Computational and Structural Biotechnology Journal 17 (2019) 1113–1122

COMPUTATIONAL AND STRUCTURAL BIOTECHNOLOGY JOURNAL

journal homepage: www.elsevier.com/locate/csbj

LAS: A Lipid Annotation Service Capable of Explaining the Annotations It Generates☆

Mariano Fernández-López [a,b,*], Alberto Gil-de-la-Fuente [a,b], Joanna Godzien [b,c], Francisco J. Rupérez [b], Coral Barbas [b], Abraham Otero [a,b]

[a] Department of Information Technology, Escuela Politécnica Superior, Universidad CEU-San Pablo, Campus Montepríncipe, Boadilla del Monte, Madrid 28668, Spain
[b] Centre for Metabolomics and Bioanalysis (CEMBIO), Facultad de Farmacia, Universidad CEU-San Pablo, Campus Montepríncipe, Boadilla del Monte, Madrid, 28668, Spain
[c] Clinical Research Centre, Medical University of Bialystok, Poland

ARTICLE INFO

Article history:
Received 7 January 2019
Received in revised form 19 July 2019
Accepted 26 July 2019
Available online 30 July 2019

Keywords:
Lipid annotation
Explanations in natural language
Knowledge based system
REST API service

ABSTRACT

The Lipid Annotation Service (LAS) is a representational state transfer (REST) application programming interface (API) service designed to aid researchers performing lipid annotation. It assigns certainty levels (very unlikely, unlikely, likely, and very likely) to the putative annotations received as input and explains the rationale of such assignments. Its rules, obtained from the Centre for Metabolomics and Bioanalysis (CEMBIO) and from a literature review, enable LAS to extract evidence to support or refute the annotations automatically by checking the inter-rule relationships.

LAS is the first metabolite annotation tool capable of explaining in natural language (English) the evidence that supports or refutes the annotations. This facilitates the understanding of the results by the user and, thus, increases the user's confidence in the results. Concerning its performance, in an evaluation of blood plasma samples whose compounds had previously been identified using well-established standards, LAS yielded an F-measure higher than 80%.

1. Introduction

Untargeted metabolomics aims to identify as many of the metabolites present in a sample as possible. The most popular approach is the detection of the changes between two or more experimental groups over the complete metabolic profile without any prior hypothesis. The metabolites associated with these changes are potential biomarkers [5]. Untargeted metabolomics is ideal for novel discoveries because the coverage of the metabolome is only limited by the sample preparation and the analytical techniques employed. Nevertheless, the huge amount of data generated makes metabolite identification a cumbersome and time-consuming task.

In untargeted metabolomics, the main bottleneck is still compound annotation and identification [22]. Liquid chromatography coupled to mass spectrometry (LC-MS) is a powerful analytical approach [6]. However, the annotation of metabolites is limited because of the variety of methodologies. Ideally, the putative annotations are confirmed using hyphenated setups (LC-MS/MS) by comparing the experimental spectra with the those of commercially available standards (identification confidence level 1 (if available)) or with the spectra available in metabolomic databases (identification confidence level 2) [4]. However, the standards are limited, and, in some cases, the amount of sample or the time available makes MS/MS analysis unfeasible [12]. In these cases, researchers must extract as much information from other analyses, such as MS analysis, retention times (RTs), in-source fragmentation patterns, and isotopic patterns, as possible.

Several software tools have been developed to annotate features using MS^n information [9,21,24,25], but the number of MS^1 metabolite annotation tools remains limited. At the same time, the number of experimentally measured and in silico generated compounds in the databases has quickly increased. This increase

Abbreviations: API, Application Programming Interface; CEMBIO, Centre for Metabolomics and Bioanalysis; CER, Ceramide; CMM, Ceu Mass Mediator; CMM-ES, CMM Expert System; CS, Composite Spectrum; DG, Diradylglycerols; EM, Experimental Mass; FA, Fatty Acid; HMDB, Human Metabolome Database; InChI, IUPAC International Chemical Identifier; LAS, Lipid Identification System; LPC, Lysoglycerophosphocholine; LPE, Lysoglycerophosphoethanolamine; LPG, Lysoglycerophosphoglycerol; LPS, Lysoglycerophosphoserine; MB, MassBank; MZ, MZedDB; MG, Monoradylglycerols; PA, Glycerophosphate; PC, Glycerophosphocholine; PE, Glycerophosphoethanolamine; PI, Glycerophosphoinositol; PG, Glycerophosphoglycerol; PS, Glycerophosphoserine; REST, REpresentational State Transfer; RT, Retention Time; SM, Phosphosphingolipid; ST, Sterol (Cholesterol ester); TG, Triradylglycerols.

☆ Fully documented templates are available in the elsarticle package on CTAN.

* Corresponding author at: Department of Information Technology, Escuela Politécnica Superior, Universidad CEU-San Pablo, Campus Montepríncipe, Boadilla del Monte, Madrid 28668, Spain.
E-mail address: mfernandez.eps@ceu.es (M. Fernández-López).

https://doi.org/10.1016/j.csbj.2019.07.016

makes MS^1 metabolite annotation tools even more important because the number of candidates for a unique *m/z* has increased correspondingly. Some of the most popular databases devoted to metabolomics that are being actively maintained include the Cyc collection [8], KEGG [17], HMDB [31], LipidMaps [29], MassBank [15], the Metabolomics Workbench [28], Metlin [26], MINE [16], and mzCloud [21].

