Supplementary Material for: Clustering Millions of Tandem Mass Spectra

Ari M. Frank[1*] Nuno Bandeira[1] Zhouxin Shen[2] Stephen Tanner[3]
Steven P. Briggs[2] Richard D. Smith[4] Pavel A. Pevzner[1*]

October 4, 2007

1 Spectral Similarity

In order to cluster mass spectra we need to determine the similarity between them. We use the normalized dot-product, which has previously been found to work well by several groups that have approached similar problems [1, 2, 3, 4, 5, 6, 7, 8].

To calculate the normalized dot-product of two mass spectra S and S', we first reduce each spectrum to a vector. Since the computation of the spectral similarity is a major part of the clustering algorithm, restricting the size of theses vectors can reduce the running time. To construct such vectors we first select the k strongest peaks from S and S' (we assume that S and S' have similar precursor masses). Joining these two sets of masses yields a set of masses $M = \{m_1, \ldots, m_t\}$, where $k \leq t \leq 2k$. M may contain less than $2k$ masses because duplicate masses are removed (we consider two peaks to have a similar mass if they are within 0.5 Da from each other). Finally, we reduce the spectrum S to a vector $s = s_1, \ldots, s_t$ by assigning to each s_i the intensity found at mass m_i in S if m_i was one of the top k peaks in S, otherwise 0 is given to that position. Similarly, we fill s' using the intensities of the peaks in S'. In our experiments we found that for these similarity computations it is optimal to set k to a value that corresponds to 15 peaks per 1000 Da of peptide mass. Once spectra S and S' are converted to vectors, their normalized dot-product is given by

$$Similarity(S, S') = \frac{\sum_{i=1}^{t} s_i \cdot s'_i}{\sqrt{\sum_{i=1}^{t} {s_i}^2 \cdot \sum_{i=1}^{t} {s'_i}^2}} \quad (1)$$

The normalized dot-product takes values between 0 (when spectra do not share any selected peaks) and 1.

Dot-products were initially used for measuring similarity between mass spectra of chemical compounds, whose mass spectra typically contain a small number of peaks [1]. Directly applying this measure to spectra of peptides can yield suboptimal results since a small number of strong peaks in the spectrum can dominate the outcome of the spectral similarity computation. Scaling peak intensities has been shown to improve the quality of the similarity computations [1]. One method that has been suggested is to scale a peak's intensity according to the square root of the intensity [1, 9, 7]. The scaling method we found most suitable for our data was to first normalize the peak intensities to bring the total spectrum's intensity to 1000 and then fill the dot-product vectors with the natural logarithm of the selected peaks' intensities.

2 Consensus Spectra

A common approach for creating a representative spectrum for a cluster is to use a consensus spectrum [4, 5, 10, 7, 8, 11], which is generated by "summing" the spectra in the cluster. Our method for creating a consensus spectrum is as follows. Given the cluster's mass spectra, we create a single merged peak list for all the spectra, and sort the list according to the peaks' masses. The list is then scanned and when a pair of adjacent peaks having a mass difference below a specified tolerance is detected, the peaks are consolidated to a single peak with a mass that equals the weighted average of the joined peaks' masses and an intensity that equals the sum of the joined peaks' intensities. To increase the accuracy of the peak joining, the process is repeated several times with an increasing tolerance threshold (the final threshold we used was 0.4 Da). This is done to avoid erroneous peak merging due to isotopic peaks, etc.

To increase the peptide's signal in the spectrum, we take advantage of the fact that peaks corresponding to genuine fragments are likely to appear in many of the cluster's spectra. Thus for each peak i in the consensus spectrum, we take note of the number peaks from the original spectra that were merged to create i and divide it by the total number of spectra to obtain the peak probability p_i. We then multiply the peak i's intensity by a scaling factor $\alpha = 0.95 + 0.05 * (1 + p_i)^5$. This function gives α a value close to 1 for peaks with low probability, but increases as the probability nears 1 to a maximal value of 2.55. Finally the list of peaks in the consensus spectrum is filtered using a sliding window to filter out weak peaks (in our experiments we kept the top 5 peaks in a window of 100 Da).

