Matching peptide mass spectra to EST and genomic DNA databases

Jyoti S. Choudhary, Walter P. Blackstock, David M. Creasy and John S. Cottrell

The use of mass spectrometry data to search molecular sequence databases is a well-established method for protein identification. The technique can be extended to searching raw genomic sequences, providing experimental confirmation or correction of predicted coding sequences, and has the potential to identify novel genes and elucidate splicing patterns.

The first draft assembly of the human genome was announced on 26 June 2000 (Refs 1 and 2). This is the first vertebrate genome to have been sequenced and, at 3.2 Gbp, it is also the largest. As of 7 June 2001, more than 50% of the sequence remained in draft form. Nevertheless, the assembly provides a comprehensive view of the genome and is an extraordinary resource.

Intriguingly, the estimated 28 000–40 000 human genes are encoded by only 2–5% of the genome. The broad range of this estimate reflects the limitations of current gene-prediction methods in annotating genes from eukaryotes. Gene-discovery algorithms based on statistical and homology methods facilitate the identification of known genes, but are inadequate for locating novel genes[3,4]. The application of alternative methods, particularly the use of expressed sequence tags (ESTs), has proved to be valuable in verifying and defining coding regions[5]. However, this approach also has its drawbacks, as EST representation tends to be biased by expression level, tissue, and cell type. In this review, we consider the use of mass spectrometry (MS) data in direct database searching of the human genome, and also discuss its application in gene mining.

Methodology

Protein identification by searching MS data against a transformed database of molecular sequences is a core technology in proteomics[6,7]. One approach, peptide-mass fingerprinting, compares a set of measured peptide molecular-mass values from a proteolytic digest against values calculated by *in silico* digestion of sequences from a protein database. Discrimination depends on the specificity of the protease and the constraint that the mass values originate from a defined protein sequence. With limited exceptions, this method cannot be applied to short stretches of sequence, such as ESTs, or long, continuous sequences, such as genomic DNA. In such cases (or when the sample is a protein mixture), the preferred approach is to identify discrete peptides using MS–MS data. The sequence tag method of Mann and Wilm depends on prior interpretation of the MS–MS spectrum to obtain a short stretch of amino acid sequence[8,9]. Alternatively, uninterpreted MS–MS data can be searched directly[10–12].

In addition to protein identification, searching MS–MS data against raw, unmasked genomic DNA can provide primary experimental verification and correction of predicted coding sequences, together with the possibility of identifying novel genes and elucidating splicing patterns. In a eukaryotic genome, where coding sequences are divided into exons and introns, a gene may extend over 100 kbp or more. The spatial distribution of peptide matches will rapidly locate gene-containing regions and, in most cases, each cluster of matches can be expected to correspond to a single gene. Even the shortest exon can be identified, as long as it spans at least one good peptide match. However, matching a peptide that spans the splice between two exons is difficult (Fig. 1). In general, such matches will be missed when searching raw genomic DNA, and can only be found by an iterative process once the general location of the gene has been identified.

Jyoti S. Choudhary and Walter P. Blackstock
Cell Mapping Project,
GlaxoSmithKline R&D,
Stevenage, UK SG1 2NY.

David M. Creasy and John S. Cottrell*
Matrix Science,
30 Harcourt Street, London,
UK W1H 4HT.
*e-mail:
jcottrell@matrixscience.com

0167-7799/01/$ – see front matter ©2001 Elsevier Science Ltd. All rights reserved. PII: S0167-7799(01)01795-4

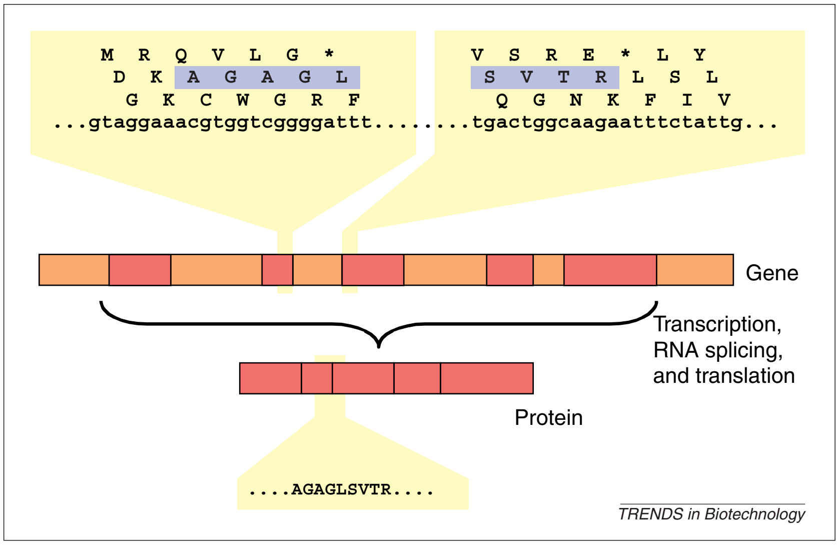


Figure 1. Peptides that span exon splices will be missed when matching uninterpreted MS–MS data to genomic DNA

