Master Degree Project:

Simultaneous Identification of Species and Tissue Composition in Unknown Samples

Summary

We are looking for a student interested in mass spectrometry-based proteomics and bioinformatics, for example from the **Master Programme in Molecular Biotechnology Engineering (X)**, to take part in an exciting 6-month research project using libraries of tandem mass spectra to identify both the biological species and the organ/tissue composition of unknown and mixed samples, such as food products or animal feed. The student will be based at BMC (Department of Analytical Chemistry) and jointly supervised by Professor Jonas Bergquist, Analytical Chemistry, Uppsala University, and Magnus Palmblad, Associate Professor and Head of Bioinformatics at the Center for Proteomics and Metabolomics of the Leiden University Medical Center in Leiden, the Netherlands. The start date is flexible, but no later than January 15, 2018. The student should have excellent laboratory skills and experience with, or at least the willingness to learn, PHP scripting and GPU computing.

For more information, contact Magnus Palmblad: n.m.palmblad@lumc.nl.

Background

Fifty-five years ago, Zuckerkandl, Jones and Pauling unlocked the field of molecular phylogenetics by comparing the patterns formed by hemoglobin peptides from different primates by two-dimensional electrophoresis and chromatography on paper (1). As large quantities of nucleic acid sequence information - billions of base pair reads – are now obtained even from minute or fossil samples, molecular phylogenetics is now mostly based on DNA. Direct sequencing or characterization of proteins, for example by mass spectrometry, is used in special cases, such as the rapid identification of bacteria (2). Any method used in molecular phylogenetics can in principle be applied to identify the biological origin of an unknown sample, provided the right references. A good example is the Barcode of Life Data System, BOLD (3).

In 2012, we published the first direct comparison of large numbers of tandem mass spectra for molecular phylogenetics (4). This was followed in 2013 by collecting reference material for a number of food species, building libraries of tandem mass spectra, and comparing these with similarly acquired data for the unknown (5). In 2015, we launched a web interface to a species identification engine with applications to human pathogens and food species identification. In 2016, we extended the range of the food species identification to cover mammalian and avian species in a collaboration between the universities in Uppsala and Leiden, and the Technical University of Denmark (6). This work also demonstrated how to determine the relative composition of mixed-species products using spectral counting. In recently published work we applied the identification method to bacterial strains (7) and aquafeed (8), and compared intra- and interlaboratory reproducibility and the use of different types of mass spectrometers (9).

The Project

For confident species identification of an unknown sample, it is in most (but not all) cases necessary to have a reference sequence or mass spectrometry dataset for that species. This also holds for tissue or cell type identification. In theory, it is *not* necessary to collect reference material from all tissues from every species – an exercise that would quickly grow into the tens or hundreds of thousands of references. This is the key principle for concurrent species

and tissue identification: we only need to find the best matching tissue for a comparable species along the tissue axis and the best matching species along a single-tissue species axis (Figure 1).

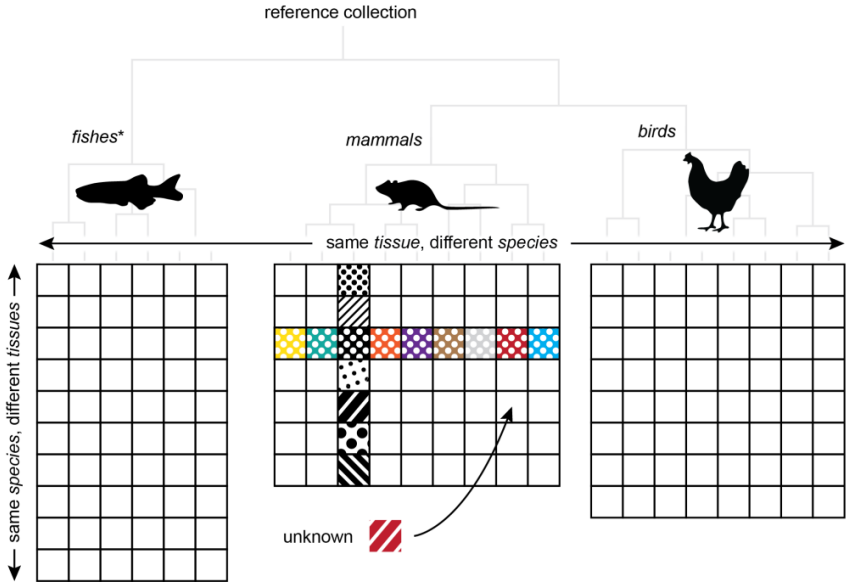

Figure 1. The key idea of simultaneous species- and tissue-typing is the use of two orthogonal reference axes. One axis is constructed from the same tissue across all species and the other from all tissues for a single species (here simplified to show 7 tissues from 9 mammalian species). The unknown species is identified as the best matching species, even if this is from a different tissue, and the unknown tissue is identified as the best matching tissue, even if this is from a different species. Tissue references will be collected for at least one (ray-finned) fish, one mammal and one bird species. *"Fish" is paraphyletic but here refers exclusively to ray-finned fish.

