AStream: an R package for annotating LC/MS metabolomic data
Arnold Alonso¹, Antonio Julià¹,*, Antoni Beltran²,³, Maria Vinaixa²,³, Marta Diaza³,⁴, Lourdes Ibañez³,⁴, Xavier Correig²,³ and Sara Marsal¹,∗
¹Grup de Recerca de Reumatologia, Institut de Recerca de l’Hospital Universitari Vall d’Hebrón (Univ. Autònoma de Barcelona), Barcelona, ²Metabolomics Platform, URV-IISPV, Tarragona, ³CIBERDEM (CIBER de Diabetes y Enfermedades Metabólicas Asociadas), ⁴Endocrinology Unit, Hospital Sant Joan de Déu (UB), Barcelona, Spain

ABSTRACT
Summary: AStream, an R-statistical software package for the curation and identification of feature peaks extracted from liquid chromatography mass spectrometry (LC/MS) metabolomics data, is described. AStream detects isotopic, fragment and adduct patterns by identifying feature pairs that fulfill expected relational patterns. Data reduction by AStream allows compounds to be identified reliably and subsequently linked to metabolite databases. AStream provides researchers with a fast, reliable tool for summarizing metabolic data, notably reducing curation time and increasing consistency of results.
Availability: The AStream R package and a study example can be freely accessed at http://www.urv.cat/AStream/AStream.html.
Contact: smarsal@ir.vhebron.net; ajulia@ir.vhebron.net
Supplementary information: Supplementary data are available at Bioinformatics online.
Received on December 16, 2010; revised on February 25, 2011; accepted on March 9, 2011

1 INTRODUCTION
Recent improvements in chromatographic technologies and spectrometry analysis techniques have increased the use of metabolomics for disease diagnosis and in systems biology (Kaddurah-Daouk et al., 2008). Liquid chromatography mass spectrometry (LC/MS) is a powerful and widely used method for identifying and quantifying metabolites in biological samples. After applying a data pre-processing algorithm, a large set of feature peaks is usually obtained for each sample; each feature is characterized by a particular retention time, mass-to-charge (m/z) value and an intensity measurement that is correlated with the relative concentration of the particular metabolite. Bioinformatic methods such as XCMS (Smith et al., 2006) can perform these complex pre-processing steps. However, once the set of feature peaks is obtained, a time-consuming and error-prone manual curation step is required, particularly when an untargeted metabolic analysis needs to be performed. The main objectives of the curation step are to reduce redundancy between the various chemical forms detected in one compound (e.g. carbon isotopes, adducts, fragments) and to identify characteristics with a higher probability of being true biological features rather than being generated by noise artifacts or potential sample contamination.

© The Author 2011. Published by Oxford University Press. All rights reserved. For Permissions, please email: journals.permissions@oup.com 1339
An LC/MS dataset obtained from 20 human serum samples was AStream summarizes them and exports them in a tabulated file. This although the presence of at least one isotopic peak is an important confidence data processing analysis had been applied using the XCMS package, only the significance value for differential expression). After raw randomization does not influence AStream metabolite annotation, analysis (Supplementary Material and Supplementary Table S1). Searching for metabolite adducts is the last analytical step of AStream. Once the features with isotopic patterns have been identified, the m/z values are compared against m/z values of all the features of the previously filtered features. In this perspective, AStream provides a default list of common adducts. However, the group of expected metabolite adducts is related to the particular LC/MS experimental setup, and the user can import their own reference list. Using prior knowledge of the expected m/z differences between protonated compounds (i.e. [M+H]+) and their corresponding adducts, AStream can identify them by searching for peaks within the expected m/z values of each adduct with respect to the m/z value of the protonated feature. The filtered feature set is created by excluding the feature pairs with low correlation and/or high retention time differences. Therefore, the feature pairs matching the expected adduct m/z differences have an increased probability of being true positives (Supplementary Figure S3). Although the presence of at least one isotopic peak is an important confidence measure to identify true biological compounds, AStream also allows the user to search for adducts without restricting the identification to the features having isotopic patterns.

In order to facilitate visualization of results relating to curated data, AStream summarizes them and exports them in a tabulated file. This results file demonstrates that AStream has fulfilled two main objectives of metabolomics data workflow: (i) reduction of the excessive redundancy present in typical metabolomics datasets; and (ii) discrimination of features with a higher probability of representing true biological compounds from those caused by noise artifacts or experimental contamination.

3 RESULTS

An LC/MS dataset obtained from 20 human serum samples was used to evaluate AStream performance (Braithwaite, X., and Correig, X., unpublished data). For illustration purposes, the samples were randomly assigned as n=10 controls and n=10 cases (of note, randomization does not influence AStream metabolite annotation, only the significance value for differential expression). After raw data processing analysis had been applied using the XCMS package, a total of n=3148 features were generated. This pre-processed metabolomic dataset is included in the AStream library. Using this dataset, AStream was used to process the large feature list into a curated, non-redundant metabolite set.

The quality control step allowed one serum sample to be identified displaying an outlying feature distribution and this was automatically discarded from further analyses (Supplementary Figure S1). In addition, the observed retention time differences and correlation histograms of all feature pairs (Supplementary Material) were used to select the optimal feature pair selection thresholds. On the basis of this information, a maximum retention time difference of 3 s and a minimum correlation value of 0.75 were used in the present analysis (Supplementary Material and Supplementary Table S1).

...