In most eukaryotes, genes are divided into protein-coding exons (red) and non-coding introns (orange). After transcription to RNA, the coding sequences are spliced together before translation. Hence, when the mature protein is digested for analysis by mass spectrometry, a proportion of the peptides will correspond to coding sequences that span exon splice sites.

Searching ESTs

Searching uninterpreted MS–MS data against nucleic acid sequences was first demonstrated by Yates and colleagues[13]. In general, the reading frame for translation is unknown, and the nucleic acid sequence must be translated in all six frames before searching. This generates a large quantity of effectively random sequence, within which some degree of matching to the experimental data will occur by chance. Yates and co-workers were able to show that representative MS–MS data contained sufficient information to discriminate a positive match from the background of random matches. Of course, in 1995, the databases were considerably smaller than today; the EST database contained just 65 112 entries.

There are now many examples in the literature of searching EST databases. For example, Neubauer and colleagues used this approach to characterize the human multi-protein spliceosome complex[14]. For several important organisms, EST databases offer more comprehensive sequence coverage than any protein database. The error rate might be high, but extensive redundancy means that the chance of finding the correct sequence for the correct variant is good. A disadvantage is that entry descriptions can be unhelpful for protein recognition, so that it is often necessary to BLAST the EST sequence against a protein database to comprehend what one has found.

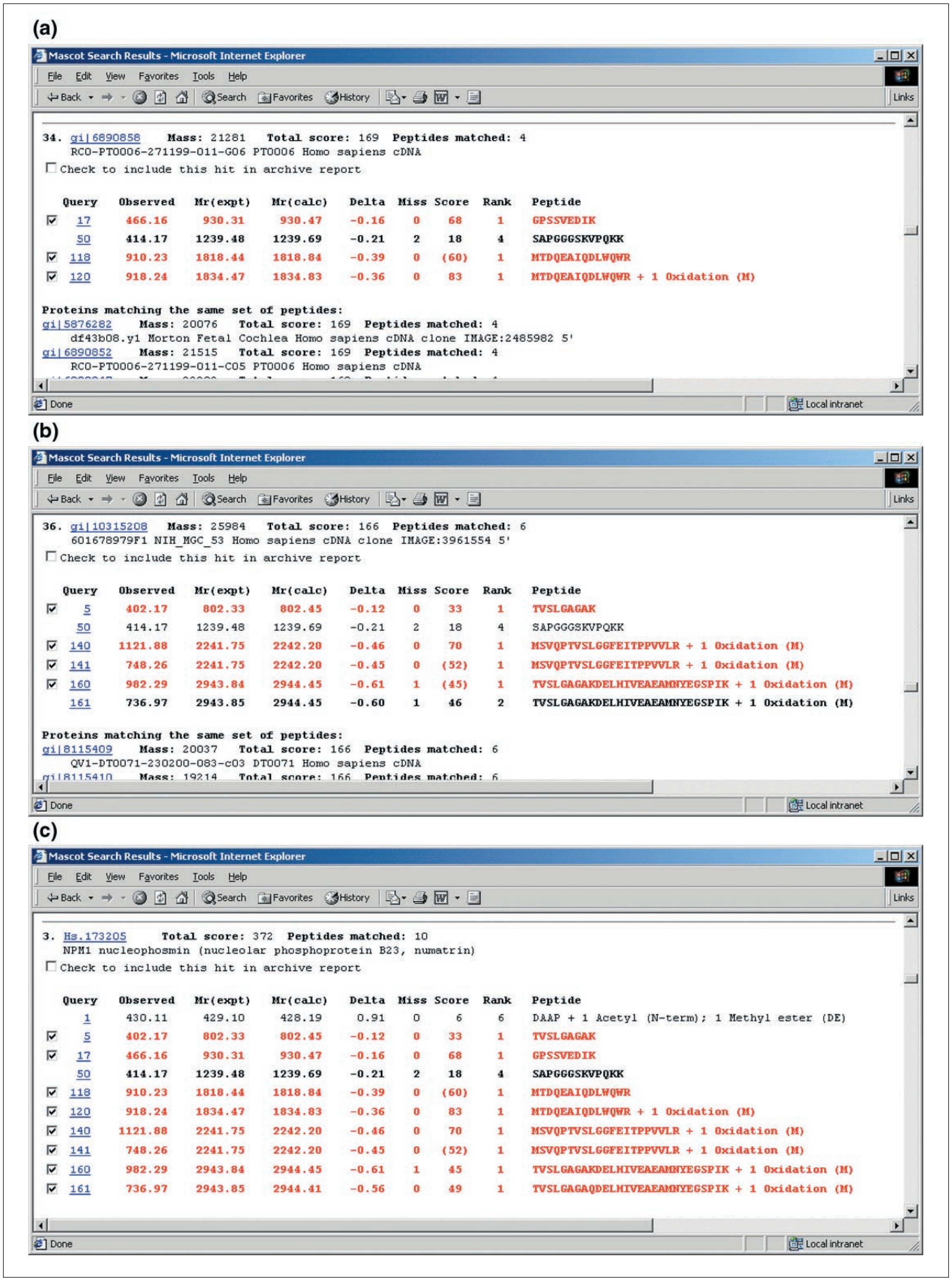
A more fundamental limitation relates to the grouping of peptide matches into protein matches. A liquid chromatography (LC–MS–MS experiment, in which the digestion products of a protein mixture are analysed, could produce hundreds or even thousands of MS–MS spectra. The search process attempts to match these spectra to peptide sequences. Following the search, a report will then try to group the matched peptides into proteins. These assignments might sometimes be arbitrary and ambiguous, but they serve two important functions. First, grouping makes the search results easier to comprehend, and second, grouping can influence judgement as to whether a weak match is correct or not. If ten strong peptide matches are found in one particular protein, there will be greater confidence that an eleventh, weaker match to the same protein is correct, than if the weaker match was an isolated one.

