Peptide Identification from Mixture Tandem Mass Spectra

Jian Wang[1], Josué Pérez-Santiago[1], Jonathan E. Katz[2], Parag Mallick[2], Nuno Bandeira[3]

[1]Bioinformatics Program, University of California, San Diego, USA

[2]Dept. of Chemistry and Biochemistry, University of California, Los Angeles, USA

[3]Center for Computational Mass Spectrometry, University of California, San Diego, USA

Supplementary Materials

Data pre-processing and simulation of mixture spectra

We downloaded the NIST human dataset (*ver. 6/06*) from PeptideAtlas (http://www.peptideatlas.org/speclib/). Peak masses were rescaled by multiplying every mass by 0.9995 to better center these around integer values [1]. Every spectrum was then converted to a vector by binning peaks masses at integer values (i.e., bin width of 1, bins at $[m - 0.5, m + 0.5]$). Peaks that fall into the same bin are combined by adding their intensities. We also applied a square-root transform to all peak intensities to reduce the disproportionate influence of high-intensity peaks on spectral similarity. After this transformation, each vector was normalized to norm 1 by dividing each element in the vector by the vector's euclidian norm. A mixture spectrum is modeled as a linear combination of two single-peptide spectra: $M = A + \alpha B$, where M is a mixture spectrum, A and B are two single-peptide spectra and α is a predetermined mixture coefficient.

Windowed peak filtering

In order to reduce noise, we applied a windowed peak filter as follows: for each peak in the spectrum we scan neighboring peaks within $\pm W$ Daltons from the current peak and retain it if it has rank $\leq N$ (by peak intensity) in its neighborhood. This method is used to filter all spectra in the experimental dataset (W=50, N=15). This filter was also applied to simulated mixture spectra to benchmark scenarios where low intensity peaks from single-peptide spectra would be missing. The performance in the identification of mixture spectra for different parameters of N and W is shown in Table 1 and 2.

		Accuracy for different mixture coefficients			
Top peaks kept	Window size(Daltons)	$\alpha = 1.0$	$\alpha = 0.5$	$\alpha = 0.2$	$\alpha = 0.1$
5	25	98.5	98.4	95.3	79.9
10	25	98.6	98.5	96.1	90.5
15	25	98.7	98.5	96.9	91.3
5	50	98.7	98.2	86.9	52.4
10	50	98.7	98.3	94.9	81.7
15	50	98.7	98.4	95.2	88.5
5	100	97.2	95.3	63.7	20.7
10	100	98.5	97.8	88.5	41
15	100	98.5	98	93.6	75.2

Table 1: M-SPLIT accuracy in the identification of mixture spectra for different levels of missing MS/MS peaks.

		Accuracy for different mixture coefficients			
Top peaks kept	Window size(Daltons)	$\alpha = 1.0$	$\alpha = 0.5$	$\alpha = 0.2$	$\alpha = 0.1$
5	25	98.4	98.3	93.9	70
10	25	98.5	98.5	94.8	80.4
15	25	98.8	98.4	94.5	79.4
5	50	98.9	98.1	84.5	43.1
10	50	98.5	98.3	93.4	69.8
15	50	98.5	98.5	94.2	77.1
5	100	97.3	95.5	61.5	18.5
10	100	98.6	98	85.7	45.6
15	100	98.7	98.4	91.6	63.3

Table 2: Iterative-approach accuracy in the identification of mixture spectra for different levels of missing MS/MS peaks.

Filtering out low-complexity spectral matches

During our experiments, we observed that low-complexity spectra (those that are dominated by a few peaks), can have articially high-cosine match to the library. In order to filter out these cases, we used a measure similar to F-score used in [2]. We consctructed two datasets, the true dataset that contain true matches and a negative control dataset which contain errorneous matches including those due to low-complexity of the spectra. The negative control dataset come from cases where the precursor m/z difference between the query spectrum and the top-matched spectrum is larger than tolerance. Then we trained a SVM (gaussian kernel, with bandwidth equal to five) using three feature: 1) cosine, 2) dot-bias (see [2]) and 3)projected-cosine to separate the two datasets. For each top matches return by MSPLIT, we computed the SVM score, if it is lower than 70 we discarded such match, since it suggets that the high cosine similarity may be due to

fortuitiou match of a few dominant peaks.