The student should collect and prepare tissue samples from a wide range of species, optimize the data acquisition and build an open resource for biological tissue and species identification. Statistical methods will be used to estimate the confidence level in the tissue and species assignment, including confidence levels at higher taxonomical ranks. This will be done by integrating existing models for peptide observability (10) with a simulation of the experimental workflow, from sample to spectral counting (Figure 2). To speed up the spectral matching, GPU-computing may be used, pre-loading the reference libraries in the GPU memory.

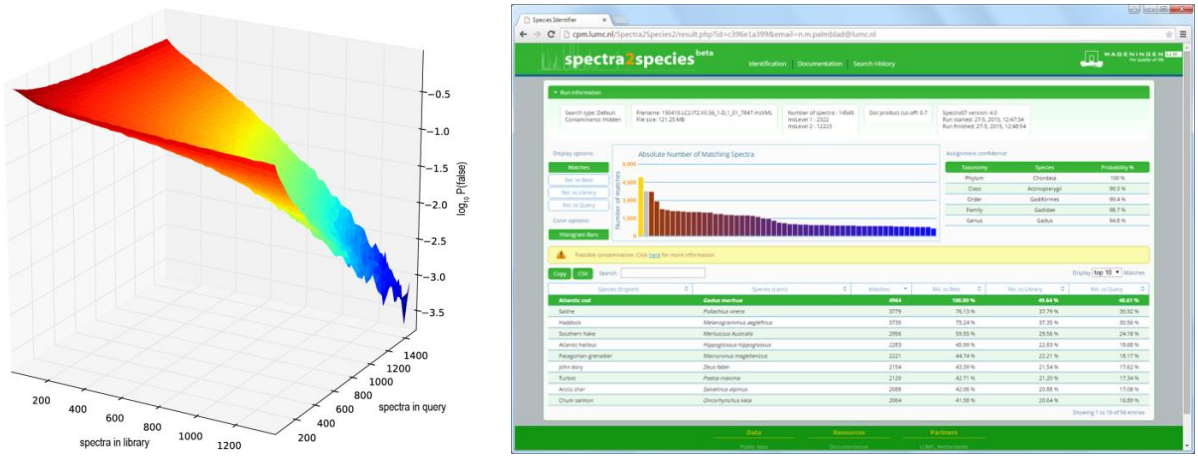

Figure 2. Results from preliminary Monte-Carlo simulation (left) of species identification showing the error rate as a function of the number of spectra in the library and unknown. State-of-the-art mass spectrometers can acquire > 1,000 spectra/minute. The results from simulations will be used to calculate confidence levels for assignment at different taxonomical ranks as in BOLD, and presented with the results (right).

References

1. Zuckerkandl E, Jones RT, Pauling L. A Comparison of Animal Hemoglobins by Tryptic Peptide Pattern Analysis. *Proc. Natl. Acad. Sci. U. S. A.* 1960;46(10):1349-60.

2. Ho YP, Reddy PM. Identification of Pathogens by Mass Spectrometry. *Clin. Chem.* 2010.

3. Ratnasingham S, Hebert PD. bold: The Barcode of Life Data System (http://www.barcodinglife.org). *Mol. Ecol. Notes.* 2007;7(3):355-64.

4. Palmblad M, Deelder AM. Molecular phylogenetics by direct comparison of tandem mass spectra. *Rapid Commun. Mass Spectrom.* 2012;26(7):728-32.

5. Wulff T, Nielsen ME, Deelder AM, Jessen F, Palmblad M. Authentication of fish products by large-scale comparison of tandem mass spectra. *J. Proteome Res.* 2013;12(11):5253-9.

6. Ohana D, Dalebout H, Marissen RJ, Wulff T, Bergquist J, Deelder AM, Palmblad M. Identification of meat products by shotgun spectral matching. *Food Chem.* 2016;203:28-34.

7. Plas-Duivesteijn SJ, Wulff T, Klychnikov O, Ohana D, Dalebout H, van Veelen PA, de Keijzer J, Nessen MA, van der Burgt YE, Deelder AM, Palmblad M. Differentiating samples and experimental protocols by direct comparison of tandem mass spectra. *Rapid Commun. Mass Spectrom*. 2016;30(6):731-8.

8. Rasinger JD, Marbaix H, Dieu M, Fumiere O, Mauro S, Palmblad M, Raes M, Berntssen MHG. Species and tissues specific differentiation of processed animal proteins in aquafeeds using proteomics tools. *Journal of Proteomics*. 2016;147:125-31.

9. Nessen MA, van der Zwaan DJ, Grevers S, Dalebout H, Staats M, Kok E, Palmblad M. Authentication of Closely Related Fish and Derived Fish Products Using Tandem Mass Spectrometry and Spectral Library Matching. *J Agric Food Chem*. 2016.

10. Fusaro VA, Mani DR, Mesirov JP, Carr SA. Prediction of high-responding peptides for targeted protein assays by mass spectrometry. *Nat. Biotechnol.* 2009;27(2):190-8.