                      |
| 4.1     | TY/I/LER                                        | TYLER                                      | PSD                                      |
| 4.2     | H(I/L)Q/K/FAGK                                  | HLQFAGK                                    | PSD                                      |
| 4.3     | NTR(I/L)A/I/UL/VKG                             | NTLALVGK                                   | PSD                                      |
| 4.4     | NQYDITDTTTWSPTGR                                | NQYDITDTTWSPTGR                            | Quadropole orthogonal acceleration time of flight (QqoaTOF) |
| 4.5     | GPFACD(I/L)DE/I/UL/VM-oX(HA/I/L)R               | GPFACDLDDELVMHALR                           | Quadropole orthogonal acceleration time of flight (QqoaTOF) |
| 4.6     | AFK(I/L)DEHGV-oXAI/I/UL/G(S/I/L)VADGR           | AFKLDHEVGLMSVADGR                          | Quadropole orthogonal acceleration time of flight (QqoaTOF) |
| 5.0     | TTTFSPEGR                                       | TTTFSPEGR                                  | PSD                                      |
| 5.1     | (QV)AP/I/LTQK                                  | VQAPLTQK                                   | PSD                                      |
| 5.2     | (I/L)EAYAV(I/L)/Q/KR                           | LEVAVLQR                                  | PSD                                      |
| 5.3     | (Q/L)YYQGGYR                                   | QLYTQQGYVR                                 | PSD                                      |
| 5.4     | (I/L)T/DEE/I/UL/QPHDEVANR                      | LI/DEELQPHDEVANR                           | PSD                                      |
| 5.5     | M-oX/VPSFK                                      | MVPSPFK                                    | PSD                                      |
| 5.6     | (I/L)(L/I)DPYVNPK                               | LLPDPYVNPK                                 | Quadropole orthogonal acceleration time of flight (QqoaTOF) and PSD |
| 6.0     | WPNI/L/JPR                                      | WPNIJPR                                    | PSD                                      |
| 6.1     | (I/L)SGGSI/UL/LJ/G/I/LK                        | LSGGGVIgLK                                 | Quadropole orthogonal acceleration time of flight (QqoaTOF) and PSD |
| 6.2     | KAI/I/LE/OS/KP/PG/I/LSR                        | KALEQPGLR                                  | Quadropole orthogonal acceleration time of flight (QqoaTOF) |
| 6.3     | YNGVG/I/(I/L)ASDT/I/UL/(I/L)YSYGI/(I/L)AK      | YNGVGIALDTISLYGLAK                         | Quadropole orthogonal acceleration time of flight (QqoaTOF) |
| 6.4     | EEAIVE/I/LK                                    | EEAIVELK                                   | Quadropole orthogonal acceleration time of flight (QqoaTOF) |
| 7.0     | TTS/I/ULADS/I/UL/L/I/HK                       | TTSSPLADSLTLHK                              | Quadropole orthogonal acceleration time of flight (QqoaTOF) |
| 7.1     | S-Acetyl-ESSYGI/I/L/TTTFS/PG                   | SESSYGLTTFS/PSGR                           | Quadropole orthogonal acceleration time of flight (QqoaTOF) |
| 7.2     | (AT or DG)DIGVL/I/LEAEQK                       | ATDGVLIAEJK                                 | PSD                                      |
| 7.3     | (I/L)(Q/Q)(I/L)LE/YAT/AAASK                   | LVQLEYATTAASK                              | Quadropole orthogonal acceleration time of flight (QqoaTOF) |
| 8.0     | NTL/(I/Q)(Q/I)(A/I)R                           | NTLQLAR                                    | PSD                                      |
| 8.1     | (L/I)ETT/I/LQR                                  | LTTFLQR                                    | PSD                                      |
| 8.2     | M-oX/I/UL/SHATASFM-oX/R                        | MLSTHTASFM                                 | PSD                                      |