As ESTs mostly correspond to protein fragments, extended groupings are rare. This can be rectified by using an index, such as UniGene, to group EST matches into gene families. UniGene (http://www.ncbi.nlm.nih.gov/UniGene) is an index created by automatically partitioning GenBank sequences using BLAST, to produce a non-redundant set of gene-oriented clusters. In an example taken from Choudhary *et al.*[15], one grouping from a search of dbEST (an EST database) using the Mascot search engine[12] contained three strong peptide matches and one weak match [Fig. 2(a)]. Another grouping contained five strong peptide matches and one weak match [Fig. 2(b)]. Casual inspection of the report might easily fail to see a connection between these two hits, in which the only common peptide was the weak match to spectrum 50. However, when the ESTs were grouped by using UniGene to translate dbEST accession strings into gene identifiers, both sets of matches were found to correspond to the same entry – Nucleophosmin [Fig. 2(c)]. Therefore the net effect of UniGene grouping was to simplify the dbEST report, and make it equal in clarity to the report from a protein database search.

Searching genomes

There are relatively few reports in the literature of searching raw genomic DNA. Most are microbial studies, such as *Porphyromonas gingivalis* (2.2 Mbp) (Ref. 16), *Haemophilus influenzae* (1.8 Mbp) (Ref. 17), and *Mycoplasma pneumonia* (0.8 Mbp) (Ref. 18). A more common approach is to search compilations of nucleic acid sequences corresponding to open reading frames[19].

Raw, unmasked eukaryotic genome sequences present a particular challenge because of both their size and the arrangement of the coding sequence into exons and introns. Küster and colleagues used a variation on the sequence tag approach to search a database containing some 95 Mbp of genomic sequence data from *Arabidopsis thaliana*, believed to represent approximately 75% of the complete genome[20]. Interpreted peptide sequence tags were translated into sets of degenerate oligonucleotide sequences for searching against the database. Peptides identified in this way were used to refine exon predictions from a variety of gene-finding programs. Having located a gene using one or more sequence tags, high


(a)
Mascot Search Results - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History Links
Done
Local intranet
(b)
Mascot Search Results - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History Links
Done
Local intranet
(c)
Mascot Search Results - Microsoft Internet Explorer
File Edit View Favorites Tools Help
Back Search Favorites History Links
Done
Local intranet