We considered five alternatives for a cluster's representative.

1. "best spectrum": the spectrum that maximizes a certain score, e.g., percent of explained intensity or percent of explained b/y ions (this is the optimal spectrum that could be selected from amongst the cluster members).

2. "consensus spectrum": a virtual spectrum constructed by consolidating all spectra in the cluster.

3. "most similar spectrum": the spectrum that has the highest average similarity to the other cluster members [3, 9].

4. "de novo spectrum": the spectrum that has the highest score when submitted to de novo sequencing.

5. "average spectrum": a spectrum chosen from the cluster at random.

We start off by evaluating different methods for selecting a cluster representative. Figure 1 shows plots in which we examine the relation between the cluster size and the quality of different types of cluster representatives. The plots were generated from 250 clusters each containing at least 100 spectra from the Human dataset which were identified with high confidence by InsPecT. The spectra were filtered using a sliding window to maintain a peak density of approximately 50 peaks per 1000 Da of peptide mass. For each cluster size, we repeatedly drew random subsets (clusters) varying in size from 1 to 100, taken from the spectra of the original 250 large clusters. For each drawn cluster of spectra corresponding to a peptide P, we examined the percent of explained intensity (i.e., the sum of the intensities of peaks belonging to fragment ions of P), the proportion of P's b- and y-ions that were observed in the spectra and the score given to the spectrum by InsPecT when annotated with the peptide P. These three statistics were recorded for five different methods for selecting cluster representatives.

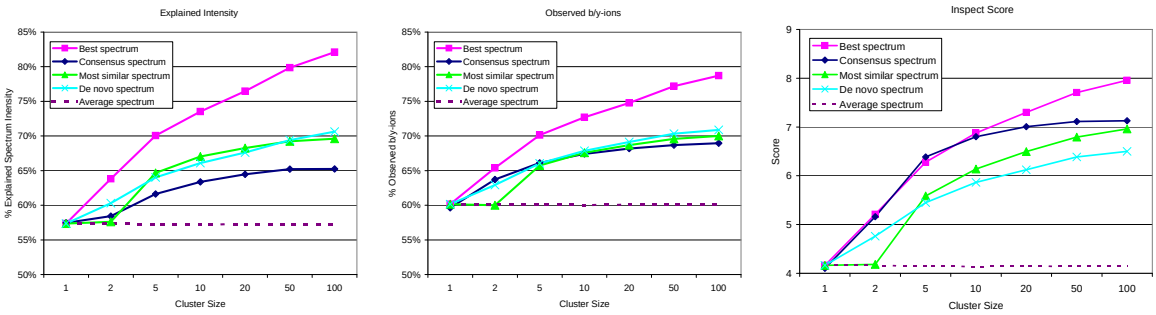

Figure 1: Cluster size and spectrum quality. Clusters of various sizes were evaluated to determine the fraction of explained spectrum intensity (left), proportion of observed b- and y-ions (center), and score given to the spectrum by Inspect (right). With each cluster these statistics were collected for five different cluster representatives: 1) The best spectrum, 2) The consensus spectrum, 3) The most similar spectrum, 4) The best de novo spectrum, and 5) The average spectrum.

Figure 1 illustrates the benefits of selecting cluster representatives "wisely". Using representatives 2-4 gave spectra with a significantly higher signal-to-noise ratio than the "average" representative (5). The most similar spectrum and the top de novo spectrum have higher proportions of explained intensity (up to 5% more) than the consensus spectrum, but relatively similar proportions of observed b- and y-ions. However ultimately, when the spectra are submitted to a database search, the consensus spectra have higher Inspect scores than the other methods (except for selecting the "best" which we only know how to identify after searching all cluster members). In fact with clusters of up to 10 spectra, the consensus spectra and the best spectra in the clusters have similar Inspect scores (with a slight advantage for consensus spectra at size 5 which get a score of 6.4 compared to best spectrum's score of 6.3). We therefore decided to use consensus spectra as the cluster representatives for our clustering algorithm.

