USE R FOR PEPTIDE MASS FINGER PRINTING.

E. W. Wolski[1 2], T. Kreitler[1], H. Lehrach[1], J. Gobom[1], K. Reinert[2]

Abstract. We use R in order to generate a greater specificity and sensitivity of protein identification by Peptide Mass Fingerprinting (PMF). This is achieved through analysis, detection and correction of measurement errors, by filtering of MS data prior to database searches and by analysis of the search results. These functions are implemented as an add-on packages to the freely-available statistical software, R.

1. Introduction

Protein identification using PMF data is performed by comparing experimentally determined peptide masses of a protein cleaved by a specific protease to in silico generated PMFs of known protein sequences[4]. Mass lists are generated by assigning m/z values to each monoisotopic signal in the mass spectra. The mass lists are then used for protein identification by search engines.

Using R[2] we

- analyze the measurement error and calibrate the masses
- find contaminants and remove them
- submit peak-lists to the identification software Mascot[5], and analyse the search result.

For this tasks we use functionality provided by R e.g.: spline functions (`R/modreg`), linear regression, functions for matrix computation, descriptive statistic functions (`R/base`) and clustering algorithms (`R/mva`). In addition we implemented functions for pairwise protein sequence comparison in the package `R/pairseqsim` and for communication with web servers in the package `R/httpRequest`.

2. R Packages for Peptide Mass Fingerprinting

R/msbase. The m/z and intensity value pairs assigned to each peak in a spectrum are stored along with the position of the sample on the MALDI sample support in the `Massvector` object. State of the art mass spectrometers can analyze several hundred samples on a single sample support. The resulting collection of `Massvectors` is modeled by the class `Massvectorlist`. The class `Massvector` provides methods for peak-list manipulation e.g. fuzzy union or fuzzy intersect. They have to be fuzzy because the masses have an measurement error. Also methods to compute distance and similarity measures on peak intensities e.g. correlation, spectral angle, similarity index, or binary measures like relative mutual information and many more, are implemented.

[1] Max Planck Institute for Molecular Genetics.

[2] Free University of Berlin, Institute of Informatics.

R/mscalib. adds methods for calibration and filtering to the `Massvector` and `Massvectorlist` classes. The masses can be calibrated by internal, external[1], pre-[6] and set based internal calibration. The data can be analyzed and filtered for abundant masses, chemical noise and significant mass differences.

R/msmascot. The calibrated and filtered peak-list can be submitted to the Mascot search software out of R. The package provides classes to store and visualize the search results. For example the class `Mascotsearch`, which stores the result of a single search, implements a plot function which visualizes the scores, number of matched peptides and the sequence coverages of fetched hits. To help to interpret the identification result, in cases when one peak-list matches multiple proteins, the summary method of the `Mascotsearch` class clusters the protein sequences and the theoretical digest. The search result obtained by submitting a `Massvectorlist` for a search is modeled by the class `MascotsearchList` which implements methods which generate summaries of e.g. the number of identification.

3. Summary

The *R/PMF* packages provide a variety of visualization, summary, analysis, filtering, and calibration methods. The combination of different calibration and filtering methods can significantly improve protein identification. Combining the implemented functionality with *Sweave*[3] enables to generate experiment and protein identification reports in high throughput. Further prospects for extension of the *R/PMF* packages include functions for peak picking in mass spectra.

References

1. J. Gobom, M. Mueller, V. Egelhofer, D. Theiss, H. Lehrach, and E. Nordhoff, *A calibration method that simplifies and improves accurate determination of peptide molecular masses by maldi-tof-ms*, Analytical Chemistry **74** (2002), no. 8, 3915–3923.
2. Ross Ihaka and Robert Gentleman, *R: A language for data analysis and graphics*, Journal of Computational and Graphical Statistics **5** (1996), no. 3, 299–314.
3. Friedrich Leisch, *Sweave and beyond: Computations on text documents*, Proceedings of the 3rd International Workshop on Distributed Statistical Computing (DSC 2003), 2003.
4. D. J. C. Pappin, P. Hojrup, and A. J. Bleasby, *Rapid identification of proteins by peptide-mass fingerprinting*, Curr. Biol. **3** (1993), 327–332.
5. D. N. Perkins, D. J. Pappin, D. M. Creasy, and J. S. Cottrell, *Probability-based protein identification by searching sequence databases using mass spectrometry data*, Electrophoresis **20** (1999), no. 18, 3551–3567.
6. A. Wool and Z. Smilansky, *Precalibration of matrix-assisted laser desorption/ionization-time of flight spectra for peptide mass fingerprinting*, Proteomics **2** (2002), no. 10, 1365–1373.