(a)

```
34. gi|6890858     Mass: 21281   Total score: 169  Peptides matched: 4
    RC0-PT0006-271199-011-G06 PT0006 Homo sapiens cDNA
□ Check to include this hit in archive report

   Query  Observed  Mr(expt)  Mr(calc)  Delta  Miss Score  Rank  Peptide
☑   17    466.16    930.31    930.47   -0.16   0     68    1    GPSSVEDIK
    50    414.17   1239.48   1239.69   -0.21   2     18    4    SAPGGGSKVPQKK
☑  118    910.23   1818.44   1818.84   -0.39   0    (60)   1    MTDQEAIQDLWQWR
☑  120    918.24   1834.47   1834.83   -0.36   0     83    1    MTDQEAIQDLWQWR + 1 Oxidation (M)

Proteins matching the same set of peptides:
gi|5876282    Mass: 20076   Total score: 169  Peptides matched: 4
    df43b08.y1 Morton Fetal Cochlea Homo sapiens cDNA clone IMAGE:2485982 5'
gi|6890852    Mass: 21515   Total score: 169  Peptides matched: 4
    RC0-PT0006-271199-011-C05 PT0006 Homo sapiens cDNA
```

(b)

```
36. gi|10315208    Mass: 25984   Total score: 166  Peptides matched: 6
    601678979F1 NIH_MGC_53 Homo sapiens cDNA clone IMAGE:3961554 5'
□ Check to include this hit in archive report

   Query  Observed  Mr(expt)  Mr(calc)  Delta  Miss Score  Rank  Peptide
☑    5    402.17    802.33    802.45   -0.12   0     33    1    TVSLGAGAK
    50    414.17   1239.48   1239.69   -0.21   2     18    4    SAPGGGSKVPQKK
☑  140   1121.88   2241.75   2242.20   -0.46   0     70    1    MSVQPTVSLGGFEITPPVVLR + 1 Oxidation (M)
☑  141    748.26   2241.75   2242.20   -0.45   0    (52)   1    MSVQPTVSLGGFEITPPVVLR + 1 Oxidation (M)
☑  160    982.29   2943.84   2944.45   -0.61   1    (45)   1    TVSLGAGAKDELHIVEAEAMNYEGSPIK + 1 Oxidation (M)
   161    736.97   2943.85   2944.45   -0.60   1     46    2    TVSLGAGAKDELHIVEAEAMNYEGSPIK + 1 Oxidation (M)

Proteins matching the same set of peptides:
gi|8115409    Mass: 20037   Total score: 166  Peptides matched: 6
    QV1-DT0071-230200-083-c03 DT0071 Homo sapiens cDNA
gi|8115410    Mass: 19214   Total score: 166  Peptides matched: 6
```

(c)

```
3. Hs.173205      Total score: 372  Peptides matched: 10
   NPM1 nucleophosmin (nucleolar phosphoprotein B23, numatrin)
□ Check to include this hit in archive report

   Query  Observed  Mr(expt)  Mr(calc)  Delta  Miss Score  Rank  Peptide
     1    430.11    429.10    428.19    0.91   0      6    6    DAAP + 1 Acetyl (N-term); 1 Methyl ester (DE)
☑    5    402.17    802.33    802.45   -0.12   0     33    1    TVSLGAGAK
☑   17    466.16    930.31    930.47   -0.16   0     68    1    GPSSVEDIK
    50    414.17   1239.48   1239.69   -0.21   2     18    4    SAPGGGSKVPQKK
☑  118    910.23   1818.44   1818.84   -0.39   0    (60)   1    MTDQEAIQDLWQWR
☑  120    918.24   1834.47   1834.83   -0.36   0     83    1    MTDQEAIQDLWQWR + 1 Oxidation (M)
☑  140   1121.88   2241.75   2242.20   -0.46   0     70    1    MSVQPTVSLGGFEITPPVVLR + 1 Oxidation (M)
☑  141    748.26   2241.75   2242.20   -0.45   0    (52)   1    MSVQPTVSLGGFEITPPVVLR + 1 Oxidation (M)
☑  160    982.29   2943.84   2944.45   -0.61   1     45    1    TVSLGAGAKDELHIVEAEAMNYEGSPIK + 1 Oxidation (M)
☑  161    736.97   2943.85   2944.41   -0.56   0     49    1    TVSLGAGAQDELHIVEAEAMNYEGSPIK + 1 Oxidation (M)
```

Figure 2. The UniGene index can be used to group ESTs that correspond to the same gene

Upper and middle screenshots show two hits from a search using the Mascot search engine of a standard liquid chromatography–mass spectrometry–mass spectrometry (LC–MS–MS) dataset against dbEST. The lower screenshot shows the same peptide matches grouped into a single hit using UniGene (http://www.ncbi.nlm.nih.gov/UniGene/).

accuracy peptide molecular weight data from matrix-assisted laser desorption–ionization (MALDI) could be used to map the exon–intron boundaries. This approach was also used to search a compilation of human genome sequence data from GenBank, representing some 80% of the complete genome. In a database of this size, short sequence tags yielded multiple matches. However, the authors were generally able to select a single match by manual reconciliation of the putative sequences with the complete MS–MS spectrum.