3 Clustering Heuristics

We use two heuristics to reduce the number of similarity computations performed by our clustering algorithm. The first heuristic evaluates how likely it is that two spectra belong to the same peptide, without explicitly computing the similarity between them. For example, spectra from the same peptides have similar sets of strong peaks: in our data, 98.2% of the pairs of spectra from the same peptide had at least one peak in common in their respective sets of the five strongest peaks. However, only 5.5% of the pairs of spectra from different peptides also have such a match in their top 5 peaks. Since testing for a common peak in the list of top 5 peaks can be done much quicker than a complete similarity computation, this heuristic can account for a significant reduction in running time by quickly eliminating the majority of the unnecessary similarity computations.

The second heuristic we use relies on the fact that our algorithm uses multiple rounds of cluster joining (with decreasing similarity thresholds τ). Instead of recomputing the similarity between pairs of consensus spectra at each round, we can carry over similarity results from one round to the next. Thus, if at a certain round a pair of clusters show extremely low similarity, we take note of this fact (by setting an appropriate indicator) and we do not examine that pair again in subsequent rounds. We use a simple bit vector to store the similarity indicators of all pairs of clusters, which for n spectra amounts to approximately $n \cdot (n-1)/2$ bits. Even when clustering large datasets (10

million spectra), the largest number of spectra simultaneously clustered is 60000, which requires 215 MB of memory to store the similarity indicators. Note that since the write operations to the bit vector always precede read operations to the same addresses, the vector does not need to be initialized at any time.

This filtration heuristic can very efficient. For example, 99.9% of pairs of spectra of the same peptide have a similarity above 0.25, while less than 1% of the pairs of spectra from different peptides have a similarity that exceeds that level. Since 0.25 is a very low threshold, we can safely assume that if a pair of clusters have a similarity below 0.25 between them, even if they have additional spectra added to them in subsequent rounds, the cluster similarity will still be way below the minimum threshold for joining clusters (in our experiments the value $\tau_{min} = 0.55$ was used).

Heuristics used		# Similarity Comparisons	(%)	Total Run time (s)	(%)
Carry Similarity	Match in Top 5				
–	–	1.89×10^9	(100.0%)	8835	(100.0%)
+	–	4.71×10^8	(24.9%)	3731	(42.2%)
–	+	5.12×10^8	(27.1%)	4009	(45.4%)
+	+	2.26×10^8	(11.9%)	2766	(31.3%)

Table 1: The algorithms performance with different combinations of heuristics. The clustering algorithm was run on 0.8M spectra from the Human to evaluate the effect of adding the heuristics of carrying similarity results between the algorithm rounds and requiring pairs of spectra to have a match in their top 5 peaks. The algorithm's performance was measured both in the total number of computations performed and the total running time.

Table 1 shows the performance of the algorithm while applying different combinations of the heuristics mentioned above. The algorithm was run with a $r = 3$ rounds, a minimal similarity threshold $\tau_{min} = 0.55$, and using 15 peaks per 1000 Da for similarity computations. On their own, each of the heuristics approximately halved the number of similarity computations that were performed. Carrying similarity results between rounds reduced the number of these computations to 24.9% of the number of computations without heuristics, and requiring spectra to have a match in their sets of top 5 peaks reduced the number of computations to 27.1%. These two heuristics are rather complimentary to each other. The filter that requires a match of a peak in the top 5 is most effective in the algorithm's first round (in which most of the similarity computations are performed). The carrying over similarity results between is naturally only applicable to subsequent rounds. Thus when these two heuristics are combined they produce a significant reduction in the number of similarity computations that are carried out to 11.9% of the number of computations performed when no heuristics are used. Note that calculating the similarities between all pairs of spectra in each mass bin amounts to 1.25×10^9 similarity computations.

