SPECLUST: a web tool for clustering of mass spectra

Peter Johansson [a], Rikard Alm [b], Cecilia Emanuelsson [b], Jari Häkkinen [a] and Markus Ringnér [c,d*]

[a]Computational Biology and Biological Physics, Department of Theoretical Physics, [b]Department of Biochemistry, [c]Division of Oncology, Department of Clinical Sciences, [d]CREATE HEALTH, Lund University, Sweden

ABSTRACT

Summary: SPECLUST is a web tool for hierarchical clustering of peptide mass spectra obtained from protease-digested proteins. Mass spectra are clustered according to the peptide masses they contain, such that mass spectra containing similar masses are clustered together. Hierarchical clustering of mass spectra with SPECLUST can in particular be useful for MS-screening of large proteomic data sets derived from 2D-gels. SPECLUST can also be used to identify masses shared by mass spectra. Masses present in the majority of the mass spectra in a data set are likely to be contaminants. With SPECLUST, MS/MS can be focused on non-contaminant shared masses in a cluster, facilitating investigations of protein isoforms. Within a cluster, shared and unique masses represent peptides from regions that are similar and different, respectively, between protein isoforms. Taken together, SPECLUST is a versatile tool for analysis of mass spectrometry data.

Availability: SPECLUST is freely available at http://bioinfo.thep.lu.se/speclust.html.

Contact: markus.ringner@med.lu.se

1 INTRODUCTION

Proteomic data sets generated using mass spectrometric instrumentation often contain protein isoforms for many proteins. For example, as many as 52 spots with the same protein were found in a study using high resolution two-dimensional polyacrylamide electrophoresis to separate proteins (Klose *et al*., 2002). The identification of protein isoforms is one of the major challenges in proteomics. Instead of comparing mass spectra to databases, as is normally done for protein identification, spectra could be compared to other spectra in the data set. One way to do this is to perform hierarchical clustering of the spectra. We have previously described a method for hierarchical clustering of lists of peptide peak masses, typically extracted from protein mass spectra (Alm *et al*., 2006). Here, we present a web tool, SPECLUST, implementing the method. SPECLUST is designed to generate clusters of mass spectra coming from similar proteins and protein isoforms. Thereby, data sets containing very many mass spectra can be reduced to a set of clusters aiding the subsequent analysis. In addition, SPECLUST enables the identification of peak masses in common for mass spectra in a cluster.

Following the initial success of hierarchical clustering when analyzing gene expression data (Khan *et al*., 1998; Eisen *et al*., 1998), it has become the most popular method for microarray data analysis. Clearly, it is of interest to be able to apply hierarchical clustering also to proteomic data. Consequently, hierarchical clustering of mass spectra has been used for many applications in proteomics (see for example Schmidt *et al*. (2003); Jacquemier *et al*. (2005)).

There are numerous software packages and web tools available for hierarchical clustering. However, none of them can be directly used for mass spectrometry data. Hierarchical clustering requires a distance measure between the items to be clustered. Commonly used distance measures are not easily applicable to mass spectra. For SPECLUST we have developed a distance measure shown to perform well for clustering of protein mass spectra (Alm *et al*., 2006), to bring hierarchical clustering to investigations of mass spectrometry data.

2 DESCRIPTION OF THE WEB TOOL

SPECLUST consists of three applications. Additional details for each application is available at the SPECLUST web site.

Clustering. This application has a simple user interface that consists of four input fields: upload of lists of peak masses to be analyzed, selection of metric to use for distances between peak lists, specification of measurement uncertainty in mass (σ) to use when matching two peaks from different peak lists, and selection of linkage method to use for joining clusters. The mass spectra are uploaded as a zip file containing peak list files, each containing peak masses for a spectrum. The distance metrics between peak lists all depend on the number of peaks in one spectrum that are matched to peaks in the other spectrum: the more peaks that are matched, the smaller the distance between the peak lists. The matching of peaks is based on a peak match score that reflects the probability that two peaks originate from the same peptide. The score is zero for measurements infinitely apart and unity for measurements being identical. The measurement uncertainty, σ, is used such that for a given mass difference higher peak match scores are obtained as sigma is increased. The matching of peaks is done using the Needleman and Wunsch algorithm (Needleman and Wunsch, 1970). The application supports the three most commonly used linkage methods (Quackenbush, 2001).

We have previously shown that using the liberal metric, average linkage, and σ set to one Dalton performs well when clustering MALDI-MS mass spectra for protein isoforms separated on two-dimensional electrophoresis gels (Alm *et al*., 2006). The output of the application is a dendrogram generated by the hierarchical clustering. Peak lists that share many masses end up close to one another in the dendrogram. The dendrogram can be downloaded as vector graphics in a pdf file providing flexible access to the dendrogram

*to whom correspondence should be addressed

when preparing figures for publication. The dendrogram can be seamlessly submitted to the cluster identification application.

Cluster identification. This application uses the dendrogram resulting from the clustering as a starting point. The input to cluster identification is a cut-off value for the distance in the dendrogram. Mass spectra joined in nodes at distances below this cut-off are considered a cluster. The output of the application is a zip file for each cluster that contains files with lists of peak masses for the mass spectra in the cluster. The user interface to this application has input fields that allows the user to manually reassign to which cluster each spectrum belongs. Each cluster can be submitted to the peaks in common application for further analysis.