| 9.0     | NAFT/(Q/K)AER                                   | NAFTQAER                                   | PSD                                      |
| 9.1     | (I/L)(Q/K)/(Q/K)(Q/K)/M/(I/L)K                 | LQLQ/QMLK                                  | PSD                                      |
| 9.2     | (I/L)(Q/I)(Q/I)ASSR/GM/(Q/K)ADR                 | LFLASSGMQADR                               | Quadropole orthogonal acceleration time of flight (QqoaTOF) |
| 9.3     | IQQYQQYNNGQK                                   | IQQYQQYNNGQK                               | PSD                                      |
| 10.0    | GKS/(I/L)FVEK                                   | GKSILLFVEK                                 | Quadropole orthogonal acceleration time of flight (QqoaTOF) |
| 10.1    | (I/L)STYDA/I/L/DVD/I/UL/VR                     | LSYEDAIDVDIIVR                             | Quadropole orthogonal acceleration time of flight (QqoaTOF) |
| 10.2    | MKV/L/V/K/M                                | KVLV/KFEMK                                 | Quadropole orthogonal acceleration time of flight (QqoaTOF) |
| 10.3    | PS/TS/SAWAAAAA/I/UL/GR                        | PS/TS/SAWAAAAA/UL/GR                      | Quadropole orthogonal acceleration time of flight (QqoaTOF) |
| 10.4    | A/I/UL/LEAVES/GSK                              | ALLEAVES/GSK                               | Quadropole orthogonal acceleration time of flight (QqoaTOF) |
| 10.5    | (I/L)DEH/I/UL/YI/ULASAG/I/UL/SADAR             | LDEHLYLASAGLSADAR                         | Quadropole orthogonal acceleration time of flight (QqoaTOF) and PSD |
| 12.0    | ARNEASEYK                                       | ARNEASEYK                                  | PSD                                      |
| 12.1    | PFHIT/I/L/S/Q/K/                            | PFHITSQK                                   | PSD                                      |
| 12.2    | YGM-oXPM-oXIV/SV/(I/L)AK                      | YGMPDPMVLAK                                | Quadropole orthogonal acceleration time of flight (QqoaTOF) |
| 12.3    | DM-oX/I/QL/VR                                  | DMAQVR                                    | Quadropole orthogonal acceleration time of flight (QqoaTOF) |
| 12.6    | Q/UL/VVS/I/UL/L/HA                          | QLVLSALHK                                  | Quadropole orthogonal acceleration time of flight (QqoaTOF) |
| 12.7    | VSEAI/I/UL/T/I/UL/(Q/K)R                      | VSEAI/I/UL/T/I/UL/(Q/K)R                  | Quadropole orthogonal acceleration time of flight (QqoaTOF) |
| 12.8    | (I/L)(I/L)P/PTTPV/I/UL/R                     | LLPGTPTPVLR                                | Quadropole orthogonal acceleration time of flight (QqoaTOF) and PSD |
| 15.0    | S/I/L/JHATYR                                   | SLHJATYR                                   | Quadropole orthogonal acceleration time of flight (QqoaTOF) and PSD |
| 15.1    | ASSQ/(Q/K)(I/L)S/SS/(Q/K)/TVM-oXK             | ASSQYQLSSQTVMK                             | Quadropole orthogonal acceleration time of flight (QqoaTOF) |
| 17.0    | DGGSSG(I/L)/UL/L/VR                         | DGGSSG/LLR                                 | PSD                                      |
| 17.1    | DGVV/I/LAADSR                                  | DGVV/LAADSR                                | PSD                                      |
| 17.2    | TSTGAYNNVR                                     | TSTGAYNNVR                                 | PSD                                      |
| 17.3    | YAE/I/LAVQNPYSSSAK                             | YAEALAVQNPYSSSAK                           | Quadropole orthogonal acceleration time of flight (QqoaTOF) |
| 17.4    | TTI/L/MLAVSYK                                  | TTLMLAVSYK                                 | Quadropole orthogonal acceleration time of flight (QqoaTOF) |