Parallel to this growth in metabolomic databases and software tools for metabolite annotation and identification, a number of tools for lipid identification have been launched. There are currently several software tools that enable the accurate identification of lipids, but all require the use of tandem mass spectrometry. LipidSearch Software (Thermo Scientific, San Jose, CA), LipidPro [1], Liquid [19] MSDial [30], and LipidMatch [18] are some examples of tools devoted to the identification of lipids using tandem mass spectra. The information about the fragmentation of lipids is key to their identification. However, the first step of annotation is commonly performed using just MS^1 data and, sometimes, unequivocal identification is not necessary in untargeted metabolomics. Furthermore, even when unequivocal identification is necessary, a previous filter based on MS^1 data to find which features represent likely lipids reduces the time necessary for the MS^n step of metabolite identification and helps to avoid incorrect assignments. LipidXplorer is another tool for annotation devoted to shotgun lipidomics using data-dependent MS/MS, so information from chromatography cannot be used to annotate the features [14]. VaLID (Visualization and Phospholipid Identification) is a search engine for metabolite annotation with LC/MS data [3]. Its purpose is to identify lipids by *m/z* information searching of all theoretically possible phospholipids. It addresses the need of providing a database for all possible phospholipids because some are not present in the general databases. However, VaLID does not assist the researcher in annotating features for which there is no prior knowledge of the compound type.

Tools to aid lipid annotation are of great value as a first step in lipid identification because in untargeted metabolomic studies researchers have no prior knowledge about the obtained features. The Lipid Annotation Service (LAS) is a representational state transfer (REST) application programming interface (API) service designed to aid researchers during lipid annotation based on MS^1 data. It provides certainty levels for the putative annotations using three types of rules: (1) the ionisation likelihood depending on the type of compound and the modifier used, (2) the relationships between different features corresponding to the same compound, and (3) the RT relationships between lipids within the same class (the hydrophobicity of the lipids pertaining to the same class can be used to support or refute annotations).

Before LAS, the Ceu Mass Mediator Expert System (CMM-ES) [11] provided similar functionality, but it lacked the capability of providing explanations of the evidence supporting or refuting the annotations in natural language. These natural language explanations not only contribute to increasing the users' confidence in the results of the tool because the user can understand the evidence supporting or refuting the annotations but also act as a pedagogical tool for novice researchers who might not know the details of how the evidence was generated. Implementing LAS functionality directly in CMM-ES would have been hampered by technical hurdles because the inference engine used by CMM-ES does not provide explanations about the inferences made. Therefore, it was necessary to reimplement from scratch the complete expert system using technologies that offer this functionality.

The core of LAS is a Prolog knowledge base published through a REST API service running in a Docker container. Both the LAS source code and the Docker container are available and have been documented in detail[1] to provide a ready-to-use platform for the metabolomics community.

[1] https://marianofl1971@bitbucket.org/marianofl1971/mediatorkbservice.git.

2. Methods

Statistical techniques play an important role in metabolite annotation, for example, when looking for correlations between the different signals arising from the same feature for grouping [2]. However, once these statistical techniques have been applied, a significant amount of manual work remains for the researcher [11]. Therefore, a rule-based system representing the knowledge that researchers use in this manual process was built as the core of LAS.

The phases of LAS development were *analysis*, *design*, *implementation and testing*, and *performance evaluation*. During the analysis phase, the problem was studied independently of the technology used to implement it. During the design, the architecture of the system was elaborated, technical decisions concerning programming languages and virtualisation were made, and tests were designed. During the implementation and testing, the system code and the tests were written in parallel. Thus, both have been debugged at the same time, and large deviations in the code with respect to its intended working have been prevented. The analysis (Section 2.2) and the design and implementation (Section 2.3) are presented below.

2.1. Functionality Borrowed from CMM-ES

The following knowledge and data have been borrowed from CMM-ES during the development of LAS:

- *Knowledge to annotate lipids*: the knowledge represented in the rules was already acquired from CEMBIO experts along the development of CMM-ES.
- *Compounds data*: the data on compunds were already available.

Nevertheless, as will be shown in the following sections, LAS has been implemented from scratch, and the way in which it represents both knowledge and data are completely different from that of CMM-ES. This has been the basis for LAS to contain new features like explanation of results, modular architecture, availability as a stand-alone service, or dockerization.

2.2. Analysis

LAS takes as its input a set of putative annotations. For example, each row of Table 1 is a putative annotation. Each *putative annotation* consists of an *identifier* (e.g. 3), an *empirical compound* (*m/z* and RT observation) (e.g. 400.3432@18.842525), an *adduct* (e.g. M + H), and a *theoretical compound* (e.g. L-palmitoylcarnitine with formula C23H45NO4 and lipid class fatty acyl carnitines [FA0707]). The lipid classification is performed using LipidMaps classification [10]. This classification has four levels: category, main class, subclass, and class level 4, and each compound can only belong to a single level. LAS applies the rules based on the compound subclass. The first number of the empirical compound represents its *m/z*, and the second number is its RT. The interpretation of the first putative annotation in Table 1 is that 'empirical compound 400.3432@18.842525 may be theoretical compound L-palmitoylcarnitine with adduct M+H'.

LAS uses three sets of rules to *assign certainty levels* to the putative annotations. These sets of rules will be presented in Sections 2.2.1, 2.2.2, and 2.2.3. The possible *certainty levels* (CLs) are *very unlikely*, *unlikely*, *not applicable (N/A)*, *likely*, and *very likely*. The certainty order is

very unlikely<*unlikely*<*not applicable*(*N*/*A*)<*likely*<*very likely*.

For example, Table 2 shows, for the putative annotations of Table 1, the CLs assigned according to the three set of rules (ionisation rules, adduct intensity rules, and retention time rules). Thus, the meaning of the first row in Table 2 is "the CL of putative annotation 3 is "N/A" according

Table 1
Example of input putative annotations for LAS.

Putative annotation number	*m/z*	RT	Adduct	Theoretical compound	Formula	Lipid class
3	400.3432	18.8425	M + H	L-palmitoylcarnitine	$C_{23}H_{45}NO_4$	FA0707
78	422.32336	18.8425	M + 2H	PG(P-20:0/22:2(13Z,16Z))	$C_{48}H_{91}O_4P$	GP0403
80	422.32336	18.8425	M + Na	L-palmitoylcarnitine	$C_{23}H_{45}NO_4$	FA0707
95	316.24945	8.1449	M + H	O-decanoyl-R-carnitine	$C_{17}H_{33}NO_4$	GL0301
211	281.24765	28.2695	M + H	14,17-Octadecadienoic acid	$C_{18}H_{32}O_2$	FA0103
293	496.3427	19.4689	M + H	PC(O-14:0/2:0)	$C_{24}H_{50}NO_7P$	GP0102
312	518.3226	19.4689	M + Na	PC(O-14:0/2:0)	$C_{24}H_{50}NO_7P$	GP0102

to the ionisation rules, 'very likely' according to the adduct intensity rules, and 'very likely' according to the retention time rules".

