Cross-species protein identification in proteomics *via* protein profiles

Patrick J Lester, Christian Cole, Simon J Hubbard*

Department of Biomolecular Sciences, UMIST, PO Box 88, Manchester M60 1QD, UK.

Received line

ABSTRACT

Protein identification from mass spectra of tryptic peptides relies on bioinformatic software to determine the most likely matching protein from a database that contains the protein sequence. However, enabling cross-species proteomics is equally important for the many species currently without sequenced genomes. We present a methodology to address this using profiles of related protein sequences against which to search. Using simulated data, we show that tryptic peptide conservation is enriched above random in these protein profiles, and a search algorithm developed from these shows an improvement over simple single-orthologue search methods.

Contact: Simon.Hubbard@umist.ac.uk

INTRODUCTION

The ability to study genes and proteins on a genome-wide basis is driving a paradigm shift in modern biology from hypothesis-driven to hypothesis-generating bioscience, where the state of every gene and/or gene product in different biological conditions may be studied. Driven by complete genome sequences, proteomics is one example of these techniques, which offers a key advantage over many other post-genomic functional studies in that proteins are the functional entities in cells or tissues. At the heart of proteome science lies advances in mass spectrometry and protein/peptide separation along with, crucially, attendant bioinformatics tools (Andersen and Mann, 2000).

One of the main proteomics methods is that of the peptide mass fingerprinting (PMF) approach that supports high-throughput analysis of many protein samples. Here, the digested protein (usually using trypsin) is analysed in the mass spectrometer and the peptide fragment mass-to-charge ratios (m/z) determined. These 'mass fingerprints' are then used to search a sequence database for the most likely candidate protein to have produced these peptide masses. This problem is tractable since the masses of all 20 common amino acids are known and theoretical tryptic peptide masses can therefore be calculated to high accuracy. For a correct assignment to be made confidently, the database must include the whole proteome for the species of interest. However, this often is not the case, with many species of agricultural importance and pathogenic interest not having a complete genome sequence.

Although it is straightforward to match protein or peptide sequences from even distantly related species *via* homology searching with algorithms such as BLAST, a single amino acid change can alter a peptide mass by tens of mass units, when they are normally measured to accuracies better than 0.5 Da. This makes cross-species protein identification *via* PMF particularly challenging (Lester and Hubbard, 2002; Liska and Shevchenko, 2003). Existing approaches have relied on using closely related species where the protein sequences are expected to be very similar, or where certain peptides are highly conserved (Cordwell *et al*., 1997; Wasinger *et al*., 1999). Alternatively, they have exploited tandem mass spectrometry to gain some partial (or complete) sequence information to allow sequences to be compared across species with BLAST-like algorithms (Shevchenko *et al*., 2001). However, as has been shown in bacterial proteomes, peptide masses are conserved above random even in protein pairs of low pairwise sequence identity (Lester and Hubbard, 2002).

In this paper, we introduce a novel approach for cross-species protein identification using PMF. This involves generating protein profiles of sequences from the same family against which to search. These databases are searched in the normal way using standard PMF algorithms, and final scores are generated both for the individual proteins in the database and on the protein profiles, which are then used to evaluate likely cross-species matches. The approach has been tested for differing profile systems and for two test database systems, eutherian mammals and *sensu stricto* yeasts. The algorithm demonstrates successful cross-species protein identification at an improved level above naïve single protein searches to match the closest orthologue. We suggest this provides a useful general approach to cross-species protein identification in proteomics that can be exploited for workers currently working on organisms for which a complete genome (and hence, proteome) sequence is not presently available.

METHODS

Two datasets were generated to test our cross-species profile approach. The first used 752 rodent protein sequences taken from the SwissProt TrEMBL non-redundant database. Rat/mouse sequences were excluded to avoid bias to one another, as they are too similar in sequence. PSI-BLAST profiles were generated for each test sequence *vs.* a database of 64,780 eutherian mammal sequences, made up mainly of human and mouse data and excluding the 752

search sequences. The eutherian mammal database was obtained from the European Bioinformatics Institute (EBI) proteome repository (http://www.ebi.ac.uk/proteome).