Estimation of mixture coefficients $\hat{\alpha}$

Optimal-cosine method:

Define:

$$f(\alpha) = \frac{M \cdot (A + \alpha B)}{\|M\| \, \|A + \alpha B\|} \tag{1}$$

as the cosine similarity between M and $A + \alpha B$, where M is a putative mixture spectrum and A and B are two candidates from the spectral library. We can rewrite the equation as:

$$f(\alpha) = \frac{C + \alpha G}{\|M\| \sqrt{E + 2\alpha D + \alpha^2 F}} \tag{2}$$

where $C = M \cdot A$, $D = A \cdot B$, $E = A \cdot A$, $F = B \cdot B$ and $G = M \cdot B$. We asssumed that the correct α maximize the similarity $f(\alpha)$, thus taking the derivative with respect to α and making it zero we get:

$$f'(\alpha) = \frac{\sqrt{E + 2\alpha D + \alpha^2 F}(G) - (C + \alpha G)(D + \alpha F)(E + 2\alpha D + \alpha^2 F)^{-1/2}}{\|M\| \, (E + 2\alpha D + \alpha^2 F)} \tag{3}$$

$$0 = (\sqrt{E + 2\hat{\alpha} D + \hat{\alpha}^2 F})(G) - \frac{(C + \hat{\alpha} G)(D + \hat{\alpha} F)}{\sqrt{E + 2\hat{\alpha} D + \hat{\alpha}^2 F}} \tag{4}$$

$$0 = (E + 2\hat{\alpha} D + \hat{\alpha}^2 F)(G) - (C + \hat{\alpha} G)(D + \hat{\alpha} F) \tag{5}$$

$$0 = EG + 2\hat{\alpha} DG + \hat{\alpha}^2 FG - CD - \hat{\alpha} FC - \hat{\alpha} DG - \hat{\alpha}^2 FG \tag{6}$$

$$\hat{\alpha} = \frac{EG - CD}{FC - DG} \tag{7}$$

To assure that $\hat{\alpha}$ is the maximum we have to verify that the second derivative is negative. For this, let us consider the function,

$$f(\alpha) = \frac{M \cdot (A + \alpha B)}{\|M\| \, \|A + \alpha B\|} = cos(\phi) \tag{8}$$

where ϕ is the angle between the vectors M and $A + \alpha B$. The first derivative of this function would be $sin(\phi)$ and the second derivative would be $-cos(\phi)$ which is always negative in the domain (0, π) making α our maximum.

Residual-spectrum method:

Recall in our model, a mixture spectrum M can be considered as a linear combination of two single-peptide spectra: $M = X + \alpha Y$. Here X and Y represents the two unknown single-peptide spectra that give rise to the mixture spectrum we try to identify. Since all putative mixture spectra are normalized to norm one before search we have:

$$M^{'} = \frac{M}{\|M\|} = \frac{X + \alpha Y}{\|M\|} \tag{9}$$

$$= \frac{X}{\|M\|} + \frac{\alpha Y}{\|M\|} \tag{10}$$

Let us assume by searching the query spectrum $M^{'}$ against the library we identified A as the first component of the mixture, thus A is roughly equivalent to X. We can then estimate first component in equation (9) by computing the shared peaks between $M^{'}$ and A, which is equivalent to the projection of $M^{'}$ on A (see Method section).