LAS provides, for each CL assignment, an *explanation in English*.

Along with the following subsections, the set of rules that generate multiple explained CL assignments for lipids will be presented. Each section includes the rationale and the definition of its rules and the linguistic patterns used to generate the explanations. The current version of LAS generates explanations in English. To generate explanations in other languages, more linguistic patterns should be added.

2.2.1. Ionisation Rules

2.2.1.1. Rationale and Definition of Ionisation Rules. The majority of molecules are ionised by simple protonation ($[M + H]^+$) in positive ionisation mode or deprotonation ($[M\text{-}H]^-$) in negative mode. Some compounds, because of their structure, cannot form such ions and can be ionised only by the formation of other adducts [20]. For example, phosphoinositols (PIs) cannot be ionised by protonation, $[M + H]^+$; therefore, they are not detectable in positive mode, unless they form sodium $[M + Na]^+$ or potassium $[M + K]^+$ adducts. Phosphocholines (PCs) are never ionised by deprotonation ($[M\text{-}H]^-$). Consequently, to ionise PCs, a formate or an acetate adduct is needed. Although the majority of molecules are ionisable in both polarity modes, some of them only form positive ions and others only form negative ions.

The tendency to form an adduct depends on the lipid class, ionisation mode and mobile phase modifier used [12]. For example, PCs in negative mode primarily form $[M + HCOO]^-$ or $[M + CH_3COO]^-$ depending on the modifier used (HCOOH or CH_3COOH); they may also form $[M + Cl]^-$ with lower intensity; but they never form $[M\text{-}H]^-$ or $[M\text{-}H\text{-}H_2O]^-$ [20]. These rules are applied to such lipid classes as: fatty acid (FA), PC, lysoglycerophosphocholine (LPC), glycerophosphoethanolamine (PE), lysoglycerophosphoethanolamine (LPE), glycerophosphoinositol (PI), glycerophosphoglycerol (PG), lysoglycerophosphoglycerol (LPG), glycerophosphoserine (PS), lysoglycerophosphoserine (LPS), glycerophosphate (PA), monoradylglycerols (MG), diradylglycerols (DG), triradylglycerols (TG), ceramide (CER), phosphosphingolipid (SM) and cholesterol ester (ST) according to the LipidMaps classification.

Rules are used to represent the knowledge described in the former paragraphs. A *rule* is made up of an *antecedent* (a condition to be checked) and a *consequence* (an action to carry out). In the specific case of LAS, the consequence is a CL assignment.

Table 2
Example of annotations to be returned by LAS for Table 1 (N/A: Not applicable).

Putative annotation number	Ionisation rules	Adduct intensity rules	RT rules
3	N/A	Very likely	Very likely
78	Very unlikely	N/A	N/A
80	N/A	Very likely	Very likely
95	N/A	N/A	Very likely
211	Very likely	N/A	Very likely
293	Very likely	Very likely	N/A
312	Very likely	Very likely	N/A

Definition 1. An ionisation rule is that whose antecedent has the following parts:

- *Modifier*: NH_3, HCOO, CH_3COO, $HCOONH_3$, CH_3COONH_3, or *indistinct*.
- *Ionisation*: *positive* or *negative*.
- *Levels in the lipid hierarchy*: a sequence of classes of lipids ordered from the most general to the most particular.
- *Adduct*: the assumed adduct in the putative annotation.

Two examples of ionisation rules are presented below.

Example 1. If the ionisation is negative, the levels in the lipid hierarchy are GP and GP01 and the adduct is M + 2H; thus, the putative annotation is very unlikely.

The modifier is not mentioned because it is indistinct. The CL assignment 'putative annotation 78 is *very unlikely*' in Table 2 has been deduced by applying this rule.

Example 2. If the ionisation is positive, the levels in the lipid hierarchy are FA and FA01 and the adduct is M + H; thus, the putative annotation is very likely.

The modifier is not mentioned either because it is indistinct. The putative annotation of 211 as *very likely* in Table 2 has been deduced by applying this rule.

The complete list of the 109 ionisation rules is available in *transformer/resources/lipids_ionization_ type_rules.ods* in the repository.

2.2.1.2. How Ionisation Rules Are Explained. The explanation of the execution of an ionisation rule must contain information about the empirical compound, the adduct, the theoretical compound, if the ionisation is positive or negative, and the levels in the lipid hierarchy, as well as the CL of the rule. An impersonal sentence is used in the explanation: 'it is likely that empirical compound (...) is theoretical compound (...)' or 'it is highly unlikely that empirical compound (...) is theoretical compound (...)'.

A definition obtained considering what has been said in the previous paragraph is presented below.

Definition 2. The linguistic pattern followed by the explanation of each ionisation rule is as follows:

> It is *CertaintyLevel* that empirical compound *m/z*@*RT* (in an experiment with *ionisationInEnglish* ionisation and *ModifierInEnglish*) is theoretical compound *ThName* with adduct *Adduct*, which belongs to the hierarchy *CategoryU*, *MainClassU*, *SubClassU*.

Here,

- *CertaintyLevel* is CL except *not applicable*,
- *m/z@RT* is the empirical compound of the putative annotation,
- *ionisationInEnglish* can be *positive* or *negative*,
- *ModifierInEnglish* is a modifier,
- *ThName* is the name of the theoretical compound of the putative annotation,

- *Adduct* is the adduct of the putative annotation, and
- *CategoryU*, *MainClassU*, and *SubClassU* is a sequence in the lipid hierarchy to which the theoretical compound belongs.

