|  |
| --- |
| BIOINFORMATICS APPLICATION NOTE  **OCFS: an optimal collaborative feature selection framework for high-dimensional data**  Karan Uppal1,2,3,4, Eva K. Lee\*1,2,3  1Center for Operations Research in Medicine and HealthCare, Georgia Institute of Technology, Atlanta, GA, 30332.  2NSF I/UCRC Center for Health Organization Transformation, Georgia Institute of Technology, Atlanta, GA, 30332.  3School of Industrial and Systems Engineering, Georgia Institute of Technology, Atlanta, GA, 30332.  4School of Biology, Georgia Institute of Technology, Atlanta, GA, 30332.  Received on XXXXX; revised on XXXXX; accepted on XXXXX  Associate Editor: XXXXXXX |

[[1]](#footnote-2)\*abstract

**Motivation:** Improved analytical technologies and better data extraction algorithms in analytic chemistry and related scientific disciplines enable detection of >10,000 chemicals; There are currently limited abilities to automatically identify large numbers of chemicals in these mass spectral analyses. Here we describe an automated computational framework to select In stage 2, a binary particle swarm optimization algorithm was used to identify minimum number of genes that allowed accurate classification of the samples based on 10-fold cross validation criteria using support vector machine a classifier implemented in the R e1071 package.

Availability: \*Insert URL for sourceforge

# introduction

Biomarker screening and discovery are key components of biomedical research. Most omics technologies measure thousands of variables (genes, metabolites, SNPs, etc.) and often fall under the category of n<<p problems that are prone to model over-fitting. The large amount of information requires application of data mining and technical techniques to identify most salient variables related to the question of interest. In general, the likelihood of erroneous or sub-optimal results is higher as the amount of information. Selection of relevant genes for sample classification is a common task in most high-throughput omics studies, where researchers try to identify the smallest possible set of genes that can still achieve good predictive performance. This is crucial for targeted validation experiments, designing follow-up studies, and for diagnostic purposes in clinical practice.

There are three different feature selection schemes as shown in Figure 1. that are generally used for microarray data analysis. The Filter method uses some statistical criteria such as p-value from t-test to select relevant genes and these genes are then used to build the model. The wrapper method such as binary Particle Swarm Optimization uses different combinations of subsets of features and selects the best model. The embedded methods such as recursive feature elimination based on SVM and random forest, LASSO, and ridge regression have built-in variable selection. For instance, LASSO and Ridge Regression that use a penalty function and gives a value of greater than 0 if a feature is relevant, and 0 otherwise.

Many feature selection approaches use arbitrary rank or significance thresholds to select the number of genes that could lead to irreproducible results (Nature Methods 2015). Moreover, different algorithms vary differently in performance depending on the distribution of the data and within-class variability. Advanced variable selection methods recursive feature elimination SVM and random forest are well suited for microarray data: they show excellent performance even when most predictive variables are noise, can be used when the number of variables is much larger than the number of observations and in problems involving more than two classes [1]. It is important to determine which algorithm is best suited for selecting the best set of genes from a gene expression dataset.

The main goal of our project was to compare the performance of different feature selection algorithms based on the criteria of minimum number of genes selected and maxmum classification accuracy.

Particle Swarm Optimization (PSO) is a stochastic optimization technique based on the movement and intelligence of swarms developed by James Kennedy and Russell Eberhart in 1995. It comprises of a number of agents/particles that constitute a swarm moving around in the search space looking for the best solution. Every individual moves according to its own experience as well as the experience of others. Several studies have reported the application of using a binary PSO for variable selection (References). Here we introduce a novel ensemble feature selection framework that uses a binary behavior based PSO (B3PSO) to combine the results from different selection methods.

# methods

We focused on comparing the performance of F-test, T-test, Kruskal test, Elastic net, Lasso, rfe-SVM, Random Forest, and Binary PSO in this study. The algorithm uses the internal cross-validation scheme that performs variable selection based on the every training set, traink, in the k-fold scheme and uses the left-out subset, testk, for evaluating the performance of the model. The internal cross-validation scheme is shown to out-perform the commonly used external cross-validation scheme where the model evaluation is performed after the selection using all samples. This leads to a biased estimate of the model performance (Reference).