$$\frac{X}{\|M\|} \approx M^{'}_{p(A)} \tag{11}$$

Then a residual spectrum R can be used to approximate the second term in equation (10)

$$R = M^{'} - M^{'}_{p(A)} \approx \frac{\alpha Y}{\|M\|} \tag{12}$$

$$\|R\| \approx \frac{\alpha \|Y\|}{\|M\|} \tag{13}$$

$$\alpha = \frac{\|R\| \, \|M\|}{\|Y\|} \tag{14}$$

$$\alpha = \|R\| \sqrt{1 + \alpha^2} \tag{15}$$

The last step comes from the fact that Y is norm one and we approximate the magnitue of M as $\sqrt{1 + \alpha^2}$ by assuming that $X \cdot Y$ is zero. Solving equation(15) in term of α we arrived at the final solution:

$$\alpha = \frac{\|R\|^2}{1 - \|R\|^2} \tag{16}$$

Upper Bound for Searching - extension to $\alpha \neq 1$

Note that the key to the search-and-bound method is developing an upperbound for the cosine function, so we can eliminate any candidates that will not score higher than current best estimate. In the method section we show the bound for the simple case when α=1. Here we extend the bound to case when α is not one, our objective function now becomes:

$$\cos(M, A + \alpha B) = \frac{M \cdot (A + \alpha B)}{\|M\| \, \|A + \alpha B\|} \tag{17}$$

$$= \frac{M \cdot (A + \alpha B)}{\|M\| \sqrt{A \cdot A + 2\alpha B \cdot A + \alpha^2 B \cdot B}} \tag{18}$$

$$\leq \frac{M \cdot (A + \alpha B)}{\sqrt{1 + \alpha^2}} \tag{19}$$

Notice we cannot use equation (19) as an upperbound like the case when α=1, since α is not a fixed value. However, we know that there is one alpha that maximize equation (19), hence we have:

$$f(\alpha) = \frac{M \cdot (A + \alpha B)}{\sqrt{1 + \alpha^2}} \tag{20}$$

$$f'(\alpha) = \frac{(M \cdot A + \alpha M \cdot B)(-1/2)(2\alpha) + M \cdot B(1 + \alpha^2)}{(1 + \alpha^2)^{3/2}} = 0 \tag{21}$$

$$f'(\alpha) = \frac{-M \cdot A\alpha + M \cdot B}{(1 + \alpha^2)^{3/2}} = 0 \tag{22}$$

$$\alpha^* = \frac{M \cdot B}{M \cdot A} \tag{23}$$

Checking 2nd derivative: (24)

$$f''(\alpha) = \frac{M \cdot A(2\alpha^2 - 1) - 3\alpha M \cdot B}{(1 + \alpha^2)^{5/2}} \tag{25}$$

$$f''(\alpha^*) = \frac{-(M \cdot B)^2}{M \cdot A} - M \cdot A \leq 0 \tag{26}$$

Substituting (23) into (19) we have (27)

$$\cos(M, A + \alpha B) \leq \frac{(M \cdot A)^2 + (M \cdot B)^2}{\sqrt{(M \cdot A)^2 + (M \cdot B)^2}} \tag{28}$$

Notice the upper bound in above holds true for any value of α, this enable us to use it as a bound in general.

Extension to square-root transformed spectrum

The bound in last section should work for any vectors, therefore it should also work for square-root transformed spectrum. Again, let $\sqrt{X}$ stands for element-wise square root of vector$X$:

$$\sqrt{X} = (\sqrt{x_1}, \sqrt{x_2}, \ \dots \ , \sqrt{x_n})$$

and assume $\sqrt{M}$, $\sqrt{A}$, $\sqrt{B}$ are all unit vectors, equation (19) should become:

$$\cos(\sqrt{M}, \sqrt{A} + \alpha\sqrt{B}) \quad = \frac{\sqrt{M} \cdot (\sqrt{A} + \alpha\sqrt{B})}{\left\|\sqrt{M}\right\| \left\|\sqrt{A} + \alpha\sqrt{B}\right\|} \tag{29}$$

$$\leq \frac{\sqrt{M} \cdot (\sqrt{A} + \alpha\sqrt{B})}{\sqrt{1+\alpha^2}} \tag{30}$$

