COMPREHENSIVE SHARING AND MAPPING OF RP-LC RETENTION TIME INFORMATION

Jan Stanstrup, Urška Vrhovšek

Metabolomics technological platform, Department of Food Quality and Nutrition, Research and Innovation Centre, Fondazione Edmund Mach, San Michele all'Adige (TN), Italy. Email: jan.stanstrup@fmach.it

Objectives

- ❑ To build a user-driven database of retention times (RTs) of compounds.
- ❑ To map the RT of compounds between different chromatographic systems.
- ❑ To use these mappings to predict the RT of compounds not experimentally determined in the system of interest.

Background

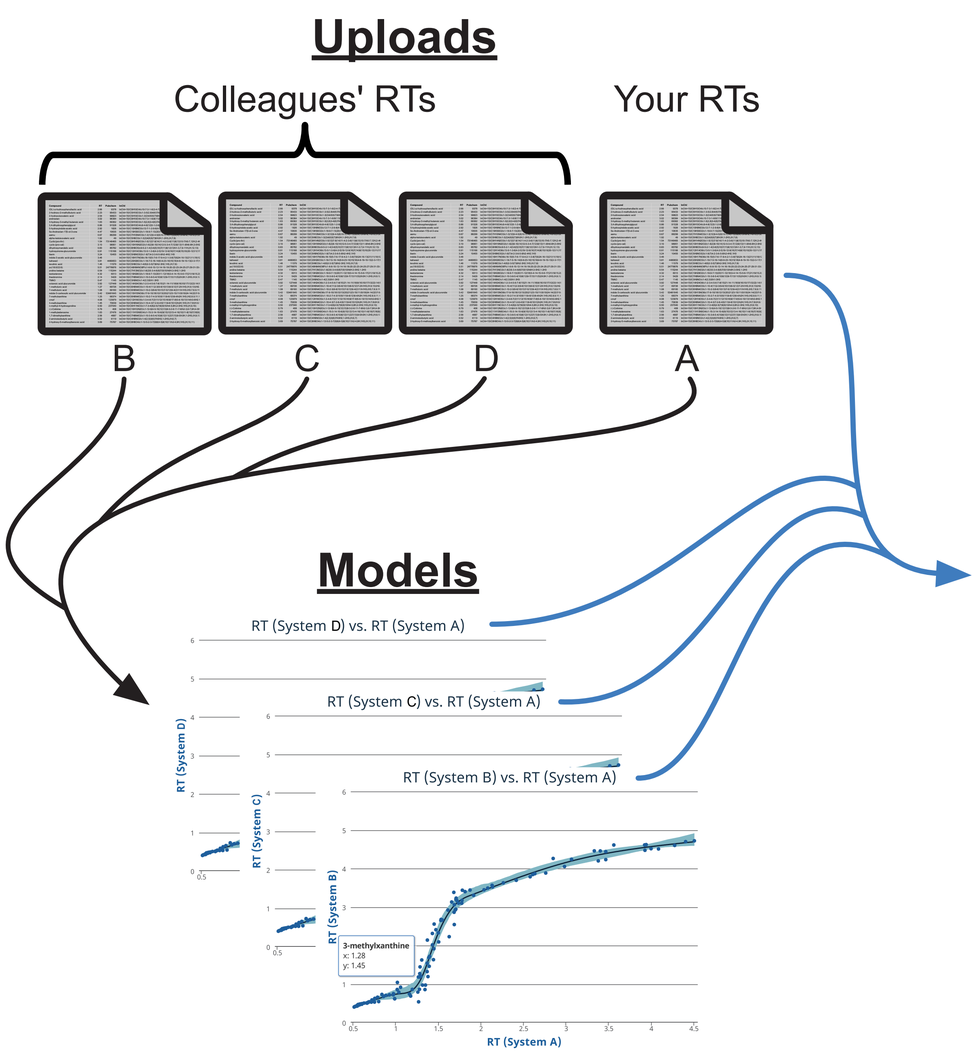
Databases of experimental LC-MS data have been developed with great success with regards to compound fragmentation and these databases have recently been used by automated tools to assist compound identification.
But utilizing *only* the fragmentation means disregarding one, equally important, half of the information in *LC*-MS.
We therefore build a database of compounds' RTs and use this database to predict the RT of compounds in systems where they have not been experimentally determined.