**1) Pattern Selection based on classification** (implemented in R):

A new feature selection algorithm based on a consensus scoring methodology was implemented. The algorithm is divided into three steps. The first two steps use the CMA package implemented in R, and the last step uses a particle swarm optimization based search strategy to find the optimal set of features that would allow maximum classification accuracy.

The user uploaded dataset is first randomly divided into distinct training and test sets based on the splitting percentage provided by the user (default: 60% train, 40%test).

In the **first step** of the feature selection process, a ranked list of features is generated using one or more feature selection algorithms selected by the user. The user can select t.test, f.test, kruskal.test, recursive feature elimination, random forest, Welch.test, Wilcox.test, lasso, elasticnet. The number of features selected impacts the performance of the classifiers and the predictive accuracy (Reunanbam 2003, Nijim and Kuhaar 2006). It is essential to find the optimal set of features.

A sequential backward or forward selection scheme is used to select the optimal set of features. For every selected algorithm, the number of features is iteratively increased from 1 to X (default 150) in increments of delta (default 5) till there is no improvement in the 10fold cross validation accuracy for I number of iterations (default 10).

This gives m subsets of top ranked features where m is the number of selected feature selection algorithms. A consensus score of a feature is defined as the number of subsets that include that feature, and its value ranges from 0 to m. Only the top X number of features (default 150; user defined) are retained for further analysis. . A gene was identified as being differentially expressed if it was selected by at least two methods. The consensus approach increases the confidence in the gene selection process as some methods are too stringent while others too lenient.

In the **second** stage, the consensus subset of features selected by at least M\_min (default=100%) algorithms is then used as an input for a novel behavior based binary PSO algorithm, B3PSO. The PSO algorithm is designed to avoid particles from getting stuck in local optima. The search criteria (local or global search) depends on the rank of a particle. In other words, the best particle will have a narrow or local search space while the worst particle will have a much broader or a global search space controlled by the inertia variable in the velocity update equation of the PSO algorithm.

B3PSO Algorithm:

Basic idea: Particles x[i][j] are 2D arrays where index i is the particle and index j is the feature. If x[i][j]=1, then the feature j is selected, otherwise it is not selected.

Initializaiton

1)Initially every particle xi has p randomly selected features where p<d

2) After time step 1, the features of a particle are set to 0, or 1 depending on its velocity. If the S measure <rand() then feature is not selected, otherwise it is selected.

3)

Each particle chooses a topology based on their behavior.

The PSO algorithm is designed to avoid particles from getting stuck in local optima. The search criteria (local or global search) depends on the rank of a particle. In other words, the best particle will have a narrow or local search space while the worst particle will have a much broader or a global search space controlled by the inertia variable in the velocity update equation of the PSO algorithm. Get the consensus features that are selected in each set.

We used the Classification for Microarrays (CMA) package - a comprehensive Bioconductor package for supervised classification with high dimensional data [2]. The package has built-in functions for performing feature selection using F-test, T-test, Elasticnet, Lasso, rfeSVM, Kruskal test, and Random Forest. We performed feature selection and cross-validation using the knn classifier from CMA package. The function has in-built feature to tune the value of k to be used.

# results

The performance of the scoring algorithm was evaluated using three publicly available gene-expression datasets for our study. The three datasets varied in their level of complexity. The NCI60 dataset had 60 training samples belonging to 9 classes, but there was no test data, and we could not split the dataset due to very few samples. The SRBCT dataset and Leukemia dataset had both training and test data.

All datasets had number of genes >> number of samples. Table 1 summarizes the details of the three datasets. The results show that the ensemble approach improves the classification results as compared to individual methods.

# DISCUSSION

# CONCLUSION

It can be concluded that the embedded feature selection algorithms perform better than the filter and wrapper methods. Rfe-SVM performed best on the Leukemia and NCI60 dataset, but Lasso gave better performance on the SRBCT dataset. Thus, the performance of a feature selection algorithm depends on the dataset being used, the number of samples in the train and test dataset as well as the number of classes.

Most feature selection algorithms (mainly the filter and wrapper methods) generate models that over-fit to the training data, and thus generalize poorly. Recursive Feature Elimination- SVM method and Lasso achieves better performance because they minimize over-fitting by penalizing the irrelevant features, or removing them at every iteration.

More samples and better classification algorithms are required for building robust models for large-class datasets like NCI60. A comparison of different classification algorithms should be considered for building the trained model.