* The sequences obtained by de novo sequencing contain ambiguous residues represented within parentheses. Modifications to specific residues in these sequences are designated by lowercase abbreviations next to the modified residues: acetylation, acetyl-X; oxidation, X-ox. For database searching, L and Q were used for the ambiguous residue pairs (I/L) and (Q/K), respectively.

b QqoaTOF, quadropole orthogonal acceleration time of flight.
be made using additional information obtained from keyword searches of the MS-Shotgun with catenation results. These assignments are also included in Table III and an accompanying discussion of keyword searching is given under “Spots That Required Additional Biological Information for Provisional Identification.”

Table IV summarizes results from MS-Shotgun with catenation analysis and keyword searches for Spots that could not be identified from sequence analysis alone, Spots 3, 5, 8, 9, and 13. To illustrate some important features of these results, a detailed discussion is also presented below for Spots 1 and 7.

### Simple Search Methods Using Individual Peptide Sequences without Congruence

These are the types of searches generally performed with sequence data obtained by tandem mass spectrometry, specific software that includes MS-Pattern (12) and the CIDentify (37) programs. The BLAST (version 1.4) (21), Gapped-BLAST (version 2.0x) (22), or FASTA (23) programs are also used (46–48). Simple analyses of single sequences of peptide fragments were compared here using MS-Pattern (detailed results not shown) and Gapped BLAST (Table II). MS-Pattern is a simple string-
Subunit designations representing both traditional names and systematic names are given for the assignments of the full-length sequences because these are used synonymously in both the literature and in accession records in sequence databases. Either systematic or traditional names are used in the assignments from MR-Shotgun with catenation analysis and keyword searches in accordance with the names that were used in the annotation lines in the BLAST output files.

| Spot no. | Suggested subunit assignment from MS-Shotgun with catenation analysis | Suggested subunit assignment from keyword searches | Traditional name | Systematic name | p values for BLAST searches using the full-length cloned sequences as the query sequences |
|----------|-------------------------------------------------|-------------------------------------------------|----------------|----------------|--------------------------------------------------|
| 1        | $\alpha_5$                                      | ND$^a$                                           | $\zeta$       | $\alpha_5$     | 1.3e-69                                          |
| 2        | C10-II (7)                                      | C10-III (5), C5                                 | $\beta_3$     | 1.5e-70         |
| 3        | NA                                              | C8                                              | $\alpha_7$    | 4.9e-75         |
| 4        | C2                                              | ND                                              | C2            | 2.2e-141        |
| 5        | NA                                              | C9, C8, C3                                      | C9            | 1.4e-78         |
| 6        | $\beta$(N3)                                     | ND                                              | N3            | 7.1e-48         |
| 7        | C3                                              | ND                                              | C3            | 4.1e-52         |
| 8        | NA                                              | C7-I                                           | C7-I          | 1.0e-42         |
| 9        | C5                                              | C5                                              | C5            | 1.2e-61         |
| 10       | Not available                                    | ND                                              | C6-I          | 6.7e-86         |
| 11       | C7-I                                            | ND                                              | N3            | 5.7e-88         |
| 12       | IOTA(C7a), C8                                   | IOTA(C7a)                                       | IOTA(C7a)     | Z              | 2.9e-92                                          |
| 13       | NA                                              | MECL-1-Lmp10$^b$                                | $X_e$         | $\beta_5$      | 2.3e-150                                          |
| 15       | $\beta$                                         | ND                                              | $\beta_1$    | 8.2e-67         |
| 17       | $\beta$                                         | ND                                              | $\beta_1$    | 8.2e-67         |

$^a$ ND, not done. Keyword searches were either not done or not reported here for these Spots because the data from sequence analysis were sufficient to identify them (determined by a $p$ value for the MS-Shotgun with catenation analysis of 1e-9 or better or from multiple alignments).

$^b$ NA, not applicable. In these cases, no proteasome sequences appeared in the top 10 hits and no other proteasome hits scored sufficiently well to distinguish it from the noise.

Subunit MECL-1 is a close homolog of subunit Z, which can be replaced by MECL-1 subunit upon interferon-γ induction.

Spots 1, represented by three peptide sequences (Table I), provides a good example of this type of information. Sequence 1.0 is highly conserved in all $\zeta (\alpha_5)$-subunit sequences in the non-redundant Static data base, thus providing sufficient information to identify this protein (Table III). Searches using peptide 1.0 revealed proteasome sequences in all of the top 10 hits in both the MS-Pattern and BLAST searches. Statistical significance values, available for only the BLAST searches, suggest that even the best match was likely to occur by chance. As shown in Table II, sequences 1.1 and 1.2 were not nearly so useful for identification of Spot 1.

In a similar manner, the subunit type represented by Spot 4 was also fairly easy to identify due to the presence of a close homolog in the data base, the full-length sequence of the C2 subunit from Trypanosoma cruzi (established later to be 81% identical to the C2 subunit from T. brucei). Four of the seven sequences available for Spot 4 were not useful for identifying the correct function, however. Again, the substantial variability among $p$ values across all of the peptide sequences available for Spot 4 suggests that these scoring functions are not dependable for these short query sequences. Similar issues pertain to the identification of Spots 7, 11, 15, and 17 from these simple searches. For Spots 2, 3, 5, 6, 8, 9, 12, and 13, neither MS-Pattern nor Gapped BLAST produced information sufficient for functional identification. For nearly all of the peptides associated with these Spots, the top 10 hits failed to include any proteasome proteins; for some Spots, many thousands of hits contained no proteasome proteins. BLAST searches using either the BLOSUM62M2 or the PAM30MS matrix also did not achieve statistically significant scores for any of these peptides.