Conclusions

- ✓ The tool is available at www.predret.com. After uploading RTs of compounds in your own systems predicted RTs for other compounds become available.
- ✓ Identification efforts can be reserved for compounds that cannot be excluded by the RT predictions.
- ✓ Community support is required to expand the database.

Results

Building models between all chromatographic systems in the database allows us to predict the RT of compounds in systems where they have not been experimentally determined. The accuracy is dependent on the number of compounds measured in both systems used in the mapping step and on the slope of the curve at the point of prediction. With the current small database (< 1000 compounds) it was possible to predict up to 350 RTs with a median error between 0.04 and 0.19 min depending on the system.



Predictions

Name	Recorded RT	Predicted RT	CI lower	CI upper	Pubchem	InChI
linolenic acid		4.64	3.98	4.75		InChI=1S/
Isoleucine	0.98	0.84	0.70	0.97	6306	InChI=1S/
l-kynurenine		1.38	1.19	1.56	846	InChI=1S/
Leucine	0.98	0.84	0.70	0.97	6106	InChI=1S/
L-lysine	0.40	0.44	0.42	0.48		InChI=1S/
L-malic acid		0.53	0.49	0.61		InChI=1S/
methione	0.70	0.66	0.59	0.73	6137	InChI=1S/C5H11NO2S/c1-9-3-2-4(6)5(7)8/h4
L-norvaline		0.58	0.52	0.65		InChI=1S/C5H11NO2/c1-2-3-4(6)5(7)8/h4H,2
l-ornithine monohydrochloride	0.41	0.42	0.40	0.46	6262	InChI=1S/C5H12N2O2/c6-3-1-2-4(7)5(8)9/h4
phenylalanine	0.50	0.74	0.61	0.86	6140	InChI=1S/C9H11NO2/c10-8(9(11)12)6-7-4-2-
Proline	0.51	0.51	0.48	0.52	145742	InChI=1S/C5H9NO2/c7-5(8)4-2-1-3-6-4/h4,6
L-saccharopine		0.49	0.45	0.52		InChI=1S/C11H20N2O6/c12-7(10(16)17)3-1-
L-serine		0.44	0.42	0.48		InChI=1S/C3H7NO3/c4-2(1-5)3(6)7/h2,5H,1,
Threonine	0.48	0.46	0.44	0.48	6288	InChI=1S/C4H9NO3/c1-2(6)3(5)4(7)8/h2-3,6
tryptophan	2.40	2.18	1.95	2.38	6305	InChI=1S/C11H12N2O2/c12-9(11(14)15)5-7-
tyrosine	0.78	0.70	0.60	0.80	6057	InChI=1S/C9H11NO3/c10-8(9(12)13)5-6-1-3-
lupeol		4.64	3.98	4.75		InChI=1S/C30H50O/c1-19(2)20-11-14-27(5)1
luteolin	3.75	3.80	3.74	4.07		InChI=1S/C15H10O6/c16-8-4-11(19)15-12(20
Valine	0.55	0.62	0.57	0.69	6287	InChI=1S/C5H11NO2/c1-3(2)4(6)5(7)8/h3-4
malonic acid		0.54	0.50	0.62		InChI=1S/C3H4O4/c4-2(5)1-3(6)7/h1H2,(H,4
malvidin 3-p-coumarylglucoside		3.73	3.65	3.96		InChI=1S/C32H30O14/c1-41-22-9-16(10-23(4
malvidin-3,5-diglucoside		3.24	3.06	3.39		InChI=1S/C29H34O17/c1-40-15-3-10(4-16(41
malvidin-3-galactoside		3.47	3.34	3.61		InChI=1S/C23H24O12/c1-31-14-3-9(4-15(32-
malvidin-3-glucoside methyl acetate		3.64	3.50	3.83		InChI=1S/C25H26O13/c1-10(26)35-9-19-21(3
mannitol	0.48	0.49	0.45	0.52		InChI=1S/C6H14O6/c7-1-3(9)5(11)6(12)4(10)
3-coumaric acid	3.53	3.44	3.35	3.59	637541	InChI=1S/C9H8O3/c10-8-3-1-2-7(6-8)4-5-9(1
melibiose		0.49	0.45	0.52		InChI=1S/C12H22O11/c13-1-3-5(14)8(17)10(
methyl arachidate		4.64	3.98	4.75		InChI=1S/C21H42O2/c1-3-4-5-6-7-8-9-10-11-
methyl behete		4.64	3.98	4.75		InChI=1S/C23H46O2/c1-3-4-5-6-7-8-9-10-11-
methyl caprylate		4.29	3.90	4.44		InChI=1S/C9H18O2/c1-3-4-5-6-7-8-9(10)11-2
methyl cis-11-eicosenoate		4.64	3.98	4.75		InChI=1S/C21H40O2/c1-3-4-5-6-7-8-9-10-11-
Methyl henicosanoate		4.64	3.98	4.75		InChI=1S/C22H44O2/c1-3-4-5-6-7-8-9-10-11-
methyl heptadecanoate		4.64	3.98	4.75		InChI=1S/C18H36O2/c1-3-4-5-6-7-8-9-10-11-
methyl jasmote		3.96	3.81	4.23		InChI=1S/C13H20O3/c1-3-4-5-6-11-10(7-8-1

