An underappreciated challenge in identifying metabolites: scoring matches between library spectra and LC-MS HRAM metabolite spectra from complex samples

Lewis Y. Geer, Yamil Simón-Manso, Xiaoyu Yang, Stephen E. Stein

NIST National Institute of Standards and Technology U.S. Department of Commerce | Mass Spectrometry Data Center

OBJECTIVE

- Improve the scoring of HRAM tandem spectral search hit lists
- A standard measure of search engine hit list relevance is the discounted cumulative gain:

$$g_p = \sum_{i=1}^{p} \frac{2^{T_i} - 1}{log_2(i+1)}$$

where p is the length of the hit list, i is a position in the hit list, and T_i is the relevance of the hit at position i. A good score will increase g_p

- Relevance in this study is defined as chemical similarity as measured by the Tanimoto score between the chemical structures of the query and hit spectra
- The Tanimoto score is the normalized fraction of the number of structural moieties shared by two chemical structures

SCORES

- The gold standard is Tanimoto scoring from a structural similarity search, which can only be performed if the structure of the query is known
- Another benchmark is the widely used NIST match factor, which is based upon the cosine similarity as described in Stein and Scott, 1994, DOI: 10.1016/1044-0305(94)87009-8
- The AI embedding score used machine learning to evaluate spectral matches
  - A deep, 1D convolutional residual network was created. It generates a 1000 element vector for each spectrum
  - This network is trained on pairs of spectra so that the dot product of the vectors generated from each pair of spectra is used to predict the Tanimoto score between the chemical structures that corresponding to the spectral pair
- A probability-based score, similar to that used in Geer et al., 2004, DOI: 10.1021/pr0499491
  - Scores using probability-based statistics have an advantage in that there are many avenues for improving the score, such are corrections for correlation

SCORES, continued

- Our probability-based score consists of two parts, intensity and match statistics
  - The intensity statistic correlates the intensity ranks between two spectra using p-values for the Spearman's rank correlation coefficient
  - The match statistic uses a sum of Poisson distributions, one for every 0.01 Dalton bin around each peak in the query product spectrum, with a probability based on the chance of a random match to a peak in any library spectrum
  - The null hypothesis is a match to a random spectrum that is as good as or better than the hit, and modeled by the survival function of the Poisson

TRAINING AND RESULTS

- To eliminate background noise as a factor in this evaluation, we used the NIST2020 HRAM tandem spectral library for both searching and querying
- We used MH1+ HCD spectra with a normalized collision energy of 35
- 332 of these 32 728 spectra were reserved as a test set used to calculate g_p
- Matches to identical spectra were removed from the hit lists
- To increase the number of possible hits to related structures, the searches did not filter on precursor m/z

	g_3	g_{10}
Structural similarity	0.70	1.07
NIST match factor	0.22	0.31
Probability score	0.22	0.31
Embedding score	0.27	0.53

CONCLUSIONS

- The probability-based score reproduces the discounted cumulative gain of the NIST match factor, opening additional avenues of improvement
- The AI embedding score has better results than the other two scores, although the comparison is not direct as this algorithm does not require the hit and query spectra to be aligned in m/z space

chemdata.nist.gov