S. Hudson and S. Maynard

Unblocking the data dam

HFL Ltd. Newmarket Road, Fordham, Cambridgeshire, UK

Overview

In the late 1980s, the utilisation of GC-MS as a front line screening technique became a reality in many anti doping laboratories. One of the major benefits of GC-MS was the selectivity derived from the mass spectral data. However, the combination of chromatographic information plus mass spectra means that the data generated can be complex, particularly when full scan mass spectral acquisition is used. Interpretation of the data has largely been a combination of expert manual inspection plus automated data processing and presentation. Mass spectral databases or 'libraries' have been compiled to help automate the analysis of the data.

Since the early 1990's HFL has developed a system that combines automated library searching and extracted ion chromatograms to present the data for final assessment by expert analysts. During this period more mass spectrometry tests have been developed for use in doping control screening protocols. This has resulted in an explosion of mass spectral data that requires assessment, particularly in animal sports testing where the range of compounds tested for is greater than in human sports testing. Consequently, the amount of mass spectral data that needs to be assessed for each doping control sample is more complex than ever and takes much more time for the human element of the process. The nature of sports testing also means that most of the data will also be negative for the presence of prohibited substances. The challenge is therefore to find a means to process the data faster while ensuring the accuracy of the results. This will improve laboratory efficiency and control the sample processing time as more screening tests are developed.

Improvements

Automated library searching routines have seen many enhancements including:-

- The use of extracted ion chromatograms to generate peaks for library searching
- The use of retention time information and other filters to refine the output from library searching.
- The use of spectral deconvolution as opposed to the more traditional background subtraction to improve the quality of spectra submitted for library searching e.g AMDIS
- Different library search algorithms – INCOS, PBM, NIST

These approaches have significantly improved the performance of library searching whilst reducing the data output requiring expert review. For example, prior to the application of retention time and quality filters, a typical Hewlett Packard Chemstation report might be as long as 15-20 pages for a single full scan GC-MS test.

Problems with current approaches

Even with these improvements each data file report is still reviewed by an expert analyst. For a laboratory that processes thousands of samples a year, with each sample subjected to multiple GC-MS and/or LC-MS analyses, the time required for expert review of the data represents a significant element in the cost of analysis. HFL currently processes around 30,000 sports drug testing samples per year. A typical sample will have 3 full scan GC-MS analyses, resulting in a total of 90,000 data set reviews each year. At the current degree of data complexity this requires a full time commitment from 3 to 4 analysts just to review this data. If current trends in method development continue this is likely to increase and is a major investment in both staff operating time and training.

Additionally, there are concerns that the quality of expert review may become compromised under an increasingly complex and heavy workload which in turn increases the risk of false negative results.

The algorithms currently used at HFL for automated library searching consider a single file in isolation. One consequence of this system is that the expert analyst is required to determine if

a common library match present in several samples in a batch is truly a prohibited substance or something that should be considered a normal component present in the majority of samples.

If the number of samples reported to contain a prohibited substance is considered to be 2% then the vast majority of samples are negative. In fact, it is likely that over 99% of tests conducted have a negative conclusion. A large amount of expert analyst time is spent to arrive at that conclusion.

Proposed Solution

HFL is working with BlueGnome Ltd. to develop a data analysis tool that has the potential to radically change the way that GC-MS and LC-MS data is processed. The project target is to develop a system that will be able to eliminate 85% of the truly negative data without human intervention. In addition the system should also help control the false negative rate by reducing the workload for the experienced analysts and presenting them with better quality data.

The data analysis tool itself is based on a Bayesian statistical framework which uses a combination of the information available from mass spectral libraries and from the expert analyst to emulate the decision making process used by that expert human operator.
The development of the system has been based upon the observation of the data review process and also detailed feedback about how the analyst determines whether or not a compound is truly present.

The system utilizes existing mass spectral databases plus any retention time information that is available. It also uses additional 'expert' information as follows

- Diagnostic ions for certain compounds may be used to aid detection and to reduce inaccurate library search results. These ions receive additional weighting in data processing making them more significant in the matching process than other ions. For example, an incorrect library search hit is commonly generated for fenfluramine by the existing data processing system, requiring expert interpretation to eliminate the need for further work.
- Data from a batch of samples can be used to eliminate incorrect library identifications due to endogenous compounds. A consideration of similar hits in other samples in the

batch would lend weight to the probability that this particular hit is a common peak that does not require further investigation.

- Attempts to overcome the effects on library searching of spectral skewing due to low scan speed or differences in instrument tune will improve the operation of the library matching process. This is achieved through the use of a function that considers that ratios between diagnostic ions of a similar m/z are less likely to be distorted (relative to the library) than ratios at opposite ends of the spectrum. This function is used to assist in the correct identification of spectra using the library.
- Algorithms are used to reduce background mass spectral noise without removing diagnostic ions. The algorithms are based on existing BlueGnome software that reduces noise in data generated from protein micro arrays.

How the proposed system works to detect the presence of a drug

- Peak detection starts with a 'reverse' type search utilizing the spectral matching algorithms.
- A score is assigned for each mass spectrum in the data against every compound in the mass spectral library.
- Noise reduction is also employed in this process.
- A compound list is generated and then combined with an array of other metrics in the Bayesian framework to make an overall judgment of whether the hit is 'true' or 'false'.
- The final output is a probability score of whether or not the sample contains a true positive hit.
- The identity of the substance is also indicated.

Future

One of the aims for future development of this software is to identify a mass spectrum that is unique to a particular sample even if a library comparison is not available. It is hoped that this will play a part in identifying modified or 'designer' substances that might be unknown at the time of analysis.

References

Hudson, S. and Maynard, S. (2002) Automated processing of GCMS data. Proc. 14th Int. Conf . racing Anal. Vet. Eds: D.W. Hill and W.T. Hill. R&W Publications, Newmarket pp163-168