However equation(29) is not the objective function that we are solving. We need to maximize the following function:

$$\cos(\sqrt{M}, \sqrt{A + \alpha' B}) \quad = \frac{\sqrt{M} \cdot \sqrt{(A + \alpha' B)}}{\left\|\sqrt{M}\right\| \left\|\sqrt{A + \alpha' B}\right\|} \tag{31}$$

$$<= \frac{\sqrt{M} \cdot \sqrt{(A + \alpha' B)}}{\sqrt{1+\alpha'}} \tag{32}$$

The last step comes from the assumption that $\sqrt{M}$, $\sqrt{A}$, $\sqrt{B}$ all have norm one. Lastly we appeal to the inequality that:

$$\sqrt{M} \cdot \sqrt{(A + \alpha' B)} \leq \sqrt{M} \cdot (\sqrt{A} + \alpha\sqrt{B}) \tag{33}$$

$$when\ setting\ \alpha^2 = \alpha' \tag{34}$$

we have the upper bound:

$$\cos(\sqrt{M} \cdot \sqrt{(A + \alpha' B)}) \leq \frac{\sqrt{M} \cdot (\sqrt{A} + \alpha\sqrt{B})}{\sqrt{1+\alpha^2}} \tag{35}$$

Thus the upper bound in previous section(eq. 28) can be still applied, just we have to applied it to the square-root transformed spectrum. Note that this procedure can be applied on any list of spectra. Thus it can be combined with any filter that reduce the size of library. In our studies we found this strategy works

very effectively when combined with the projected-cosine filter, which reduce the search space from 10^8 to a few hundreds pairs of spectra.

References

[1] S. Kim, N. Gupta, N. Bandeira, and P.A. Pevzner. Spectral dictionaries: Integrating de novo peptide sequencing with database search of tandem mass spectra. *Molecular & Cellular Proteomics*, 8(1):53, 2009.

[2] H. Lam, E.W. Deutsch, J.S. Eddes, J.K. Eng, N. King, S.E. Stein, and R. Aebersold. Development and validation of a spectral library searching method for peptide identification from MS/MS. *Proteomics*, 7(5):655–667, 2007.