Overall, the MS-Pattern and simple BLAST searches were only marginally useful for identification of most Spots, although in a few cases highly conserved peptide sequences were available that matched well with homologs in the data base. In general, neither the number of available peptide fragments nor their lengths were as important as was a high degree of conservation of a peptide sequence fragment with a homolog in the data base. Although further statistical analysis will be required to explore this issue more systematically, our experience using the limited number of peptides evaluated for the 20 S protea-

### Table IV

| Spot no. of peptides in each spot | No. of proteasome sequences in top 10 hits | Total no. of proteasome sequences found via keyword searches/total no. of sequences in output file$^a$ | Best p value of a proteasome hit/best p value of a non-proteasome hit |
|----------------------------------|----------------------------------------|-------------------------------------------------------------------------------------------------|---------------------------------------------------------------------|
| 3:4                              | 1                                      | 52/444                                                                                           | 0.064/0.036                                                         |
| 5:7                              | 0                                      | 792/3550                                                                                         | 0.0088/0.000007                                                     |
| 8:3                              | 0                                      | 1/987                                                                                             | 1.00/0.61                                                          |
| 9:4                              | 2                                      | 152/2644                                                                                          | 0.039/0.018                                                        |
| 13:3                             | 1                                      | 4/975                                                                                             | 0.99995/0.25                                                       |

$^a$ This total reflects all hits, including those with Shotgun scores = 1 (no congruence).
some dramatically illustrates the difficulties encountered in using these search methods to provide credible identification/prediction of function. By using MS-Pattern searches, for all of the most easily identifiable Spots 1, 4, 7, 11, 15, 17, the total number of hits per Spot ranged from 3 to 32,988. For these results, except for rank ordering based on the number of mismatched residues allowed, there was no way to evaluate whether a sequence listed in the top 10 hits was more likely to be a homolog than a hit further down the list. The poor p values obtained for the great majority of the single sequence fragment searches showed that BLAST is not well suited for identification of remote homologs from these data either.

The degree to which these algorithms gave both unreliable and unpredictable results was surprising given their widespread use for this purpose. The poor performance of these algorithms for the 20 S proteasome raises an important warning for the common use of these simple searching algorithms for identification of function from sequence information typically obtained from mass spectrometry. This problem can be expected to be greatly amplified for high throughput experiments where functional assignments from data base searches will depend on automated evaluation without recourse to manual review by scientists with expertise in each system in question.

Investigation of Spots Using Congruence Analysis

Searches using automated congruence analysis were performed using our program, MS-Shotgun, with and without catenation of peptide sequences.

Overall, simple congruence analysis with Shotgun across the outputs of either MS-Pattern or Gapped BLAST searches (Fig. 1B, left side) using all peptide sequences in independent searches gave no substantial improvement over BLAST searching without Shotgun analysis (data not shown), e.g. Spots that could not be identified using simple searching with MS-Pattern or BLAST were not identifiable using simple congruence analysis either. This was perhaps due to the inherent limitations of data base search programs to find true positive hits from such short peptide sequences, see below.

In comparison to searches performed for each peptide independently, and to simple congruence analysis without catenation, MS-Shotgun analysis with catenation resulted in sometimes greatly amplified scores for the correct hits for Spots 1, 2, 4, 6, 7, 11, 12, 15, and 17 (Table II). The implementation of the FASTS program used in this study showed a similar level of improvement in distinguishing true homologs (data not shown). Detailed discussion of the results for Spots 1 and 7 is provided below.

Improvement in the Ability to Identify Spot 1—For Spot 1, the additional information available from MS-Shotgun with catenation analysis produced scores with improved statistical significance (p values) relative to BLAST searches on single peptide sequences (Table II). Because it could take advantage of the information from all three of the peptide sequences available for Spot 1 in consonance, MS-Shotgun with catenation analysis resulted in an enriched clustering of proteasome sequences within the top 10 hits. The catenation analysis was also useful to distinguish the order of the peptides. Considering all hits, the best p value for a proteasome hit is 5 orders of magnitude better than the best p value for a non-proteasome hit using the BLOSUM62MS matrix and two orders of magnitude better using the PAM30MS matrix. For 9 of the top 10 hits, the most statistically significant score among all six of the possible permutations was that obtained by the catenation representing the correct sort order, 021 (data not shown).

To evaluate how similar the α5-subunit of T. brucei, as represented by Spot 1, is to α5-subunit sequences available from other species, we mapped the 3 sequences from Spot 1 onto a multiple alignment of the best hits found from MS-Shotgun with catenation (Fig. 3). This confirmed the correctness of the linear ordering of the peptides, 021, predicted by the analysis. From this alignment it is apparent that no one homolog in the data base is clearly “closest” to the T. brucei sequence; thus, inclusion of several sequences in a multiple alignment gives a more useful picture of the relationship between the α5-subunit from T. brucei to those from other species than would a pairwise comparison to any one of these homologous sequences. From pairwise comparisons, such as reported by FASTS, it would be difficult to see how either peptide 1.1 or 1.2 from Spot 1 relates to any one of these α5-subunit sequences. It is also clear that peptide 1.0, GVNTFGPPEGR, fortuitously represents a highly conserved motif among known α5-subunits and distinguishes that subunit from other related subunits. To illustrate this point, the sequence for a related subunit, the C2 subunit from Oryza sativa, is included in Fig. 3.