The second dataset, used the *S. cerevisiae* proteome (6,212 sequences) obtained from the EBI proteome repository. PSI-BLAST profiles were generated for the *S. cerevisiae* proteome against a database of 32,829 *sensu stricto* yeast proteins obtained from the Whitehead Institute and Washington University (Cliften *et al*, 2003; Kellis *et al*., 2003), consisting of the species: *S. kluyveri, S. kudriavzevii, S. castellii, S. mikatae, S. bayanus* and *S. paradoxus*.

PSI-BLAST was run for a maximum of 3 cycles with an expectation cut-off of 1e-60 to ensure only closely related proteins were included. However, due to the stringent cut-off, not all of the sequences in the datasets resulted in a hit: only 575 rodent profiles and 5,124 yeast profiles were produced. Only sequences generating profiles were used in subsequent PMF simulations. In order to reduce profile size and to avoid noise from paralogous protein sequences, only one sequence from each species was used in the final profiles, comprising a total of 18,838 proteins for the yeast data and 2,370 for the eutherian mammals. All of these sequences were then annotated with the identifier of all the profiles of which they are a member.

Ideally, a cross-species PMF database search would return the putative orthologue as the top hit (equivalent to the top BLAST hit). However, there will be an increased number of chance hits to incorrect proteins in a cross-species context and PMF scoring algorithms will differentiate them with difficulty. In principle, this may be overcome by combining orthologous proteins into profiles, which ensures that the maximum number of 'true' (non-random) peptide matches across the family is rewarded. This should result in improved scores over random matches against 'false' profiles.

When matching N search peptides against a profile, then M out N peptides may match to the first protein in the profile, termed the 'principal orthologue' (the closest relative) in the true profile. In addition, a further G search peptides may match to additional proteins in the profile, which we refer to as the gain. For the cross-species profile approach to be successful the gain, G_{match}, should be larger for the 'true' matching profile than the gain G_{random} for other 'false' profiles. We therefore calculated the net gain, $G_{net} = G_{match} - G_{random}$ for every profile in the dataset as a measure of positive enrichment due to profiles. To calculate G_{random} we selected the first protein in the profile as 'principal orthologue' against which the gain is calculated. These calculations were performed at 3 different mass accuracies, measured in ppm (parts per million) units, at which peptides were deemed to be matching.

In silico simulations were carried out on the reduced rodent and yeast datasets. Theoretical peptide masses were generated equivalent to a tryptic digest for each protein sequence and, using our in-house PMF algorithm (Sidhu *et al*., 2001), searched against either the eutherian mammals or *sensu stricto* yeast databases. A fixed number of search peptide masses (of at least 500 Da to mimic typical peptide masses obtained from MALDI-TOF experiments) were randomly selected from each protein to search against the database and was repeated eight times to ensure a decent coverage of the randomly selected peptides.

The results of each PMF run were dependent on how correct hits were scored/assigned. The first type of search, classed as a Single Orthologue Search (SOS), was a standard PMF search where a correct match was only made when the top-scoring BLASTP hit (the principal orthologue) was returned as the top hit protein. This method was extended (SOS+) to consider any protein from the 'true' profile returned as the top hit as a correct match. Both these methods only assign one protein as the correct match.

In addition to these pairwise searches, two profile searches were undertaken. In this case, a score was assigned to each profile by summing the individual PMF scores for each member protein with at least one tryptic peptide match and then calculating the mean score by dividing each profile total by the number of sequences with a score. The profiles were then ranked on these mean scores. The first profile method, classed as a Single Profile Search (SPS), considered a correct match was produced when the top ranking profile was the one containing the true orthologue of the search protein. The second profile method (SPS+) extended this definition to consider a correct match to have been made when any protein from the 'true' profile was found in the top-scoring profile.