Example 3. An explanation according to the pattern defined in Example 2 is presented below:

> Putative annotation 211 is very likely according to ionisation rules.
>
> • *Rule type*: Ionisation.
>
> • *Core information*: It is very likely that empirical compound 281.24765@28.2695 (in an experiment with positive ionisation and no modifier) is theoretical compound 14,17-octadecadienoic acid with adduct M + H, which belongs to the hierarchy FA, FA01.

2.2.2. Adduct Intensity Rules

A list of possible, impossible, and preferred ions for distinct compound types can be established, as well as relationships between the expected intensities of the different ions. For example, PCs can be ionised in positive mode by protonation ($[M + H]^+$), but they can also be ionised with sodium ($[M + Na]^+$) and potassium ($[M + K]^+$). However, the main signal is $[M + H]^+$ and all others have a lower intensity. Hence, a putative annotation for $[PC + Na]^+$ can be right only if the signal corresponding to $[PC + H]^+$ is also present, otherwise it is an incorrect assignment.

For the current version of LAS, only the intensity pattern of adducts M + H > M + Na > M + K have been considered. More intensity patterns will be added in the future. Adduct intensity rules are defined as follows.

Definition 3. An adduct intensity rule checks whether there are pairs of compounds putatively identified as the same theoretical compound, and with adducts satisfying some of the following conditions.

1. The adduct of the first compound is M + H and the adduct of the second is M + Na, or vice versa.
2. The adduct of the first compound is M + H and the adduct of the second is M + K, or vice versa.
3. The adduct of the first compound is M + Na and the adduct of the second one is M + K, or vice versa.

The CL of the consequent will be *very likely* in every case.

Example 4. For instance, 'putative annotation 3 is very likely according to the adduct intensity rules' is inferred using relationship 1 (M + H, M + Na) with putative annotation 80.

The results of the adduct intensity rules are used in some cases to increase the CL of the ionisation rules. Thus, if an ionisation rule concludes that putative annotation *p* is *likely*, but an adduct intensity rule concludes that *p* is *very likely*, then the CL of the ionisation rule applied to *p* is increased to *very likely*, and the adduct intensity rule is considered as *additional evidence*. For example, an ionisation rule infers the CL assignment 'putative annotation 312 is likely'. However, as shown in Table 2, it is transformed into 'putative annotation 312 is *very likely* according to the ionisation rules' because 'putative annotation 312 is *very likely* according to the adduct intensity rules'.

2.2.2.1. How Adduct Intensity Rules Are Explained. The explanation of an adduct intensity rule has to include the empirical compound and the adduct and the theoretical compound of each putative annotation involved in the relationship between adducts.

Definition 4. The linguistic pattern followed by the explanation of each adduct intensity rule is as follows:

> It is very likely that empirical compound *m/z1@RT1* is theoretical compound *ThName* with adduct *Adduct1* because empirical compound *m/z2@RT2* with adduct *Adduct2* may also be theoretical compound *ThName*.

where

- *m/z1s@RT1* is the empirical compound of the first putative annotation.
- *ThName* is the name of the theoretical compound of both putative annotations.
- *Adduct1* is the adduct of the first putative annotation.
- *m/z2s@RT2* is the empirical compound of the second putative annotation.
- *Adduct2* is the adduct of the second putative annotation.

Example 4 is explained below.

> Putative annotation 3 is *very likely* according to adduct intensity rules.
>
> • *Rule type*: Adduct intensity.
>
> • *Core information*: It is very likely that empirical compound 400.3432@18.8425 is theoretical compound L-palmitoylcarnitine with adduct M + H because, for example, empirical compound 422.32336@18.8425 with adduct M + Na may be theoretical compound L-palmitoylcarnitine as well.

If an adduct intensity rule causes an increase in the CL of an ionisation rule, the explanation of such ionisation rule will include a reference to the explanation the corresponding adduct intensity rule.

2.2.3. Retention Time Rules

The RT reflects the time that a particular molecule spends in the column because it is retained by the stationary phase. This time depends on the mechanisms of retention, column geometry and temperature, instrument dwell volume, mobile phase, modifier, and gradient. In the case of reverse-phase (RP) chromatography, the most common type of chromatography, polar analyte molecules interact little with the non-polar bed, and, thus, they elute very early. On the other hand, non-polar molecules will be retained for longer, thus eluting later. Although the flexibility of LC and the possibility to modify many experimental parameters (mobile phase, gradient, modifiers, flow, temperature, and type of column, as well as its length and diameter) make this technique very powerful, they also make obtaining reproducible RTs impossible. This is probably why there is only one proposal for metabolite annotation where dynamic RT prediction based on the chromatographic linear solvent strength for RP-LC data is used to support steroid identification Randazzo et al. [23]. The behaviour of molecules inside a chromatographic column under set chemical conditions is well defined, especially for compounds belonging to the same class Godzien et al. [12]. In consequence, although the absolute RT is very difficult to predict (even though it is not impossible Cao et al. [7]; Hagiwara et al. [13]; Stanstrup et al. [27]), the prediction of the relative order of elution is feasible for certain compounds belonging to the same chemical class that are analysed under the same analytical conditions. This can be a valuable aid in the analytical process, and the RTs for some types of compounds can be compared. This is especially interesting for lipids belonging

to the same class because their backbones are the same, and this RT prediction is based on two relationships: the length of the carbon chain and the degree of unsaturation (number of double bonds) Godzien et al. [12]. As the chain length increases, the hydrophobicity of a lipid molecule also increases, so the lipid will be retained longer in a RP column. On the other hand, double bonds increase the polarity of lipids, thus reducing the RT.

Consequently, the notional RT rule can be defined as follows.

Definition 5. Let p_1 be a putative annotation that establishes that a particular empirical compound, m/z_1 @ RT_1, is theoretical compound p_1 having c_1 carbons and d_1 double bonds. Let p_2 be a putative annotation that establishes that a particular empirical compound, m/z_2 @ RT_2, is theoretical compound p_2 having c_2 carbons and d_2 double bonds. Let us suppose that both p_1 and p_2 are lipid type p. Then, there is evidence supporting p_1 because of RT if and only if some of the following conditions are satisfied.