Improvement in the Ability to Identify Spot 7—For Spot 7, 81 of the top scoring 87 hits represent proteasome sequences (data not shown) with the p values for the highest scoring proteasome hit improving from statistically insignificant for all of the BLAST searches on single peptide sequences to 2.5e−11 (Table II) for the catenated analysis of the correctly ordered permutation, peptide order 1320. In contrast, the best p value of a non-proteasome hit is 6 orders of magnitude larger. In Fig. 4A, the four peptide sequences from Spot 7 are mapped onto three of the best scoring hits in the BLAST search output for the correctly ordered permutation, catenation order 1320. Despite substantial differences in the sequences of these three related hit sequences, the figure shows that the catenated peptide sequence 1320 aligns quite well with each.

Because only a limited number of preset gap penalties are available for use with Gapped BLAST, it is currently difficult to determine the advantages of gapped over ungapped alignments with this type of data. As with many other aspects of these analyses, comparisons of gapped and ungapped alignment outputs for different Spots show great variability in the statistical significance of the scores. From a preliminary examination, it appears that the usefulness of gapped alignments (generated using Gapped BLAST, e.g. BLAST 2.0x versions) over ungapped alignments (generated using BLAST version 1.4) is substantial for some proteins but can even result in less significant scores for others (data not shown). The reasons for this apparently anomalous behavior again appear to be complex, showing large variation with other parameters such as scoring matrix. One likely reason for the uneven performance of the gapping function with these data is that gap penalties are assessed both in terms of a penalty for opening a gap and a penalty for extension of a gap. Thus, when the gaps reflecting missing information in our catenations are long, the gap penalty becomes substantial, adversely affecting the score. When gaps reflecting missing information are short, gap penalties are low, allowing a scoring advantage for the gapped alignment. Since there is no way to predict the length of gaps due to missing information from the peptide sequence information itself, it is impossible to set appropriate gap penalties for any given set of catenations. Moreover, the current gap length limit

5 An updated implementation of FASTS that became available after our initial analyses were completed has apparently been optimized to identify sequences that are already present in the data bases more effectively. For us, this has resulted in poorer performance with the 20 S proteasome data, for which only homologous sequences were present in the data bases. The version of FASTS used for this study is no longer available.
for scoring a gapped alignment is 40 amino acids. As peptide sequences from our data frequently were separated by gaps of >40 amino acids due to missing information, it is not possible to evaluate the performance of the gapping function for these catenations. A solution to this problem will require modification of the search algorithm to treat unnatural gaps due to catenations. A solution to this problem will require modification of the search algorithm to treat unnatural gaps due to evolutionary divergence, which the algorithm was designed to accommodate.

Despite the complexities associated with this issue, the use of Gapped BLAST appears to provide advantages for data base searching with the "unnatural" sequences represented by our catenations. For the best scoring catenation of Spot 7, as illustrated in Fig. 4A, all four peptides were used in the alignments, with gaps between peptides 3 and 2 inserted by the Gapped BLAST algorithm in order to obtain the best alignment to each full-length homolog sequence. Note, however, that the gap between peptides 3 and 2 is relatively short.

Fig. 4B illustrates the complex and variable nature of these data with regard to which and how much of the available sequence fragments contribute to the overall scores. In this figure, the four peptide sequences from Spot 7 are shown as ordered in the three best scoring catenation orders for a single high scoring hit. Note that the alignment for sort order 1320 uses all four sequence fragments in the alignment with the hit sequence, whereas the contribution of sequence fragment 0 to the alignment in sort order 1302 is minimal, making the effective sort order for this search 1320, consistent with the best scoring sort order of 1320. The alignment for sort order 0132 does not even include sequence fragment 0 in the alignment, again making the effective sort order for this search 132. The poor contribution provided by sequence fragment 0 to these alignments is also consistent with the results for simple BLAST searches (Table II), which shows that no hits representing proteasome sequences were obtained for sequence 7.0 using simple BLAST searches.