RESULTS

Our previous work has shown that tryptic peptides are conserved across bacterial species boundaries over and above random (Lester and Hubbard, 2002) and Figure 1 shows the same is true for eukaryotic organisms (rodents) across protein profiles. This plot shows the net gain G_{net} in matching tryptic peptides due to the use of protein profiles. The net gain is greater than 1 in the majority of cases, for all ppm mass accuracy measurements considered, although there is a stronger enrichment at lower, more accurate, values. This indicates that there is a clear benefit to be had

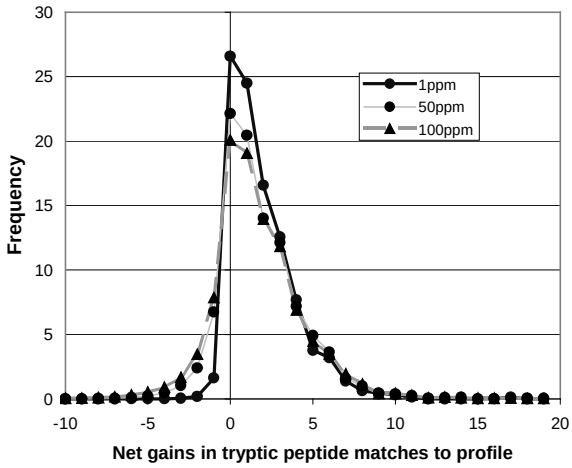

Figure 1. Net gains in peptide mass matches due to profiles compared to random profile matching using the rodent dataset.

from the use of protein profiles in cross-species PMF experiments and that the use of profiles enhances the net conservation of tryptic peptides over and above that which would be expected by chance.

Cross-species protein identification *via* PMF is typically done using tryptic peptide data and publicly available

databases aiming to find an orthologous protein. However, this often yields unsatisfactory results as the correct match is difficult to find amongst the noise of non-specific hits. Indeed, searches we have conducted with standard search tools (Mascot and ProFound) rarely produce a match with a significant score (data not shown). It can be seen in Figure 2 that this naïve approach of searching for a single 'top-matching orthologue', which we call the Single Orthologue Search method (SOS), can be improved upon significantly by the implementation of homologue 'profiling' methods. This is particularly true for the rodent data. The SOS method finds the correct orthologue as the top hit in 45% of the cases. Improving the measured mass accuracy to 5 ppm from 50 ppm only improves this to 47%. However, the use of Single Profile Searching (SPS) significantly improves the correct identification of the protein to 67% at 50 ppm error and 73% at 5 ppm error. The best profile result is achieved using the extended SPS method (SPS+) at 5 ppm error, which yields the correct orthologue ranked as the top hit 84% of the time (Figure 2).

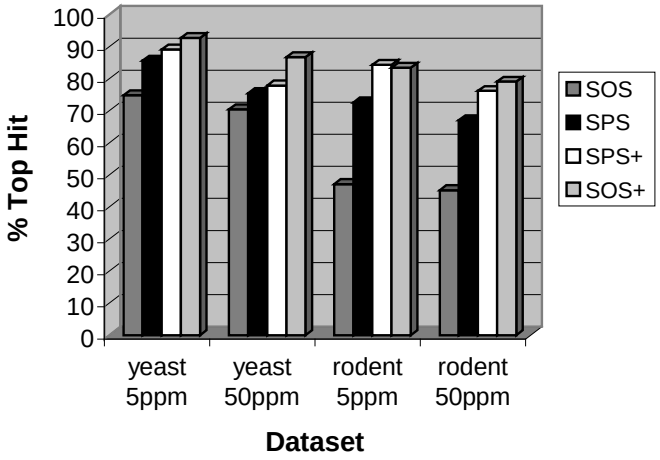

Figure 2. Relative performance of cross-species protein identification strategies at different mass accuracies

The improvement over SOS by the profiling methods is not as large for the yeast species, but is nonetheless significant. The baseline results of 70% and 75% correctly ranked top hits are improved to 87% and 93% using the extended SOS method (SOS+) for 50 and 5 ppm error, respectively. The profile methods, SPS and SPS+, also improved significantly on SOS but not to the same extent as SOS+. Although the SOS+ method is based upon the SOS method, it is important to stress that it is also dependent on the generation of profiles (as used in the SPS and SPS+ methods). Additionally, we have yet to fully optimise the scoring system for the profiling methods, which promises further improvement over SOS+ as a significant enrichment over random matching is observed.

