Identification of Unknown Proteins Through Mining Distant Protein Homology With Mass Spectral Information

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

The explosion of public gene and EST databases for human and other species has accelerated the identification of large numbers of proteins. Hence, the matching and retrieval of proteins extant in these databases using mass fingerprints and partial peptide sequence information is now routine. However, current strategies for database interrogation using mass spectral sequence information are limited in their ability to detect sequence homology permitting identification of novel proteins as well as gaining clues to their function. Therefore, we have developed new bioinformatics tools to identify unknown proteins through effective detection and identification of their remote homologs in databases.

We have first investigated the 20 S proteasome macromolecular complex from *Trypanosoma brucei*. The protein complex was separated by 2-D gel electrophoresis into 15 major and several minor spots using previously described methods (1). In order to identify all the proteasome subunits from *T. brucei*, 15 major spots were excised from the gel and digested. The list of experimental tryptic peptide masses obtained by MALDI from each protein digest was first submitted for database searching using MS-Fit program (Protein Prospector). Since cDNA sequences of 20S proteasome subunits from *T. Brucei* were not present in any of the publicly databases available at the time this experiment was performed, mass mapping was limited to searches for homologous proteins with conserved tryptic peptides. Not surprisingly, no proteasome entries from other species were found. Therefore, *de novo* sequencing was performed to obtain 2-8 sequence fragments ranging from 5-18 residues per sequence. These sequence fragments were compared to the proteins in a static version of the Genbank non-redundant protein database downloaded on 3/11/99 using the MS-Pattern string-searching program from the ProteinProspector software package (2) or the MS-Shotgun program developed for this work, which was based on the

Shotgun program(3). The results have shown MS-Shotgun to be the most effective of the methods evaluated for identifying distant homologs of the 20S proteasome from *T. Brucei*. MS-Pattern analysis was unable to identify proteasome components in any fragments for 8 of the 15 spots except as buried in the noise of the output. When MS-Pattern searches were analyzed for congruence in a procedure analogous to that used in MS-Shotgun, 4 of these spots gave promising results. However, congruence analysis of Gapped Blast results using MS-Shotgun provided sufficient information for identification of 12 of 15 spots and additional information potentially useful for the identification of an additional spot. Gapped Blast of single sequence fragments alone proved superior to MS-Pattern as well but was unable to identify proteasome or other convincing homologs for several of the spots solved by congruence analysis (MS-Shotgun) (data not shown). The protein identity obtained from MS-Shotgun analysis for each spot was evaluated from the appropriate multiple alignment context. For each spot, the query sequence fragments were mapped onto multiple alignments of the top-scoring MS-Shotgun hits, which was used to evaluate the protein identity. Figure 2 shows an example alignment for Spot 7, using a subset of the hits with MS-Shotgun scores of 3(Figure 1). The associated sequences were labeled as spot7a0-3 in Figure2. Therefore spot7 was identified as proteasome C3 subunit. As a result, 14 major spots were identified as the 14 essential subunits of 20S proteasome from *T. brucei*, the same as in eukyrotic cells. While none of these functional assignments have been

unequivocally verified by experimental characterization of the proteins, they have been provisionally confirmed by independent methods.

References:
1. Huang, L. Shen, M., Chernushevich, I.V., Burlingame, A.L., C.C.Wang, Robertson, C.D. *Molecular and Biochemical Parasitology*, 1999, 102, 211-223.
2. K. Clauser, P. Baker and A. L. Burlingame, *Anal Chem*, 1999, 71, 2871-2882.
http://prospector.ucsf.edu.
3.Pegg, S. C.-H. and P. C. Babbitt. *Bioinformatics*. 1999, **15**(9): 729-740.
Supported by NCRR P41 RR01614 and NCRR S10 RR12961.

```
gnl|PID|e1334470 - (AL031966) 20s proteasome component ...          Score: 3
------Query----------Score-----Probability-----Segments------------
spot7a2.save            52         0.997           1
spot7a1.save            44      1.000000           1
spot7a3.save            37      1.000000           1
sp|P25787|PRC3_HUMAN - PROTEASOME COMPONENT C3 (MACROPAIN S...      Score: 3
------Query----------Score-----Probability-----Segments------------
spot7a1.save            50         0.999           1
spot7a2.save            42      1.000000           1
spot7a3.save            37      1.000000           1
sp|P17220|PRC3_RAT - PROTEASOME COMPONENT C3 (MACROPAIN S...        Score: 3
------Query----------Score-----Probability-----Segments------------
spot7a1.save            50         0.999           1
spot7a2.save            42      1.000000           1
spot7a3.save            37      1.000000           1
sp|P49722|PRC3_MOUSE - PROTEASOME COMPONENT C3 (MACROPAIN S...      Score: 3
------Query----------Score-----Probability-----Segments------------
spot7a1.save            50         0.999           1
spot7a2.save            42      1.000000           1
spot7a3.save            37      1.000000           1
```

Figure 1. The top four hits from MS-Shotgun analysis of the 4 sequence fragments obtained for spot 7.

```
 P23639  ~MTDRYSFSL TTFSPSGKLG QIDYALTAVK QGVTSLGIKA TNGVVIATEK
spot7a2  ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~A TDGVVLAAEQ
 P25787  ~AERGYSFSL TTFSPSGKLV QIEYALAAVA GGAPSVGIKA ANGVVLATEK
 P17220  ~AERGYSFSL TTFSPSGKLV QIEYALAAVA GGAPSVGIKA ANGVVLATEK
 P49722  ~AKRGYSFSL TTFSPSGKLV QIEYALAAVA GGAPSVGIKA ANGVVLATEK
 P24495  ~AERGYSFSL TTFSPSGKLV QIEYALAAVA AGAPSVGIKA TNGVVLATEK
 P40301  MATERYSFSL TTFSPSGKLV QLEYALAAVS GGAPSVGIIA SNGVVIATEN
spot7a1  ~~~SESSYGL TTFSPSGR~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
spot7a3  ~~~~~~~~~~ ~~~~~~~~LV QLEYATTAAS K~~~~~~~~~ ~~~~~~~~~~
spot7a0  ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~

 P23639  KSSSPLAMSE TLSKVSLLTP DIGAVYSGMG PDYRVLVDKS RKVAHTSYKR
spot7a2  K~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
 P25787  KQKSILYDER SVHKVEPITK HIGLVYSGMG PDYRVLVHRA RKLAQQYYL.
 P17220  KQKSILYDER SVHKVEPITK HIGLVYSGMG PDYRVLVHRA RKLAQQYYL.
 P49722  KQKSILYDER SVHKVEPITK HIGLVYSGMG PDYRVLVHRA RKLAQQYYL.
 P24495  KQKSILYDEQ SAHKVEPITK HIGMVYSGMG PDYRVLVRRA RKLAQQYYL.
 P40301  KHKSPLYEQH SVHRVEMIYN HIGMVYSGMG PDYRLLVKQA RKIAQTYYL.
spot7a1  ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
spot7a3  ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
spot7a0  ~TTSPLADSL TLHK~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~ ~~~~~~~~~~
```

Figure 2. Multiple alignment of novel sequences from spot7a0-3 onto the"best" homologs found by MS-shotgun as in Figure 1.