1. $c_1 < c_2$ and $d_1 = d_2$ and $rt_1 < rt_2$
2. $c_1 = c_2$ and $d_1 < d_2$ and $rt_1 > rt_2$

The first condition means that, if the number of double bonds is the same for two lipids of the same class, that with a longer chain will have a higher RT. The second condition means that, if the length of the chains is the same for two lipids of the same class, the one with the least number of double bonds will have a higher RT.

Definition 6. Let p_1 be a putative annotation that establishes that a particular empirical compound, m/z_1 @ RT_1, is theoretical compound p_1 having c_1 carbons and d_1 double bonds. Let p_2 be a putative annotation that establishes that a particular empirical compound, m/z_2 @ RT_2, is theoretical compound p_2 having c_2 carbons and d_2 double bonds. Let us suppose that both p_1 and p_2 belong to lipid type p. Then, there is evidence against p_1 because of RT if and only if some of the following conditions are satisfied.

1. $c_1 < c_2$ and $d_1 = d_2$ and $rt_1 \geq rt_2$
2. $c_1 = c_2$ and $d_1 < d_2$ and $rt_1 \leq rt_2$

Definition 7. Let p_1 be a putative annotation; the following RT rules can be applied.

1. If there is only evidence supporting p_1, then *putative annotation p_1 is very likely*.
2. If there is only evidence against p_1, then *putative annotation p_1 is unlikely*.
3. If there is both evidence supporting and against p_1, then *putative annotation p_1 is likely*.

Example 5. Putative annotation 95 is O-decanoyl-R-carnitine, a CAR lipid with 10 carbons and no double bond. Its RT is 8.1449. Moreover, putative annotation 3 is L-palmitoylcarnitine, another CAR lipid with 16 carbons and no double bond. Its RT is 18.8425. Therefore, condition 1 of Definition 5 is satisfied and, consequently, there is evidence supporting both putative annotations.

2.2.3.1. How Retention Time Rules Are Explained. As shown in the following definition, the explanation of a retention time rule includes the empirical compound and the theoretical compound with the number of carbons atoms and double bonds. If there is evidence supporting or refuting the putative annotation, it will be explained as well, showing the information about the corresponding putative annotations.

Definition 8. The linguistic pattern followed in the explanation of each RT rule is presented below; moreover, explanations of the supporting or refuting evidence are provided.

> It is *CertaintyLevel* that empirical compound *m/z*@*RT* is theoretical compound *ThName* with adduct *Adduct*. The theoretical compound is a *LipidType* lipid with *NCarbons* carbons and *NDBE* double bonds.

Here,

- *m/z@RT* is an empirical compound,
- *ThName* is the name of a theoretical compound,
- *Adduct* is an adduct,
- *LipidType* is a type of lipid,
- *NCarbons* is a number, and
- *NDBE* is another number.

If there are no double bonds or there is only one, the linguistic pattern for this part of the sentence must be *no double bond* or *one double bond*, respectively.

Example 6. Example 5 is explained below.

> Putative annotation 95 is *very likely* according to the retention time rules.
>
> • *Rule type*: Retention time.
>
> • *Core information*: It is very likely that empirical compound 316.24945@8.1449 is theoretical compound O-decanoyl-R-carnitine with adduct M + H. The theoretical compound is a CAR lipid with 10 carbons and no double bond.
>
> • *Evidence supporting*:
>
> 1. It is very likely that compound 400.3432@18.8425 is theoretical compound L-palmitoylcarnitine with adduct M + H, another CAR lipid with 16 carbons and no double bond.
>
> 2. It is very likely that compound 422.32336@18.8425 is theoretical compound L-palmitoylcarnitine with adduct M + Na, another CAR lipid with 16 carbons and no double bond.

2.3. Design and Implementation

The logic of the system is supported by a knowledge base (KB) in SWI-Prolog[2] (see Fig. 1). Its most basic component is *the CL module*, which allows the representation of knowledge through a multi-valued logic according to the values specified at the beginning of Section 2.2. Using this multi-valued logic, the *meta-ontology* implements frame oriented primitives: *instance of*, *subclass of*, etc. The *Compound Ontology* represents the CMM database of compounds based on the lower layers. The *module of ionisation rules* are generated from a spreadsheet that contains the rules. The *main module of the KB* loads the ionisation rules and implements the rest of the rules. This module builds a trace for explanations, which is natural language independent. That is, the formalization of the trace has been thought so that generators for different natural languages (English, Spanish, Polish, etc.) can be developed. Finally, the *generator of explanations in English* obtains the explanations from the trace built by the main module.

[2] www.swi-prolog.org.

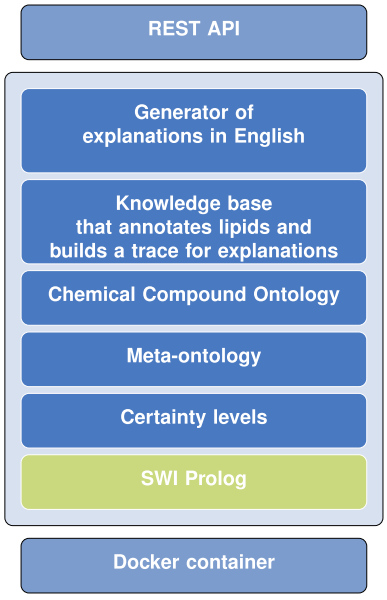

Fig. 1. Software architecture.

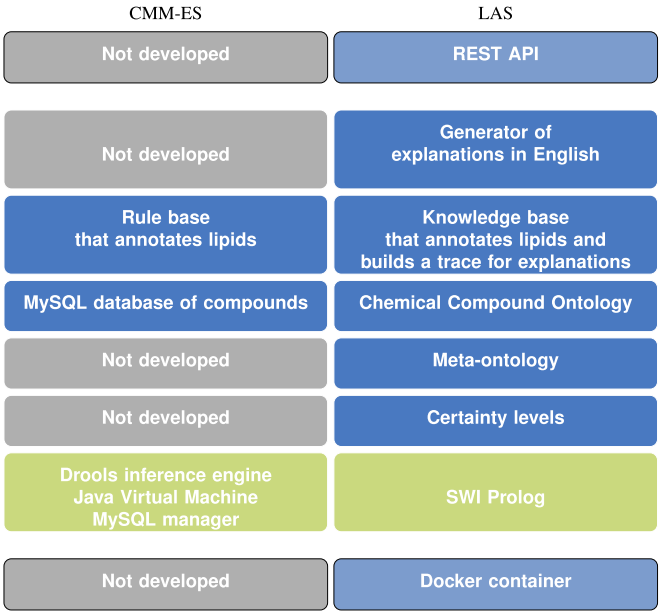

Fig. 2. LAS versus CMM-ES.