**Fig. 1.**Scoring algorithm workflow

**Table 1.**Benchmark results

b) Binary PSO

**For n iterations:**

Initialize the swarm with N particles by randomly selecting p features for every particle

For every particle, evaluate the fitness function:

Build SVM model using the train data (140 samples)

Evaluate the model using the validation data (70 samples) and determine the classification accuracy

Determine the local best and global best particle that gives highest classification accuracy

Update the velocity of every particle,



Update the position of every particle,

S=1/(1+exp(-vi))

xi =1 if S>random number

0 otherwise

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| **Feature Selection Algorithm** | **Accuracy (%) using the top n=3 genes** | **Accuracy (%) using the top n=5 genes** | **Accuracy (%) using the top n=10 genes** | **Accuracy (%) using the top n=15 genes** | **Accuracy (%) using the top n=25 genes** | **Accuracy (%) using the top n=50 genes** |
| Elasticnet | 50 | 60 | 55 | 60 | 60 | 65 |
| Ttest | 60 | 50 | 50 | 65 | 60 | 60 |
| rfe-svm | 40 | 35 | 55 | 70 | 50 | 50 |
| Random Forest | 50 | 45 | 60 | 75 | 55 | 70 |
| F-test | 50 | 50 | 70 | 68 | 55 | 55 |
| Kruskal | 50 | 50 | 40 | 60 | 60 | 65 |
| Lasso | 70 | 75 | 50 | 60 | 60 | 65 |

acknowledgements

*Funding*:

References

Eg: Alexandrescu,A. (2001) Modern C++ Design: Generic Programming and Design

# [Patti GJ](http://www.ncbi.nlm.nih.gov/pubmed?term=Patti%20GJ%5BAuthor%5D&cauthor=true&cauthor_uid=22343432), [Tautenhahn R](http://www.ncbi.nlm.nih.gov/pubmed?term=Tautenhahn%20R%5BAuthor%5D&cauthor=true&cauthor_uid=22343432), [Siuzdak G](http://www.ncbi.nlm.nih.gov/pubmed?term=Siuzdak%20G%5BAuthor%5D&cauthor=true&cauthor_uid=22343432). Meta-analysis of untargeted metabolomic data from multiple profiling experiments. [Nat Protoc.](http://www.ncbi.nlm.nih.gov/pubmed/22343432) 2012 Feb 16;7(3):508-16. doi: 10.1038/nprot.2011.454.