Peaks in common. This application identifies masses shared by a set of mass spectra: peaks in common. The user interface consists of five input fields: upload of lists of peak masses to be analyzed, specification of measurement uncertainty in mass (σ) to use when matching two peaks from different peak lists, pairwise-, multiple-, and consensus score cut-offs. The lists of peak masses should be a zip file containing mass spectrum files as for the clustering application. The zip file can either be uploaded or submitted through the cluster identification application. The measurement uncertainty is used when matching peaks in the same way as in the clustering application. The three different score cut-offs are used to generate three different outputs.

First, all pairs of peak lists are investigated for shared peaks. Peaks are considered pairwise in common if they are matched and the peak match score is larger than the pairwise cut-off. For each pair of peak lists, peaks that are pairwise in common are printed.

Second, peak lists are investigated for peaks shared with multiple other peak lists. The peak match scores for a peak are summed across all other peak lists for which the peak is pairwise in common. A multiple score is constructed by dividing this sum with the number of peak lists investigated. A peak is considered in common with multiple other peak lists if its multiple score is larger than the multiple cut-off. For each peak list the peaks in common with multiple other peak lists are printed

Third, a consensus peak list is generated for the set of peak lists investigated. The consensus peak list is obtained by partitioning all peaks for all peak lists into sets of peaks. The sets are built up such that the more peak lists that share a peak, the larger the set with the peak. Sets containing more than consensus cut-off peaks are considered consensus peaks. Statistics such as average and standard deviation are printed for consensus peaks.

DISCUSSION

We have developed a web tool, SPECLUST, for hierarchical clustering of mass spectra. A major benefit of the clustering is the independence of protein databases, which makes it suitable for organisms with unsequenced genomes. SPECLUST integrates clustering with identification of peak masses in common for mass spectra in a cluster. In addition to the overview and data reduction provided by the dendrogram generated by the clustering, we have found SPECLUST to be useful for four different purposes. First, peaks shared by many spectra are likely to be contaminants (Levander *et al*., 2004). Such peaks can be identified by running peaks in common for all spectra under investigation with the consensus cut-off set to, for example, 10-30% of the total number of spectra (Alm *et al*., 2006). In this way a consensus list of all repeatedly occurring peaks is generated. These peaks, likely to stem from contaminants can be removed from the mass spectra. Second, peaks in common can be used to improve identification of the proteins in a cluster. This improvement can be achieved both by submitting a peak list containing only peak masses found to be in common for a cluster to standard peptide mass fingerprinting methods (Alm *et al*., 2006), and by focusing further MS/MS analysis of peptides on such shared peaks (Alm *et al*., 2007). Third, peaks in common can be used for analysis of protein isoforms by identifying shared and unique masses within a cluster. Such masses represent peptides from regions that are similar and different, respectively, between protein isoforms. Fourth, peaks in common that do not match theoretical peptide masses may represent modified peaks that can be used to characterize the post-translational modifications of the protein isoforms (Alm *et al*., 2006). These examples suggest that SPECLUST is a versatile tool for analysis of in particular MALDI-MS mass spectra derived from two-dimensional electrophoresis gels.

ACKNOWLEDGMENT

RA and CE were supported by the Swedish Research Council for Environment, Agricultural Science and Spatial Planning (FORMAS). JH was supported by the Knut and Alice Wallenberg Foundation through the Swegene consortium. MR was supported by the SSF Strategic Center for Clinical Cancer Research - CREATE Health.

Conflict of Interest: none declared.

REFERENCES

Alm,R. *et al.* (2006) Detection and identification of protein isoforms using cluster analysis of MALDI-MS mass spectra, *J. Proteome Res.*, **5**, 785-792.

Alm,R. *et al.* (2007) Proteomic variation is larger within than between strawberry varieties, submitted to *J. Proteome Res.*

Eisen,M.B.*et al.* (1998) Cluster analysis and display of genome-wide expression patterns, *Proc. Natl. Acad. Sci. USA*, **95**, 14863-14868.

Jacquemier,J. *et al.* Protein expression profiling identifies subclasses of breast cancer and predicts prognosis. *Cancer Res.* **65**, 767-779.

Khan,J. *et al.* (1998) Gene expression profiling of alveolar rhabdomyosarcoma with cDNA microarrays, *Cancer Res.*, **58**, 5009-5013.

Klose,J. *et al.* (2002) Genetic analysis of the mouse brain proteome, *Nat. Genet.*, **30**, 385-393.

Needleman,S.B. and Wunsch,C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J. Mol. Biol.* **48**, 443-453.

Quackenbush,J. (2001) Computational analysis of microarray data, *Nat. Rev. Genet.* **2**, 418-427.

Levander,F. *et al.* (2004) Automated methods for improved protein identification by peptide mass fingerprinting. *Proteomics*. **4**, 2594-2601.

Schmidt,F. *et al.* (2003) Iterative data analysis is the key for exhaustive analysis of peptide mass fingerprints from proteins separated by two-dimensional electrophoresis. *J. Am. Soc. Mass. Spectrom.* **14**, 943-956.