The feasibility of searching a complete eukaryotic genome using uninterpreted MS–MS data was first investigated by Choudhary *et al.*[15] An LC–MS–MS dataset from a tryptic digest of human embryonic kidney cell lysate

Table 1. Sequence database sources

Type	Name	Compiler	Download URL
Non-redundant protein	nr	National Center for Biotechnology Information (Bethesda, MD, USA)	ftp://ncbi.nlm.nih.gov/blast/db/nr.Z
Non-identical protein	MSDB	Proteomics Department, Imperial College London (London, UK)	ftp://ncbi.nlm.nih.gov/repository/MSDB/msdb.fasta.Z
Non-redundant protein	NRP	National Cancer Institute's Advanced Biomedical Computing Center (Frederick, MD, USA)	ftp://ftp.ncifcrf.gov/pub/nonredun/protein.nrdb.Z
Expressed sequence tag	dbEST	National Center for Biotechnology Information	ftp://ncbi.nlm.nih.gov/blast/db/est.Z
Human genome draft assembly	HG	International Human Genome Sequencing Consortium (University of California at Santa Cruz, CA, USA)	http://genome.cse.ucsc.edu/goldenPath/12dec2000/bigZips/

Table 2. Peptide matching statistics for Mascot searches of a standard LC–MS–MS dataset against three types of database[a]

Category	MSDB	dbEST	HG
Top match with significant ions score	74	56	33
Top match, but ions score not significant	26	37	13
Not top match and ions score not significant	10	11	11
No match because of higher scoring, non-significant matches	0	6	11
No match because peptide sequence not found in MSDB	4	0	0
No match because peptide sequence not found in dbEST	0	2	0
No match because coding sequence substantially missing from HG	0	0	15
No match because coding sequence poorly aligned in HG	0	0	10
No match because peptide spans exon/intron boundary in HG	0	0	19
No match because peptide results from non-tryptic post-translational processing	0	2	2

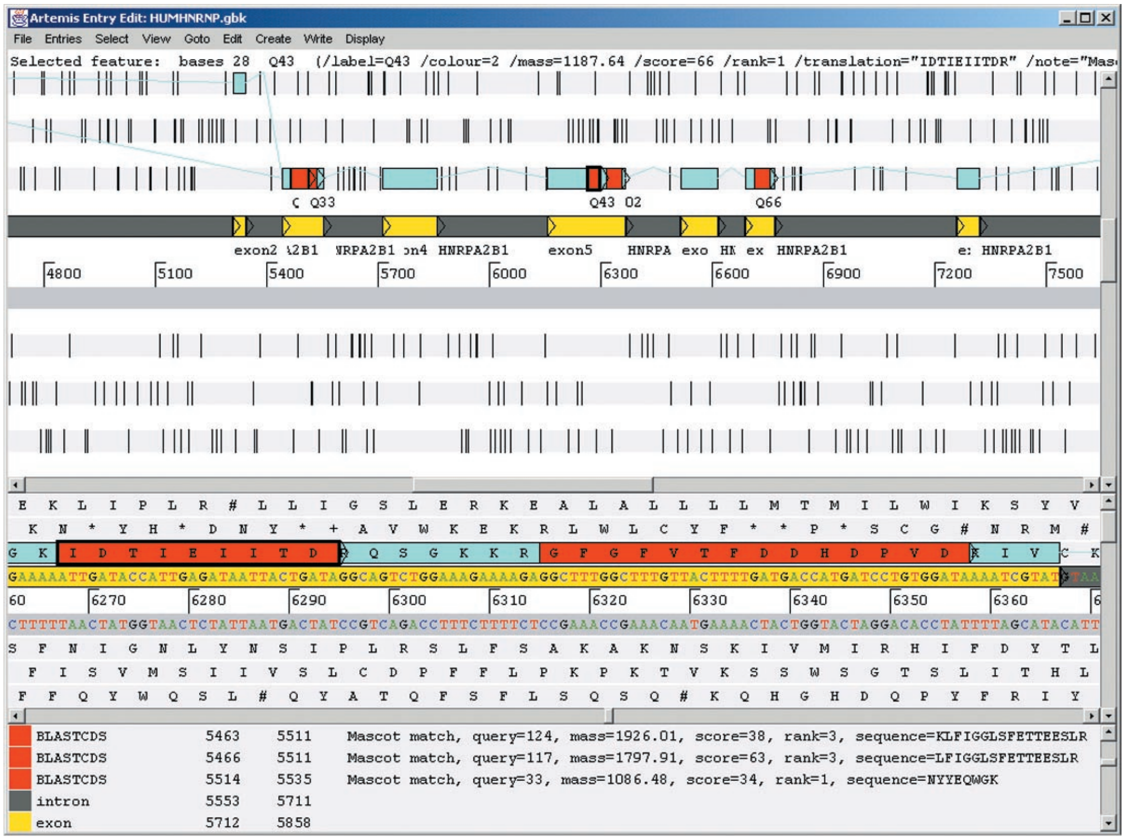
[a] Abbreviations: dbEST, a database of expressed sequence tags; HG, the International Human Genome Project draft assembly of the human genome; MSDB, a comprehensive, non-identical protein database; LC, liquid chromatography; MS, mass spectrometry. Adapted from Ref. 15.