Another important way in which MS-Shotgun with catenation analysis improves the ability to identify function from sequence analysis is its ability to separate correct from incorrect scoring orders. This advantage was substantial for Spots 4, 6, 7, 11, and 17. The difference in p values obtained for correct and incorrect sort orders is illustrated graphically for Spot 7 in Fig. 5 which shows that the score for the catenation representing the correct sort order, 1320, and variants of the effective sort order, 132, have the best p values and stand out well from the background scores representing other catenations.

**Spots That Required Additional Biological Information for Provisional Identification**—For the unknown proteins in Spots 3, 5, 8, 9, and 13, identification of the best candidate homologs was not possible using sequence information alone, even with the aid of MS-Shotgun with catenation. Table IV shows that few proteasome hits appeared among the top 10 hits for any of these Spots using MS-Shotgun with catenation. The p values for statistical significance are also unimpressive and in all cases worse than the values for non-proteasome hits. These results are not surprising and help to remind us that function cannot always be assigned from sequence information. This is especially true for proteomics studies in which only a small amount of sequence information, representing varied and unknown coverage of the parent protein, is available.

Despite the failure of these methods to produce any high scoring proteasome hits for these Spots, congruence analysis gave some information that in conjunction with biological information allowed provisional assignments of function. Because this experiment was designed to isolate only proteins from the 20 S proteasome complex, the appearance of known proteasome proteins in our search outputs helped to assign a subunit type for each Spot. This was implemented through
Fig. 4. A, a portion of the Gapped BLAST output for three top scoring hits from the MS-Shotgun with catenation search using the catenation order 1320. The 1st lines for each hit show the sequence identifier information, BLAST scores, and statistics for that hit as it appeared in the alignment portion of the BLAST output. The 1320 sequence catenation (designated as the "Query" sequence) is shown in color aligned with the regions of each top scoring hit (designated as the Sbjct sequence) that matched it, as determined by the BLAST algorithm. The unnumbered line between the Query and Sbjct sequences shows the matched regions that were used to generate the scores for each hit. Exact matches are shown as capital letters and conservative substitutions are shown using +. Hyphens designate the gaps that were inserted in the query sequence, relative to the subject sequence, to obtain an optimal alignment score. In order to show how each individual sequence fragment in catenation 1320 aligns along each hit, the sequence fragments 1, 3, 2, 0 are distinguished by different colors keyed to coloring of the peptide order "1320." B, a portion of the Gapped BLAST outputs for one high scoring hit from the three best scoring MS-Shotgun with catenation searches for Spot 7. This hit scored well in the three searches representing the top scoring catenation orders. Elements of the alignments are as described for A except that the sequence identifier information is not provided. Shown are the Gapped BLAST alignments for the three different database searches using the catenation orders 1320, 1302, and 0132 as query sequences. The effective catenation orders for each search, as described in the text, are shaded in yellow. Sequence fragments and their corresponding peptide fragment designation in the sort orders 1320, 1302, or 132 are colored the same.
automated keyword searches of the database search outputs using the keywords “proteasome,” “multicatalytic,” “endopeptidase,” and “macropain.” Table III gives tentative assignments of subunit identity for Spots 3, 5, 8, 9, and 13, based on the best scoring hits from the MS-Shotgun with catenation analysis and/or the keyword searches. Although these assignments all include the correct subunit identification among the likely solutions obtained from this strategy, it should be emphasized that the partial sequence information available for these Spots was insufficient for credible and definitive assignment of their functions. The tentative nature of such assignments is emphatically illustrated in Table IV by the small number of proteasome sequences found from keyword searches compared with the total number of hits in the search outputs and the fact that hits representing non-proteasome sequences scored with better statistical significance than proteasome sequences. Although keyword searching is easily automatable, the time-intensive individual analysis required to evaluate the data for these difficult to identify Spots would preclude the use of this approach in a high throughput setting.

A Note on the Importance of the Scoring Matrix in Database Searches Using Peptide Sequences Obtained from Tandem Mass Spectrometry

Information theory analysis (24) suggests that the PAM30 matrix would be more effective for generating statistically significant scores than the BLOSUM62 matrix for the searches using the short peptide fragments typified by our data. However, except for those few peptides in our data set that are highly similar to a region of a homologous full-length protein in the data bases, we found only slight improvement using either the PAM30 or the PAM30MS matrix compared with their BLOSUM62 counterparts (data shown in the Tables only for the -MS versions of these matrices). As shown in Table II, use of the PAM30MS matrix for the searches using individual peptide sequences resulted in 3–5 orders of magnitude improvement in p values over the results obtained from use of the BLOSUM62MS matrix for peptides 1.0, 4.4, 4.5, 4.6, and 15.1. These peptides are the highest scoring peptides from all database searches of this type that we performed and were among the few peptides containing sufficient information for identification of the proteasome subunit from a simple search of a single peptide. For the more difficult searches, neither PAM30 nor PAM30MS performed sufficiently better than BLOSUM62MS to allow identification of those proteins from simple database searches.