The reduction in running time is also quite impressive, using both heuristics reduces the running time less than a third of the time it takes without employing heuristics. It is worth noting that the clustering results with and without heuristics are very similar. For instance, without heuristics 71.4% of the spectra fell into non-singleton clusters compared to 70.8% when both heuristics were used.

4 Using Clustering to Focus Efforts On Interesting Spectra

In typical large-scale MS/MS experiments only 10%-20% of the spectra get identified. When these datasets are clustered, the number of spectra is reduced tenfold but the majority of these clusters do not get identified in the database search. Though many clusters can correspond to unidentifiable peptides (for instance spectra with very poor fragmentation patterns), these clusters can also belong to peptides with mutations/PTMs or alternative splice variants. Below we describe a process in which we use clustering to isolate from a large dataset of 14.5 million spectra a relatively small group of unexplained spectra that are good candidates for further investigation. This set of spectra is then processed using spectral networks [11] to obtain additional peptide identifications and further reduce the number of unidentified spectra that are left to be investigated. Table 2 summarizes the steps taken in this process.

Analysis Stage	# Explained Clusters	# Clusters That Remain Unexplained
Initial Dataset	-	14.5 M
After Clustering	-	1.29 M
Identified by InsPecT	278914	1.02 M
Identified by MS-Alignment	85430	935779
Removal of unidentified singletons	-	190091
Identified by alignment to annotated spectral network components	28915	**161176**

Table 2: Reducing number of unexplained clusters. The table describes different steps used to isolate a small subset of "interesting" unexplained spectra (clusters) from a large 14.5 million spectra dataset.

We started off with the complete Shewanella dataset of 14.5 million spectra that has been recently analyzed [12]. Clustering this data resulted 1.29 M clusters (of which 848418 were singletions), over a tenfold reduction compared to the original dataset size. Following that we used InsPecT to perform a database search of the clusters against a six frame translation of the Shewanella genome, which confidently assigned non-modified peptides to 278914 of the clusters (false discovery rate of 5% at the peptide level). These identified clusters mapped back to 2.97 million of the spectra in the original 14.5M dataset (20.5% of the spectra in the dataset), compared to 1.4 million spectra that were identified without clustering [12]. The clustered search identified 41220 peptides in the forward database, of which 94% were mapped back to known annotated proteins.

At this stage we were left with 1.02 million clusters that evaded identification by MS/MS database search, a 14-fold reduction in the number of spectra needed to be searched. We proceeded to run a "blind" MS-Alignment search of these clusters, which led to the identification of additional 85430 clusters which could be mapped to 10048 modified peptides (from the list of 10758 putative modifications identified in ref [12]). We remained with 935779 clusters of spectra that were not identified in MS/MS database searches.

We continued the analysis using spectral networks [11]. First we removed 745688 unidentified singleton clusters, and were left with 189988 unidentified clusters (along with an additional 364344 identified clusters). We chose to remove the singletons because of their large number and the fact that they are less likely to be spectra with a strong signal-to-noise ratio. The spectral network graph was created by representing each cluster as a graph vertex. The graph's edges were created by connecting all pairs of vertices that displayed statistically significant alignments between their peaks using a single arbitrary mass shift, which typically corresponded to a mass difference of up