containing peptides from at least 22 human proteins, was searched against a comprehensive, non-identical protein database (MSDB), dbEST, and the International Human Genome Project draft assembly of the human genome (HG)[1] (Table 1). The search engine was Mascot[12], and the same set of search parameters was used for all searches.

HG was searched as both intact chromosome length sequences and also as 600 Kbp segments with small (600 bp) overlaps. Table 2 contains summary statistics for the observed peptide matches. After data reduction, the LC–MS–MS dataset contained 169 spectra. Of these, 114 spectra were matched to peptides from human proteins and 11 spectra were matched to a non-human protein (bovine trypsin). A balance of 44 spectra remained unmatched. Reasons for failing to match a spectrum when searching a protein database are as follows:

- the peptide sequence is not in the database;
- the presence of an unsuspected post-translational modification;
- the peptide is a result of non-specific cleavage;
- the spectrum is of poor quality; and
- the spectrum is of a non-proteinaceous contaminant.

Several factors caused a significant reduction in the number of matches found in HG. One was that several well-characterized mRNAs were partially or completely missing from the November 2000 assembly, such as transaldolase (gene TALDO1, Swiss-Prot TAL1_HUMAN, GenBank mRNA L19437). Another factor was that dbEST contains extensive redundancy, whereas HG represents a single consensus sequence. The sequences in both dbEST and HG contain 'errors' of various types: experimental sequencing errors, misalignments, polymorphisms, and so



```
Selected feature:  bases 28  Q43  (/label=Q43 /colour=2 /mass=1187.64 /score=66 /rank=1 /translation="IDTIEIITDR" /note="Mas
```

```
E  K  L  I  P  L  R  #  L  L  I  G  S  L  E  R  K  E  A  L  A  L  L  L  L  M  T  M  I  L  W  I  K  S  Y  V
 K  N  *  Y  H  *  D  N  Y  *  +  A  V  W  K  E  K  R  L  W  L  C  Y  F  *  *  P  *  S  C  G  #  N  R  M  #
G  K  I  D  T  I  E  I  I  T  D  R  Q  S  G  K  K  R  G  F  G  F  V  T  F  D  D  H  D  P  V  D  K  I  V  C  K
GAAAAATTGATACCATTGAGATAATTACTGATAGGCAGTCTGGAAAGAAAAGAGGCTTTGGCTTTGTTACTTTTGATGACCATGATCCTGTGGATAAAATCGTATTAA
60       6270      6280      6290      6300      6310      6320      6330      6340      6350      6360      6
CTTTTTAACTATGGTAACTCTATTAATGACTATCCGTCAGACCTTTCTTTTCTCCGAAACCGAAACAATGAAAACTACTGGTACTAGGACACCTATTTTAGCATACATT
S  F  N  I  G  N  L  Y  N  S  I  P  L  R  S  L  F  S  A  K  A  K  N  S  K  I  V  M  I  R  H  I  F  D  Y  T  L
 F  I  S  V  M  S  I  I  V  S  L  C  D  P  F  F  L  P  K  P  K  T  V  K  S  S  W  S  G  T  S  L  I  T  H  L
F  F  Q  Y  W  Q  S  L  #  Q  Y  A  T  Q  F  S  F  L  S  Q  S  Q  #  K  Q  H  G  H  D  Q  P  Y  F  R  I  Y
```

```
🟥 BLASTCDS    5463   5511   Mascot match, query=124, mass=1926.01, score=38, rank=3, sequence=KLFIGGLSFETTEESLR
🟥 BLASTCDS    5466   5511   Mascot match, query=117, mass=1797.91, score=63, rank=3, sequence=LFIGGLSFETTEESLR
🟥 BLASTCDS    5514   5535   Mascot match, query=33, mass=1086.48, score=34, rank=1, sequence=NYYEQWGK
⬛ intron      5553   5711
🟨 exon        5712   5858
```