The REST API (also written in Prolog) eases the shareability of LAS as a Web resource and its integration in different systems. The whole system has been dockerised so that it can be installed on any machine regardless of the configuration, such as library version and environment variables. All the software is publicly available and documented in detail (see Footnote 1).

3. Results and Discussion

This section presents the evaluation of LAS considering its precision, recall, F-measure, the role of each group of rules (ionisation, adduct relationship, and RT), and the performance in terms of execution time, as well as the quality of the explanations obtained.

The performance of LAS in terms of precision, recall, and F-measure was determined using a gold standard provided by a CEMBIO expert. The gold standard was made up of 30 experimental mass signals (plasma samples,[3]) previously identified by MS/MS or using commercially available standards (Sigma–Aldrich and Fluka Analytical). Plasma samples were prepared by simple deproteinisation with a cold methanol/ethanol (1/1) mixture. Standards were prepared in methanol with concentrations between 2 and 10 ppm. To obtain these data, analysis of the samples was carried out using a high-performance LC (HPLC) system (1200 series, Agilent Technologies) connected to an Agilent quadrupole time-of-flight QTOF (6520) MS detector. Data were collected in positive and negative electrospray ionisation (ESI) ion modes in separate runs on a QTOF MS operated in the m/z range of 50 to 1000 with an acquisition rate of 1 scan/s. The capillary voltage was set to 3000 V for positive and 4000 V for negative ionisation mode; the nebuliser gas flow rate was 10.5 L/min. Accurate mass measurements were obtained by means of an automated calibrant delivery system using a dual-nebuliser ESI source that continuously introduces a calibrating solution. All 30 experimental masses correspond to lipids.

We searched for putative annotations for the 30 experimental masses using the Ceu Mas Mediator tool, which returned a total of 891 putative annotations. These annotations were provided as input to LAS. The set of annotations generated by the LAS service, together with their corresponding evidence, was compared with the gold standard provided by CEMBIO.

[3] https://bitbucket.org/marianofl1971/mediatorkbservice/src/master/evaluation/all_datasets_CMM.xls.

3.1. Previous Definitions

Definition 9. Let p_1 and p_2 be two putative annotations. It is said that p_1 is supported by more evidence than p_2 if and only if the following conditions hold.

1. There is some set of rules (ionisation rules, adduct intensity rules, or RT rules) that supports p_1 as *likely* or as *very likely*.
2. There is at least a set of rules that supports p_1 with more evidence than p_2.
3. There is no set of rules that supports p_2 with more evidence than p_1.

Example 7. In Table 2, putative annotation 80 is supported by more evidence than 78 because 80 is very likely for both adduct relationship and retention time rules; meanwhile, 78 is N/A for both. Moreover, 78 is very unlikely according to the ionisation rules.

Let us note that there may be non-comparable putative annotations. For example, if the ionisation rules assign more evidence to p_1 but the RT rules assign more evidence to p_2.

Definition 10. Let $p_1, \ldots, p_n$ putative annotations refer to the same empirical compound. p_1 is the putative annotation with the strongest evidence if and only if there is no putative annotation $(p_2, \ldots, p_n)$ supported by more evidence than p_1.

Example 8. In the set of putative annotations made by 78 and 80, both referring to empirical compound 422.32336@18.8425, 80 is the putative annotation supported by the strongest evidence.

Definition 11. Two theoretical compounds are considered to be the same if and only if some of the following conditions are true.

1. Both have the same name.
2. Their names are synonyms according to Pubchem.[4]
3. They belong to the same type of lipid with the same number of carbons and the same number of double bonds.

Example 9. The compounds MG(18:1(11Z)/0:0/0:0) and MG(0:0/18:1(9Z)/0:0) are considered the same because they satisfy condition 3.

[4] https://pubchem.ncbi.nlm.nih.gov/search/.

Fig. 3. Presentation of the results of the LAS service.

Metabolites found for mass: 399.3367, retention time: 18.842525 and adduct: M+H -> 3

Hypothesis Id	Id	Name	Formula	Molecular Weight	Retention Time	error PPM	Evidences
1	32675	L-palmitoylcarnitine	C23H45NO4	399.334859	18.842525	5	1. **Annotation EXPECTED, Rule Type: ADDUCT_RELATION** **Core Info**: The empirical compound 399.3367@18.842525 is expected to be the theoretical compound L-palmitoylcarnitine with adduct M+H, because, for example, the empirical compound 421.31686@18.842525 with adduct M+Na may be the theoretical compound L-palmitoylcarnitine as well. 2. **Annotation EXPECTED, Rule Type: RETENTION_TIME** **Core Info**: The empirical compound 399.3367@18.842525 is expected to be the theoretical compound L-palmitoylcarnitine with adduct M+H. The theoretical compound is a _CAR(_ lipid with 16 carbons and no double bond **Evidences For**: For example, the compound 315.2424@8.144917 may be the theoretical compound Decanoylcarnitine with adduct M+H, another _CAR(_ lipid with 10 carbons and no double bond.
2	32694	O-palmitoylcarnitine	C23H45NO4	399.334859	18.842525	5	1. **Annotation EXPECTED, Rule Type: ADDUCT_RELATION** **Core Info**: The empirical compound 399.3367@18.842525 is expected to be the theoretical compound O-palmitoylcarnitine with adduct M+H, because, for example, the empirical compound 421.31686@18.842525 with adduct M+Na may be the theoretical compound O-palmitoylcarnitine as well.