Comparison of cosine and projected-cosine as filters

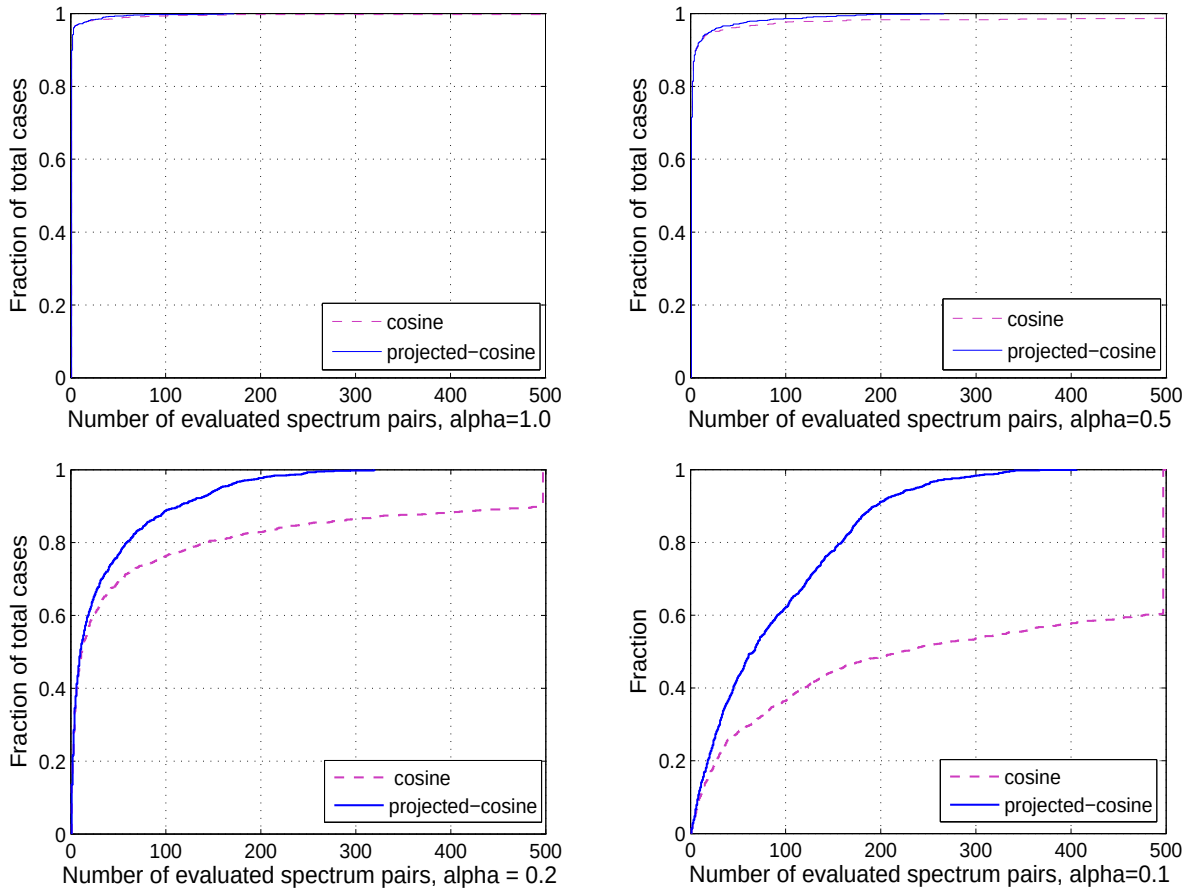
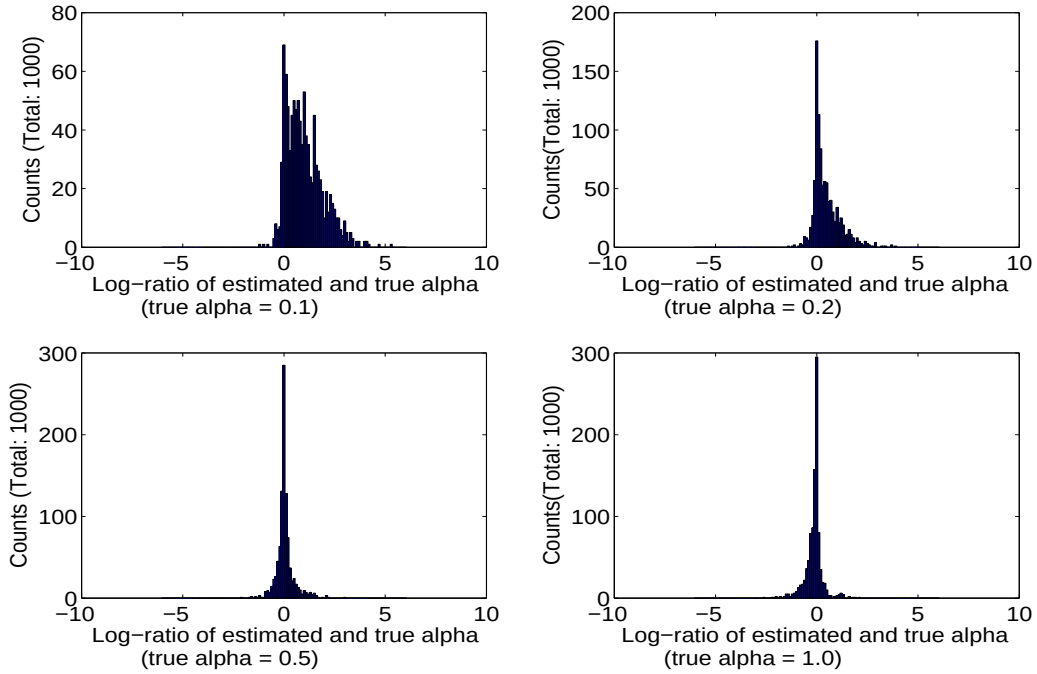