One reason that use of a scoring matrix predicted by theory to be optimal with these data was not more advantageous likely lies, again, in the manner in which the peptide sequence fragments represent coverage of the parent protein. Altschul (24) has predicted the ranges of local alignment lengths for which specific PAM matrices are appropriate. For example, the minimum sequence length required to distinguish a statistically significant alignment from one likely to occur by chance is 12 amino acids using the PAM30 matrix and 31 amino acids using a PAM120 matrix. Although the BLOSUM series matrices cannot be cleanly correlated with the PAM matrices with respect to these predictions, it has been suggested that the BLOSUM series does not include any matrices suitable for the shortest queries (see the BLAST tutorial provided by National Center for Biotechnology Information). But the short sequences available from tandem mass spectrometry correspond not to local alignment regions found by algorithms such as BLAST but rather to stretches of sequence that may only fortuitously and occasionally correspond to regions that would be matched in a BLAST alignment of two full-length sequences. This means that matches between our query sequences and even close homologs in the databases will frequently score with p values that are below the level of statistical significance whether PAM30 or BLOSUM62 is used.

An example is provided by the sequence fragments available for Spot 1, as discussed above and shown in the multiple alignment given in Fig. 3. One of these sequences, peptide 1.0, matches perfectly or nearly perfectly a highly conserved region in all of the homologs shown in Fig. 3. Thus, this peptide, which fortuitously corresponds to a high scoring alignment region for
Using PAM30MS, it is 4.4e matrix than the BLOSUM62 matrix, which is indeed the case. BLAST, would be expected to score better with the PAM30 matrix than the BLOSUM62 matrix. Moreover, in some cases, the results using PAM30MS were worse than the results using BLOSUM62MS. For Spot 2.4, for example, a simple BLAST search provided results that, while not statistically significant, at least represented proteasome sequences as 6 of the top 10 hits using BLOSUM62MS. In contrast, the search using the PAM30MS matrix contained only two proteasome sequences in the top 10 hits, with neither hit showing an improvement in p value relative to the BLOSUM62MS matrix.

These results and those provided in Table II for other Spots show that neither the use of BLOSUM62MS nor of PAM30MS provides statistically significant scores for many of the peptides available for the T. brucei proteasome. Thus, our results reflect the need for more sophisticated solutions than those provided by standard approaches to data base searching which were developed, after all, for distinctly different purposes. Moreover, the variability in performance of both of these matrices with these data leads to great uncertainty in the identification of these proteins and emphasizes the importance of using the longest possible sequence lengths for data base searching. By using our protocol, this can be achieved by catenation of all available sequence fragments from each protein to provide the longest possible query sequence. As described above, using all possible permutations of those catenations as queries in data base searches provides a distribution of scores from which the correctly ordered catenation may, in some cases, be distinguished. But although the improvement provided by this approach over that performed by searching with single sequence fragments is substantial, it is not yet a mature solution. Based on all of the results described in this study, several areas in which future research may prove fruitful are described below.

**Future Directions**

As the volume of data from high throughput proteome analysis increases, automated assignment of function from sequence information inferred from mass spectrometry will be required. Our results suggest that current informatics approaches do not provide an adequate solution. Although the new approaches described in this work enhanced our ability to assign subunit identities to the T. brucei proteasome, the complexities associated with analysis of this type of data suggest that substantial further development of the bioinformatic approaches is required. In particular, the variability in how well any set of peptide sequences from a given protein represents conserved elements of a homolog in the data base confounds a systematic and rigorous solution to the problem. Our experience with these data does, however, suggest several directions in which improvement may be achieved. Thus, whereas the nature of the available data may obviate an elegant solution, practical improvements may be sufficient to allow us to find more effectively the relevant relationships that will help us establish identity or obtain clues about function of unknown proteins. We envision that a better tuning of our congruence approach might be achieved in several ways as follows.

First, alter the data base search algorithm to treat the missing information in our catenated sequences, e.g., regions of the associated full-length sequence for which no sequence information was obtained, differently from the normal types of gaps expected in an alignment of a query sequence and a putative data base homolog. The Gapped BLAST algorithm was designed to handle the latter types of gaps but not the former. Implementation of such a strategy could also remove the problems associated with the current allowed maximum gap length of 40 amino acids. We are exploring ways in which such changes could be implemented with respect to both the availability of BLAST source code and to the consequences of these changes for the number of false positives that might be obtained at statistically significant scores.