Fig. 4. Input form for the LAS service.

3.2. Calculation of the Performance in Terms of Precision, Recall, and F-Measure

Given the definitions in the previous section, the numbers of true positives (TP) and false positives (FP) are calculated as follows.

1. The KB is queried with the set of putative annotations generated by querying CMM. For each annotation the multiple certainty level assignment with the strongest evidence is taken.
2. Then, each annotation of the gold standard is compared to the putative annotation with the strongest evidence proposed by LAS. The following cases are possible.
   (a) All the putative annotations with the strongest evidence refer to theoretical compounds considered to be same and match the gold standard. In such a case, the set of putative annotations with the strongest evidence are counted as TPs.
   (b) Some putative annotations with the strongest evidence refer to theoretical compounds considered to be the same and match the real identification criteria, whereas others do not. In that case, the set of putative annotations with the strongest evidence are counted as both TPs and FPs.
   (c) There are some putative annotations with the strongest evidence, but there is no putative annotation with the strongest evidence matching the gold standard. In such a case, we count an FP.

The precision, recall, and F-measure have been calculated,[5] yielding 0.76, 0.93, and 0.84, respectively. In the next section, we will analyse in detail the contribution that each of the three groups of rules has played to achieve these results.

3.3. Contribution of Each Group of Rules

For the 30 compounds that made up the gold standard, the following rules of each type were executed.

- For 23 compounds, ionisation rules were executed; this was the only type of rule executed for 2 compounds.
- Some adduct intensity rules were executed for 10 compounds. Adduct intensity rules were always executed in conjunction with other rules. However, there is one case (400.3432@18.8425 with *m/z* 422.32336 and adduct M + Na) where the application of this type of rule was decisive in achieving the correct annotation.
- Some RT rules were executed for 25 identifications. RT rules were the only type of rule executed for three compounds.

Based on these results, it can be concluded that the RT rules were the most often executed and, thus, contributed most to the results, followed by the ionisation rules and the adduct intensity rules.

3.4. Execution Time Performance

Given that both the adduct intensity rules and RT rules check conditions with all pairs of putative annotations, the time required for LAS is quadratic with respect to the number of the putative annotations. Concerning real execution times, using an HP Envy 15 with an Intel i7, 16 GB RAM, and a 1 TB hard disk to host the service, LAS provided an answer in approximately 3 s for an input of 1300 putative annotations.

3.5. Integration Into the Ceu Mass Mediator On-line Tool

Before the implementation of LAS, the CMM tool also had an expert system based on knowledge that used rules similar to those presented in this paper. However, there are significant differences between the functionality previously provided by this expert system and LAS (see Fig. 2). CMM cannot provide an explanation of the evidence supporting or refuting the annotations. LAS explanations, which are provided in natural language, will increase the users' confidence in the results of the tool because they can understand the evidence supporting or refuting the annotations. Furthermore, the LAS results could be used as a pedagogical tool for novice researchers who might not know how the evidence was generated. Providing this functionality required the reimplementation of the expert system from scratch because the technology used in CMM does not have an explanatory capability.

Furthermore, LAS formalises the CLs through a meta-ontology and an ontology of chemical compounds with the purpose of obtaining a modular, extensible, and theoretically sound system. In fact, both the meta-ontology and the ontology could be reused in other systems. CMM-ES was designed in an ad hoc manner and, thus, cannot be easily reused. Finally, LAS is accessible via a REST service. In contrast, CMM-ES is embedded in a larger Java system, and it cannot be used in an independent way. Nevertheless, LAS has been integrated as a CMM service in the on-line tool accessible through the web page `http://ceumass.eps.uspceu.es/prolog_batch_advanced_search.xhtml` (see Figs. 3 and 4). The service takes advantage of all the previous features that CMM offers for the identification of compounds and the automatic detection of adducts based on composite spectra, as well as the possibility of restricting the search to specific databases or specific compound types, the usage of a chemical alphabet, and the selection of different adducts, thus reducing the number of putative annotations for each feature. LAS has been used internally in CEMBIO since late 2018. This has allowed us to debug and provide feedback about the tool before its public launch.

4. Conclusions and Future Work

In this paper, we have presented LAS, a publicly available service that aids in the identification of lipids. The core of the system is a KB in Prolog that executes rules acquired from a group of experts belonging to CEMBIO. These rules can be divided into three large groups: ionisation rules, adduct intensity rules, and RT rules. A unique feature of the system is that it provides explanations in natural language (English) for the evidence supporting or refuting the annotations. In the evaluation carried out on a set of 30 compounds that had been identified with the use of standards or MS/MS, the LAS expert system obtained an F-measure of 0.84. LAS has been made available as a REST API service, which facilitates sharing and integration in different systems. The service has been dockerised to make it platform independent.

Regarding future work, the KB can be improved by taking into account intensities and relationships between more types of adducts in the adduct intensity rules. Concerning the explanations, they can be extended to other natural languages (Spanish, Polish, etc.) by adding linguistic patterns. Finally, content negotiation could be a useful functionality so that the results of the service can be presented in different formats (e.g. HTML).

Acknowledgments

This work was supported by grants from the NOVELREM project (Comunidad de Madrid Ref: B2017/BMD3751). AGF acknowledges Fundación Universitaria San Pablo CEU for his PhD fellowship.

References

[1] Ahmed Z, Mayr M, Zeeshan S, Dandekar T, Mueller MJ, Fekete A. Lipid-pro: a computational lipid identification solution for untargeted lipidomics on data-independent acquisition tandem mass spectrometry platforms. Bioinformatics 2015;31:1150–3.
[2] Alonso A, Marsal S, Julia A. Analytical methods in untargeted metabolomics: state of the art in 2015. Front Bioeng Biotechnol 2015;3.

[5] `evaluation/t_vs_o.ods` in the repository (see Footnote 1).