Methods

The prediction tool is made available as a web application at www.predret.com. The user will upload a spreadsheet with RTs of compounds measured in their system along with molecular identifiers such as PubChem CIDs or InChIs.
The system will then map the RT of compounds between systems by building monotonically increasing smooth generalized additive models between RTs experimentally determined in two different systems. This model can then be used to predict the RT of a compound if the RT is known in one system, but not in the other. This mapping and prediction can be done between all systems added to the database.

COMPREHENSIVE SHARING AND MAPPING OF RPLC RETENTION TIME INFORMATION

Jan Stanstrup*, Urška Vrhovšek*

*Department of Food Quality and Nutrition, Research and Innovation Centre, Fondazione Edmund Mach (FEM), San Michele all'Adige (TN), Italy.

Data repositories have been developed with great success to the benefit of the scientific community with regards to the fragmentation of compounds[1,2]. In addition, recently efforts have been made to use automated tools, often assisted by these databases of experimental data, to assist compound identification[3–5]. However, these tools and databases only focus on one aspect of the experimental data: the fragmentation of the compounds formed in mass spectrometers. But utilizing the fragmentation is only using half the available information.

In GC analysis retention indexes are routinely used to make different systems comparable but for LC systems there are currently no coordinated focused efforts to share and exploit information regarding the retention time (RT) of compounds. The reason RT information has been neglected in LC systems is that the RT is specific to a specific chromatographic setup and there exists no agreed upon RT references.

We have therefore sought to rectify this by building a database of compounds' RTs. With this database we are able to map the RT of compounds between systems. RTs, experimentally determined in two different systems, of a number of compounds is used to build monotonically increasing smooth generalized additive models between the RTs in the two systems using the mgcv package[6] for R. This model can then be used to predict the RT of a compound if the RT is known in one system, but not in the other. Building these models between all chromatographic systems in the database thus allows us to predict the RT of a high number of compounds in systems where they have not been experimentally determined.

The tool is completely web-based and available at www.predret.org. On the website it is possible to upload a spreadsheet containing RT information and subsequently download predicted RTs for other compounds based on the data available in the database.

We believe that this tool will greatly help the identification process since compounds that are not compatible with the observed RT can be disregarded. Confirmatory experiments can then be reserved for compounds that *could* have the observed RT. This will allow researchers to complete the feature annotation and compound identification process in a faster and more rational manner and thus safe time and resources, both monetary and environmental.

References

1. Horai, H. *et al.* MassBank: a public repository for sharing mass spectral data for life sciences. *J. Mass Spectrom.* **45,** 703–714 (2010).
2. Wishart, D. S. *et al.* HMDB: a knowledgebase for the human metabolome. *Nucleic Acids Res.* **37,** D603–D610 (2009).
3. Gerlich, M. & Neumann, S. MetFusion: integration of compound identification strategies. *J. Mass Spectrom.* **48,** 291–298 (2013).
4. Wolf, S., Schmidt, S., Müller-Hannemann, M. & Neumann, S. In silico fragmentation for computer assisted identification of metabolite mass spectra. *BMC Bioinformatics* **11,** 148 (2010).
5. Peironcely, J. E. *et al.* Automated Pipeline for De Novo Metabolite Identification Using Mass-Spectrometry-Based Metabolomics. *Anal. Chem.* **85,** 3576–3583 (2013).
6. Wood, S. N. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73,** 3–36 (2011).