Figure 1: Comparing cosine and projected-cosine as filters. The library was prefilter with both cosine and projected-cosine, to keep the top 500 candidates. Then the branch-and-bound strategy was applied to search for the optimal pairs. The number of pairs of spectra considered before M-SPLIT terminates and find the optimal solution were counted. It is observed that projected-cosine is a more effective filter when combined with branch-and-bound strategy. It result in less numbers of pair of spectra that needed to be consider as compare to that of using cosine as a filter.

Quantification of relative abundance of peptides

a)Residual spectrum method



b)Optimal cosine method

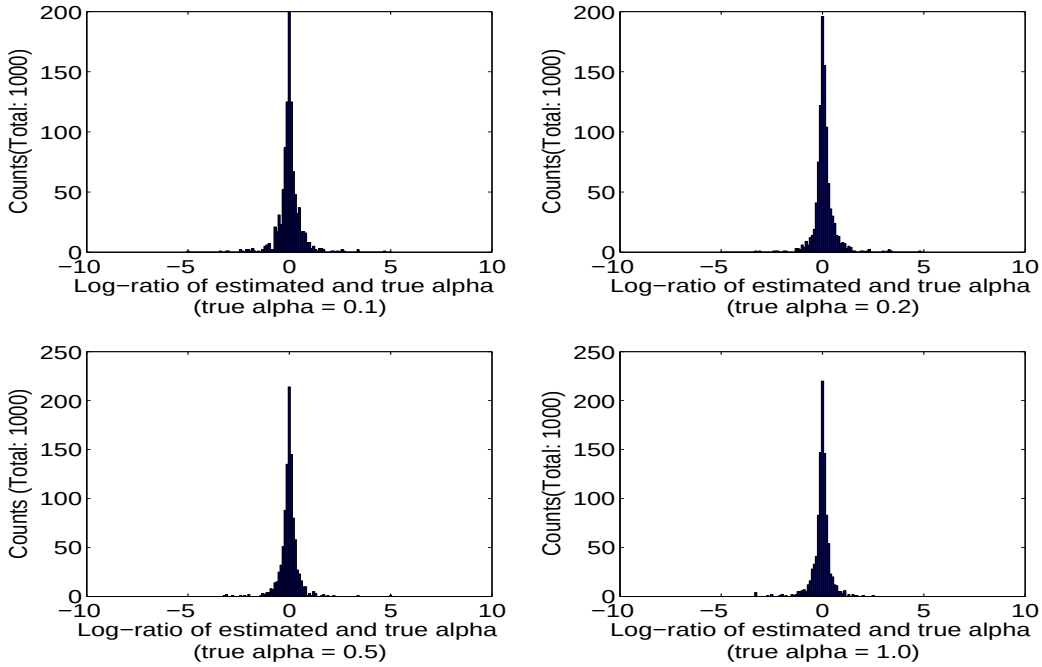


Figure 2: Log-2 ratios of estimated ($\hat{\alpha}$) and true (α) mixture coefficient when a)using the residual spectrum method and b)maximizing the cosine similarity between the query spectrum and the two library candidates Note for the case when $\alpha = 1$ both peptide has equal abundance. Due to error in estimation, it is possible for $\hat{\alpha}$ to be greater than one in some cases. Since α is a relative value, in M-SPLIT we take inverse of estimated α whenever it is greater than one to make it consistent with our definition that $0 \leq \alpha \leq 1$