[3] Blanchard AP, McDowell GSV, Valenzuela N, Xu H, Gelbard S, Bertrand M, et al. Visualization and phospholipid identification (valid): online integrated search engine capable of identifying and visualizing glycerophospholipids with given mass. Bioinformatics 2013;29:284–5.
[4] Blazenovic I, Kind T, Ji J, Fiehn O. Software tools and approaches for compound identification of lc-ms/ms data in metabolomics. Metabolites 2018;8:1989–2218.
[5] Brennan L. Metabolomics in nutrition research: current status and perspectives. Biochem Soc Trans 2013;41:670.
[6] Broeckling CD, Ganna A, Layer M, Brown K, Sutton B, Ingelsson E, et al. Enabling efficient and confident annotation of lc– ms metabolomics data through ms1 spectrum and time prediction. Anal Chem 2016;88:9226–34.
[7] Cao M, Fraser K, Huege J, Featonby T, Rasmussen S, Jones C. Predicting retention time in hydrophilic interaction liquid chromatography mass spectrometry and its use for peak annotation in metabolomics. Metabolomics 2014;11:696–706.
[8] Caspi R, Billington R, Fulcher CA, Keseler IM, Kothari A, Krummenacker M, et al. The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. Nucleic Acids Res 2016;44:D471–80.
[9] Dührkop K, Shen H, Meusel M, Rousu J, Böcker S. Searching molecular structure databases with tandem mass spectra using csi:fingerid. Proc Natl Acad Sci U S A 2015; 112:12580–5.
[10] Fahy E, Cotter D, Sud M, Subramaniam S. Lipid classification, structures and tools. Biochim Biophys Acta 2011;1811:637–47.
[11] Gil-de-la Fuente A, Fernandez Lopez M, Otero A, Godzien J, Ruperez F, Barbas C. Knowledge-based metabolite annotation tool: Ceu mass mediator. J Pharm Biomed Anal 2018;154:138–49.
[12] Godzien J, Armitage EG, Rupérez FJ, Barbas C, Ciborowski M, Jorge I, et al. A single invial dual extraction strategy for the simultaneous lipidomics and proteomics analysis of hdl and ldl fractions. J Proteome Res 2016;15:1762–75.
[13] Hagiwara T, Saito S, Ujiie Y, Imai K, Kakuta M, Kadota K, et al. Hplc retention time prediction for metabolome analysis. Bioinformation 2010;5:255–8.
[14] Herzog R, Schwudke D, Shevchenko A. Lipidxplorer: software for quantitative shotgun lipidomics compatible with multiple mass spectrometry platforms. Curr Protoc Bioinformatics 2018;43:14.12.1–30.
[15] Horai H, Arita M, Nihei Y, Ikeda T, Ojima Y, Kakazu Y, et al. Massbank: a public repository for sharing mass spectral data for life sciences. J Mass Spectrom 2010;45:703–14.
[16] Jeffryes JG, Broadbelt LJ, Tyo KEJ, Colastani RL, Henry CS, Elbadawi-Sidhu M, et al. Mines: open access databases of computationally predicted enzyme promiscuity products for untargeted metabolomics. J Chem 2015;7:44.
[17] Kanehisa M, Goto S. Kegg: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res 1999;28:27–30.
[18] Koelmel JP, Kroeger NM, Ulmer CZ, Bowden JA, Patterson RE, Cochran JA, et al. Lipidmatch: an automated workflow for rule-based lipid identification using untargeted high-resolution tandem mass spectrometry data. BMC Bioinformatics 2017;18:331.
[19] Kyle JE, Crowell KL, Casey C, Fujimoto GM, Kim S, Dautel SE, et al. Liquid: an-open source software for identifying lipids in lc-ms/ms-based lipidomics data. Bioinformatics 2017;33:1744–6.
[20] Milne S, Ivanova P, Forrester J, Brown HA. Lipidomics: an analysis of cellular lipids by esi-ms. Methods 2006;39:92–103.
[21] mzCloud. Mzcloud. https://www.mzcloud.org/; 2018.
[22] Peisl BYL, Schymanski EL, Wilmes P. Dark matter in host-microbiome metabolomics: tackling the unknowns–a review. Anal Chim Acta 2017;1037:13–27.
[23] Randazzo GM, Tonoli D, Strajhar P, Xenarios I, Odermatt A, Boccard J, et al. Enhanced metabolite annotation via dynamic retention time prediction: Steroidogenesis alterations as a case study. J Chromatogr B 2017;1071:11–8.
[24] Ridder L, Verhoeven S, Schaik RV, Van DH, Vervoort J, Vos RCHD. Substructure-based annotation of high-resolution multistage msn spectral trees. Rapid Commun Mass Spectrom 2012;26:2461–71.
[25] Ruttkies C, Wolf S, Neumann S, Schymanski EL, Hollender J. Metfrag relaunched: incorporating strategies beyond in silico fragmentation. J Chem 2016;8:3.
[26] Smith CA, O'Maille G, Want EJ, Qin C, Trauger SA, Brandon TR, et al. Metlin - a metabolite mass spectral database. Ther Drug Monit 2005;27:747–51.
[27] Stanstrup J, Neumann S, Vrhovsek U. Predret: prediction of retention time by direct mapping between multiple chromatographic systems. Anal Chem 2015;87: 9421–8.
[28] Sud M, Fahy E, Cotter D, Azam K, Vadivelu I, Subramaniam S, et al. Metabolomics workbench: an international repository for metabolomics data and metadata, metabolite standards, protocols, tutorials and training, and analysis tools. Nucleic Acids Res 2016;44:D463–70.
[29] Sud M, Fahy E, Cotter D, Brown A, Dennis EA, Glass CK, et al. Lmsd: lipid maps structure database. Nucleic Acids Res 2007;35:D527–32.
[30] Tsugawa H, Cajka T, Kind T, Ma Y, Higgins B, Ikeda K, et al. Ms-dial: data-independent ms/ms deconvolution for comprehensive metabolome analysis. Nat Methods 2015;12:523–6.
[31] Wishart DS, Jewison T, Guo AC, Wilson M, Knox C, Liu YF, et al. Hmdb 3.0-the human metabolome database in 2013. Nat Biotechnol 2013;41:D801–7.