Figure 3. Peptide match results displayed in a genome browser

Peptide matches to heterogeneous nuclear ribonucleoprotein A2/B1 (gene hnRPA2B1) are displayed using Artemis, a Java-based genome browser developed by the Sanger Centre, Cambridge, UK.

on. However, there is a good chance that one of the redundant sequences in dbEST contains the exact sequence being sought, whereas the chance of the consensus sequence in HG being correct falls off rapidly with the error density.

HG showed the largest number of matches missed because there were higher scoring (though non-significant) random matches. Although dbEST (2.3 Gbp) and HG (3.3 Gbp) are similar in size, dbEST contains a much greater degree of redundancy. In other words, a richer variety of random sequences can be found in HG.

Besides the degree of redundancy, the other factor that affects the scoring distribution for random matches is the absolute size of the database. *In silico* digestion of the six-frame translation of HG produces some 1.5 × 10^5 tryptic limit peptides of mass 1500 Da ± 0.5. Without the constraint of tryptic cleavage, this number increases to 6 × 10^7. Allowing for the possibility of non-quantitative post-translational modifications further increases the size of the search space, with a geometric dependence on the number of different modifications.

Taking the no-enzyme case, it becomes impossible to get a significant match to a peptide shorter than seven residues, because any given sequence of this length can be expected to occur by chance (there are nominally $20^6 = 6.4 \times 10^7$ possible six residue peptides). With longer peptides, a significant match is possible only when the spectrum contains sufficient information. The availability of multiple fragmentation pathways, finite mass measurement uncertainty, the mass degeneracy of certain residues and combinations of residues, and many other factors mean that an MS–MS spectrum is rarely, if ever, an unambiguous representation of a unique sequence. Hence, peptides usually need to be significantly longer than the minimum to get a positive match.

The remaining factor causing a significant loss of matches in the HG search was exon boundary crossing. The average size of a human exon is approximately 200 bp or 65 residues[21]. Taking 15 residues as a representative peptide length, the average probability of a randomly chosen peptide spanning an exon or intron boundary is approximately 0.23, in close agreement with the observed loss of 19 matches.

Result presentation and evaluation

The standard report formats of programs such as Mascot were designed for searches of databases comprising relatively short sequences[12]. The top-level report is a tabular summary of the peptide matches, grouped by protein. For each protein, there is a link to a second-level report (the 'protein view'), showing the matches for that particular protein or EST as highlights on the complete sequence, and also as highlighted entries in a table of the predicted proteolytic peptides. In the case of a nucleic acid sequence, there will often be frame shifts, necessitating multiple protein views, one for each frame in which matches were found.

Such reports become unwieldy with very long sequences, where the need is for graphical tools, capable of zooming and panning around the sequence to investigate the spatial distribution of peptide matches. We have adopted Artemis[22], a Java-based genome browser developed by the Sanger Centre (Cambridge, UK), as a tool for reviewing the results from searching MS–MS data. To transfer Mascot results into Artemis, the standard report script was modified to write out peptide match data in the format of an EMBL or GenBank feature table[23,24].

Figure 3 shows matches from a search of HG to coding sequences of the hnRPA2B1 gene, which has two splicing patterns giving rise to the A2 and B1 isoforms of heterogeneous nuclear ribonucleoprotein[15]. The upper panel of the Artemis window provides an overview. This can be zoomed out to show the entire genome as a single strip. In the centre are strips representing the sense and antisense DNA strands above and below, which are strips representing the six-frame translations. The vertical bars are stop codons. Exons from the feature table are shown as light blocks on the DNA strips, and the corresponding coding sequences are marked on the protein strips. Individual peptide matches are shown as small, dark blocks within the coding sequences.