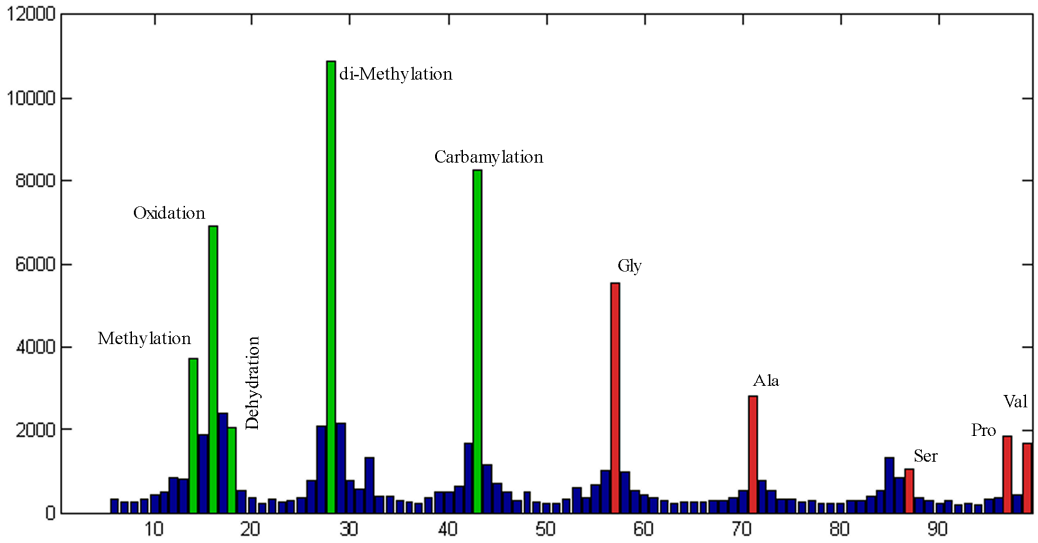

Figure 2: Histogram of absolute precursor mass differences for all detected spectral pairs on the clusters from the Shewanella dataset; the y axis represents the number of spectral pairs with a given difference in parent mass. For clarity, we only show the mass range $1 - 100$ Da.

to two amino acids or a PTM. We then used these alignments to create connected components (spectrum "stars" which correspond to different variants of the same peptide: prefix, or suffix peptides, variants with PTMs, etc.).

After extracting connected components, an additional 28915 clusters, which previously did not have an annotation, could be annotated based on their membership in a connected component with at least one annotated cluster. 161176 clusters remained unidentified, of which 32004 clusters belonged to 1313 connected components containing only unidentified clusters. The remaining 129172 did not align to any other clusters.

The process described above demonstrates how clustering can help focus the analysis efforts when dealing with large datasets. Computationally intensive searches might prove to be intractable if performed on the entire body of unidentified spectra. However, using clustering we were able to reduce the number of spectra that needed to be examined to approximately one million. This reduction both allowed us to perform time-consuming "blind searches" in a reasonable time and perform analysis using spectral networks [11] (after removing the unidentified singleton clusters), which both added new peptide identifications and reduced the number of unassigned spectra that remained to be investigated to 161176.

Our spectral network analysis leads to a surprisingly large estimate of the number of spectra that remain *unidentified* even in advanced MS/MS database searches. One may ask a question whether a large number of uninterpreted spectra forming our spectral networks is simply an artifact of many spurious alignments between unrelated spectra. Figure 2 illustrates that it is not the case by presenting the histogram of all mass offsets represented in spectral networks. Since spectral alignment has no knowledge of biologically relevant modifications and masses of amino acids, the histogram should not have any peaks in case the spectral networks are formed by spurious alignments. The fact that the histogram has prominent peaks corresponding exactly to common modifications and masses of amino acids proves that the spectral networks indeed reveal many unidentified peptides. While Bandeira et al., 2007 [11] demonstrated that spectral networks enable

accurate peptide sequencing, de novo reconstruction of these peptides remains beyond the scope of this paper and will be described elsewhere.

Clustering in conjunction with spectral networks facilitate the creation of *spectral archives* that contain both identified and *unidentified* spectra to complement the existing spectral library approaches [1, 13, 10, 9, 8]. Clusters of spectra obtained from MS/MS datasets (both identified and unidentified) are used to create spectral networks that can be stored in spectral archives. The spectral archives aggregate results from many experiments, and possibly even contain results from experiments done with closely related organisms. As results from new experiments become available, they are aligned against the clusters in the existing spectral archives to gain additional identifications both to the new data (by aligning the new results to existing annotated clusters), and also to existing unidentified clusters in the spectral archive (by aligning them with annotated clusters from the new results). As the spectral archives grow with the addition of data from more and more experiments, the spectral networks may enable accurate de novo sequencing [14] of the unidentified spectra in the archive. Our spectral network approach, for the first time, allows one to estimate the number of peptides that remain *unidentified* in MS/MS searches. Since most edges in our spectral networks represent common modifications and amino acid masses (see Figure 2) we argue that the spectral network have a potential to reveal these peptides that evade even advanced MS/MS searches.