The difference in improvement over SOS between the yeast and the rodent data is likely due to the different diversity in the two datasets. The yeast species are closely related and possess a greater average sequence similarity between orthologues, leading to better results from the naïve SOS approach. The rodent dataset however is more diverse, making the identification of the correct orthologue more difficult, especially for the SOS method. This is where the more advanced profiling methods become useful. The ability to define a set of related sequences as the optimal hit allows the matching of several (potentially) lower hits and combining them into a single top-ranked correct protein identification. Furthermore, the additional success for the more diverse rodent dataset suggests the profile approach is more beneficial in cross-species matching where the expected similarity is modest rather than very closely related species, such as in the yeast sequences.

CONCLUSIONS

In this paper we show that profile-based approaches for cross-species protein identification offers considerable benefits over naïve single-protein strategies. The chief improvement lies in the organisation of the database itself, placing proteins in profiles and removing noise from the data further by slimming the profiles down to include one representative from each species. Furthermore, once organised like this, the ability to spot correct matches to any homologous protein becomes clearer since they are now organised (and importantly, labelled) as protein family groups. This effect is made plain by the improved SOS+ searches where a correct match is recorded when the top-hit is also a member protein of the true 'top-matching' orthologue profile. Further work is underway to develop a more discriminative scoring system based on the results shown in Figure 1 which demonstrates the enrichment above random in tryptic peptide matches to protein profiles containing orthologues of the search protein.

ACKNOWLEDGEMENTS

All authors acknowledge the BBSRC for support.

REFERENCES

Andersen, JS & Mann, M. (2000) Functional genomics by mass spectrometry. *FEBS Lett.*, 480, 25-31.

Cliften, P, Sudarsanam, P, Desikan, A, Fulton, L, Fulton, B, Majors, J, Waterston, R, Cohen, BA & Johnston, M. (2003) Finding functional features in Saccharomyces genomes by phylogenetic footprinting. *Science*, 301, 71-76.

Cordwell, SJ, Wasinger, VC, Cerpa-Poljak, A, Duncan, MW, & Humphery-Smith, I. (1997) Conserved motifs as the basis for recognition of homologous proteins across species boundaries using peptide-mass fingerprinting. *J. Mass Spectrom.*, 32, 370-378.

Kellis, M, Patterson, N, Endrizzi, M, Birren, B & Lander, ES (2003) Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423, 241-254.

Lester, PJ & Hubbard, SJ. (2002) Comparative bioinformatic analysis of complete proteomes and protein parameters for cross-species identification in proteomics. *Proteomics*, 2, 1392-1405.

Liska, AJ and Shevchenko, A. (2003) Expanding the organismal scope of proteomics: Cross-species protein identification by mass spectrometry and its implications. *Proteomics*, 3, 19-28.

Wasinger, VC, Urquhart, BL & Humphrey-Smith, I. (1999) Cross-species characterisation of abundantly expressed *Ochrobactrum anthropi* gene products. *Electrophoresis*, 20, 2196-2203.

Shevchenko, A, Sunyaev, S, Loboda, A, Shevchenko, A, Bork, P, Ens, W & Standing, KG. (2001) Charting the Proteomes of Organisms with Unsequenced Genomes by MALDI-Quadrupole Time-of-Flight Mass Spectrometry and BLAST Homology Searching. *Anal. Chem.*, 73, 1917-1926.

Sidhu, KS, Sangvanich, P, Brancia, FL, Sullivan, AG, Gaskell, SJ, Wolkenhauer, O, Oliver, SG & Hubbard, SJ. (2001) Bioinformatic assessment of mass spectrometric chemical derivatisation techniques for proteome database searching. *Proteomics*, 1, 1368-1377.