Below this panel is a similar arrangement, but at higher resolution. This portion of the display has been zoomed to the point at which individual bases and residues can be seen. Finally, the lower panel of the Artemis window shows a tabular view of the feature table. When a match is selected, it is highlighted in all three views, making it possible to see detail such as the spectrum number, sequence, molecular weight, and Mascot score.

Concluding remarks

The HG assembly remains a draft, and contains artefacts that will be corrected over the coming months as it is refined and annotated. MS can play an important role in this process, providing experimental verification of predicted coding sequences and assisting in the identification of novel features. Matching MS–MS data is a more laborious method of gene mining than purely computational methods, such as sequence alignment or exon prediction, but it has the advantage of being an experimentally-based orthogonal approach, capable of finding truly novel genes. Bioinformatic tools for high-throughput work are in development, but much work remains to be done before we can extract and utilize all of the information encoded in raw MS data. Like the genome, the current generation of software could also be regarded as work in progress.

References

1 International Human Genome Sequencing Consortium (2001) Initial sequencing and analysis of the human genome. *Nature* 409, 860–921

2 Venter, J.C. *et al.* (2001) The sequence of the human genome. *Science* 291, 1304–1351

3 Burge, C.B. and Karlin, S. (1998) Finding the genes in genomic DNA. *Curr. Opin. Struct. Biol.* 8, 346–354

4 Stormo, G.D. (2000) Gene-finding approaches for eukaryotes. *Genome Res.* 10, 394–397

5 Adams, M.D. *et al.* (1991) Complementary DNA sequencing: expressed sequence tags and human genome project. *Science* 252, 1651–1656

6 Yates, J.R., III (1998) Database searching using mass spectrometry data. *Electrophoresis* 19, 893–900

7 Beavis, R.C. and Fenyö, D. (2000) Database searching with mass spectrometric information. In *Proteomics: A Trends Guide* (Blackstock, W. and Mann, M., eds), pp. 22–27, Elsevier

8 Mann, M. and Wilm, M. (1994) Error-tolerant identification of peptides in sequence databases by peptide sequence tags. *Anal. Chem.* 66, 4390–4399

9 Mann, M. (1994) Sequence database searching by mass spectrometric data. In *Microcharacterisation of Proteins* (Kellner, R. *et al.*, eds), pp. 223–245, VCH

10 Eng, J.K. *et al.* (1994) An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* 5, 976–989

11 Clauser, K.R. *et al.* (1999) Role of accurate mass measurement (+/− 10 ppm) in protein identification strategies employing MS or MS/MS and database searching. *Anal. Chem.* 71, 2871–2882

12 Perkins, D.N. *et al.* (1999) Probability-based protein identification by searching sequence databases using mass spectrometry data. *Electrophoresis* 20, 3551–3567

13 Yates, J.R., III *et al.* (1995) Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases. *Anal. Chem.* 67, 3202–3210

14 Neubauer, G. *et al.* (1998) Mass spectrometry and EST-database searching allows characterization of the multi-protein spliceosome complex. *Nat. Genet.* 20, 46–50

15 Choudhary, J.S. *et al.* (2001) Interrogating the human genome using uninterpreted mass spectrometry data. *Proteomics* 1, 652–667

16 Chen, W. *et al.* (2001) Searching the *Porphyromonas gingivalis* genome with peptide fragmentation mass spectra. *Analyst* 126, 52–57

17 Link, A.J. *et al.* (1997) Identifying the major proteome components of *Haemophilus influenzae* type-strain NCTC 8143. *Electrophoresis* 18, 1314–1334

18 Regula, J.T. *et al.* (2000) Towards a two-dimensional proteome map of *Mycoplasma pneumoniae*. *Electrophoresis* 21, 3765–3780

19 Link, A.J. *et al.* (1999) Direct analysis of protein complexes using mass spectrometry. *Nat. Biotechnol.* 17, 676–682

20 Küster, B. *et al.* (2001) Mass spectrometry allows direct identification of proteins in large genomes. *Proteomics* 1, 641–650

21 Strachan, T. and Read, A.P. (1999) *Human Molecular Genetics* 2, BIOS Scientific Publishers Ltd

22 Rutherford, K. *et al.* (2000) Artemis: sequence visualisation and annotation. *Bioinformatics* 16, 944–945

23 Baker, W. *et al.* (2000) The EMBL nucleotide sequence database. *Nucleic Acids Res.* 28, 19–23

24 Benson, D.A. *et al.* (2000) GenBank. *Nucleic Acids Res.* 28, 15–18