References

[1] S.E. Stein and D.R. Scott. Optimization and testing of mass spectral library search algorithms for compound identification. *J. Am. Soc. Mass. Spectrom.*, 5:859–866, 1994.

[2] X.K. Wan, I. Vidavsky, and M.L. Gross. Comparing similar spectra: from similarity index to spectral contrast angle. *J. Am. Soc. Mass. Spectrom.*, 13:85–88, 2002.

[3] D.L. Tabb, M.J. MacCoss, C.C. Wu, S.D. Anderson, and J.R. Yates, III. Similarity among tandem mass spectra from proteomic experiments: detection, significance, and utility. *Anal. Chem.*, 75:2470–2477, 2003.

[4] I. Beer, E. Barnea, T. Ziv, and A. Admon. Improving large-scale proteomics by clustering of mass spectrometry data. *Proteomics*, 4:950–60, 2004.

[5] D.L. Tabb, M.R. Thompson, G. Khalsa-Moyers, N.C. VerBerkmoes, and W.H. McDonald. MS2Grouper: Group assessment and synthetic replacement of duplicate proteomic tandem mass spectra. *J. Am. Soc. Mass Spec.*, 16:1250–1261, 2005.

[6] S.R. Ramakrishnan, R. Mao, A.A. Nakorchevskiy, J.T. Prince, W.S. Willard, W. Xu, E.M. Marcotte, and D.P. Miranker. A fast coarse filtering method for peptide identification by mass spectrometry. *Bioinformatics*, 22:1524–1531, 2006.

[7] J. Liu, A.W. Bell, J.J. Bergeron, C.M. Yanofsky, B. Carrillo, C.E. Beaudrie, and R.E. Kearney. Methods for peptide identification by spectral comparison. *Proteome Sci.*, 5:3, 2007.

[8] H. Lam, E.W. Deutsch, J.S. Eddes, J.K. Eng, S.E. King, N. Stein, and R. Aebersold. Development and validation of a spectral library searching method for peptide identification from ms/ms. *Proteomics*, 7:655–667, 2007.

[9] F.B. Frewen, G.E. Merrihew, C.C. Wu, W. Stafford Noble, and M.J. MacCoss. Analysis of peptide ms/ms spectra from large-scale proteomics experiments using spectrum libraries. *Anal. Chem.*, 78:5678 – 5684, 2006.

[10] R. Craig, J.C. Cortens, D. Fenyo, and R.C. Beavis. Using annotated peptide mass spectrum libraries for protein identification. *J. of Proteome Research*, 5:1843 –1849, 2006.

[11] N. Bandeira, D. Tsur, A. Frank, and P.A. Pevzner. Protein identification by spectral networks analysis. *PNAS*, 104:6140–6145, 2007.

[12] N. Gupta, S. Tanner, N. Jaitly, J. Adkins, M. Lipton, R. Edwards, M. Romine, A. Osterman, V. Bafna, R.D. Smith, and P. Pevzner. Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation. *Genome Res.*, 17:1362–1377, 2007.

[13] J.R. Yates, III, S.F. Morgan, P.R. Gatlin, C.L. amd Griffin, and J.K. Eng. Method to compare collision-induced dissociation spectra of peptides: Potential for library searching and subtractive analysis. *Anal. Chem.*, 70:3557–3565, 1998.

[14] N. Bandeira, K. Clauser, and P. Pevzner. Shotgun protein sequencing: Assembly of ms/ms spectra from mixtures of modified proteins. *Mol. Cell. Proteom.*, 6:1123–1134, 2007.