Second, adapt search engines to accommodate better the idiosyncratic nature of sequence information obtained from mass spectrometry. These could include uncertainty in the ordering of the first two residues of a sequence and the inability to distinguish Ile/Leu and Gln/Lys, although high energy CID can eliminate the ambiguity in distinguishing Ile and Leu (31, 32, 19) and accurate measurement of relevant sequential ion series can distinguish Gln and Lys (12). We have noted that the use of either the BLOSUM62MS matrix or the PAM30MS matrix, which score Ile/Leu and Gln/Lys equivalently, substantially improves our ability to identify remote homologs. Recent implementation of congruence/catenation methods in the string searching algorithm MS-Pattern, initially designed to accommodate the ambiguities inherent in sequence data obtained from Edman degradation or PSD mass spectrometry, also show great promise for improved identification of remote homologs.

Other accommodations for the deficiencies observed in CID spectral sequence fragment content, such as the common lack of observation of b1 ions, could improve data base search results. Some such accommodations, mentioned previously, were incorporated into the original design of the MS-Pattern algorithm. The types and sources of possible subsequence ambiguities, e.g., residue orders or isobaric substitutions, as well as the improvements in quality and completeness of de novo sequence being obtained from new generation tandem instrumentation will be discussed elsewhere. By using the information and the models provided by MS-Pattern and CIDentify, it should be possible to implement efficient searching protocols that allow submission of each peptide sequence in a format that accommodates the likely alternate versions.

Third, systematically investigate other scoring matrices and other scoring parameters to determine a generally more optimal set of default parameters useful with sequence information obtained from tandem mass spectrometry. Although our initial comparisons of the BLOSUM62MS and PAM30MS suggest that the problem is more complicated than that for full-length “natural” sequences, it is still likely that improvement could be obtained, perhaps through use of initial parameterizing runs followed by additional searches using those optimized parameters. More useful parameterization may also require knowing whether and how the lengths of catenated peptide sequences can be correlated to specific scoring matrices now in use.

Fourth, develop an efficient strategy for minimizing the number of catenation searches that must be performed for large numbers of peptide sequences. The number of catenations for all possible permutations of n sequences from a given Spot is n! For proteins for which several peptide sequences are available, MS-Shotgun with catenation analysis requires a large number of searches, e.g., for n = 6, 720 searches are required. For high throughput analysis of perhaps 10s of thou-

---

6 P. R. Baker, R. J. Jacobs, L. Huang, K. F. Medzhiradzsky, M. A. Baldwin, and A. L. Burlingame, manuscript in preparation.
The near future needs for high throughput mass spectrometric analysis of protein spots were described by M. L. Vestal at the meeting of the American Society for Mass Spectrometry, September 1999, Atiloma, CA, and have also been observed from recent results using MALDI-TOF TOF derived from CID spectra (L. Huang, M. A. Baldwin, K. E. Williams, C. B. Silbohm, K. F. Medzhirdashky, P. R. Baker, D. A. Malthby, M. Fallavici, J. Campbell, P. Juhasz, S. A. Martin, M. L. Vestal, and A. L. Burlingame, manuscript in preparation).

7 The near future needs for high throughput mass spectrometric analysis of protein spots were described by M. L. Vestal at the meeting of the American Society for Mass Spectrometry, September 1999, Atiloma, CA, and have also been observed from recent results using MALDI-TOF TOF derived from CID spectra (L. Huang, M. A. Baldwin, K. E. Williams, C. B. Silbohm, K. F. Medzhirdashky, P. R. Baker, D. A. Malthby, M. Fallavici, J. Campbell, P. Juhasz, S. A. Martin, M. L. Vestal, and A. L. Burlingame, manuscript in preparation).
Functional Assignment of the 20 S Proteasome from *Trypanosoma brucei* Using Mass Spectrometry and New Bioinformatics Approaches

Lan Huang, Richard J. Jacob, Scott C.-H. Pegg, Michael A. Baldwin, Ching C. Wang, Alma L. Burlingame and Patricia C. Babbitt

*J. Biol. Chem. 2001, 276:28327-28339.*

doi: 10.1074/jbc.M008342200 originally published online April 17, 2001

Access the most updated version of this article at doi: 10.1074/jbc.M008342200

Alerts:
- When this article is cited
- When a correction for this article is posted

Click here to choose from all of JBC's e-mail alerts

This article cites 43 references, 10 of which can be accessed free at [http://www.jbc.org/content/276/30/28327.full.html#ref-list-1](http://www.jbc.org/content/276/30/28327.full.html#ref-list-1)