# [Wishart DS](http://www.ncbi.nlm.nih.gov/pubmed?term=Wishart%20DS%5BAuthor%5D&cauthor=true&cauthor_uid=23161693), [Jewison T](http://www.ncbi.nlm.nih.gov/pubmed?term=Jewison%20T%5BAuthor%5D&cauthor=true&cauthor_uid=23161693), [Guo AC](http://www.ncbi.nlm.nih.gov/pubmed?term=Guo%20AC%5BAuthor%5D&cauthor=true&cauthor_uid=23161693), [Wilson M](http://www.ncbi.nlm.nih.gov/pubmed?term=Wilson%20M%5BAuthor%5D&cauthor=true&cauthor_uid=23161693), [Knox C](http://www.ncbi.nlm.nih.gov/pubmed?term=Knox%20C%5BAuthor%5D&cauthor=true&cauthor_uid=23161693), [Liu Y](http://www.ncbi.nlm.nih.gov/pubmed?term=Liu%20Y%5BAuthor%5D&cauthor=true&cauthor_uid=23161693), [Djoumbou Y](http://www.ncbi.nlm.nih.gov/pubmed?term=Djoumbou%20Y%5BAuthor%5D&cauthor=true&cauthor_uid=23161693), [Mandal R](http://www.ncbi.nlm.nih.gov/pubmed?term=Mandal%20R%5BAuthor%5D&cauthor=true&cauthor_uid=23161693), [Aziat F](http://www.ncbi.nlm.nih.gov/pubmed?term=Aziat%20F%5BAuthor%5D&cauthor=true&cauthor_uid=23161693), [Dong E](http://www.ncbi.nlm.nih.gov/pubmed?term=Dong%20E%5BAuthor%5D&cauthor=true&cauthor_uid=23161693), [Bouatra S](http://www.ncbi.nlm.nih.gov/pubmed?term=Bouatra%20S%5BAuthor%5D&cauthor=true&cauthor_uid=23161693), [Sinelnikov I](http://www.ncbi.nlm.nih.gov/pubmed?term=Sinelnikov%20I%5BAuthor%5D&cauthor=true&cauthor_uid=23161693), [Arndt D](http://www.ncbi.nlm.nih.gov/pubmed?term=Arndt%20D%5BAuthor%5D&cauthor=true&cauthor_uid=23161693), [Xia J](http://www.ncbi.nlm.nih.gov/pubmed?term=Xia%20J%5BAuthor%5D&cauthor=true&cauthor_uid=23161693), [Liu P](http://www.ncbi.nlm.nih.gov/pubmed?term=Liu%20P%5BAuthor%5D&cauthor=true&cauthor_uid=23161693), [Yallou F](http://www.ncbi.nlm.nih.gov/pubmed?term=Yallou%20F%5BAuthor%5D&cauthor=true&cauthor_uid=23161693), [Bjorndahl T](http://www.ncbi.nlm.nih.gov/pubmed?term=Bjorndahl%20T%5BAuthor%5D&cauthor=true&cauthor_uid=23161693), [Perez-Pineiro R](http://www.ncbi.nlm.nih.gov/pubmed?term=Perez-Pineiro%20R%5BAuthor%5D&cauthor=true&cauthor_uid=23161693), [Eisner R](http://www.ncbi.nlm.nih.gov/pubmed?term=Eisner%20R%5BAuthor%5D&cauthor=true&cauthor_uid=23161693), [Allen F](http://www.ncbi.nlm.nih.gov/pubmed?term=Allen%20F%5BAuthor%5D&cauthor=true&cauthor_uid=23161693), [Neveu V](http://www.ncbi.nlm.nih.gov/pubmed?term=Neveu%20V%5BAuthor%5D&cauthor=true&cauthor_uid=23161693), [Greiner R](http://www.ncbi.nlm.nih.gov/pubmed?term=Greiner%20R%5BAuthor%5D&cauthor=true&cauthor_uid=23161693), [Scalbert A](http://www.ncbi.nlm.nih.gov/pubmed?term=Scalbert%20A%5BAuthor%5D&cauthor=true&cauthor_uid=23161693). HMDB 3.0--The Human Metabolome Database in 2013. [Nucleic Acids Res.](http://www.ncbi.nlm.nih.gov/pubmed/23161693) 2013 Jan;41(Database issue):D801-7. doi: 10.1093/nar/gks1065. Epub 2012 Nov 17.

# [Wishart DS](http://www.ncbi.nlm.nih.gov/pubmed?term=Wishart%20DS%5BAuthor%5D&cauthor=true&cauthor_uid=21827274). Advances in metabolite identification. [Bioanalysis.](http://www.ncbi.nlm.nih.gov/pubmed/21827274) 2011 Aug;3(15):1769-82. doi: 10.4155/bio.11.155.

# <http://www.ncbi.nlm.nih.gov/pubmed/21827274>

# [Vinayavekhin N](http://www.ncbi.nlm.nih.gov/pubmed?term=Vinayavekhin%20N%5BAuthor%5D&cauthor=true&cauthor_uid=20373502), [Saghatelian A](http://www.ncbi.nlm.nih.gov/pubmed?term=Saghatelian%20A%5BAuthor%5D&cauthor=true&cauthor_uid=20373502). Untargeted metabolomics. [Curr Protoc Mol Biol.](http://www.ncbi.nlm.nih.gov/pubmed/20373502) 2010 Apr;Chapter 30:Unit 30.1.1-24. doi: 10.1002/0471142727.mb3001s90

# [Tautenhahn R](http://www.ncbi.nlm.nih.gov/pubmed?term=Tautenhahn%20R%5BAuthor%5D&cauthor=true&cauthor_uid=21174458), [Patti GJ](http://www.ncbi.nlm.nih.gov/pubmed?term=Patti%20GJ%5BAuthor%5D&cauthor=true&cauthor_uid=21174458), [Kalisiak E](http://www.ncbi.nlm.nih.gov/pubmed?term=Kalisiak%20E%5BAuthor%5D&cauthor=true&cauthor_uid=21174458), [Miyamoto T](http://www.ncbi.nlm.nih.gov/pubmed?term=Miyamoto%20T%5BAuthor%5D&cauthor=true&cauthor_uid=21174458), [Schmidt M](http://www.ncbi.nlm.nih.gov/pubmed?term=Schmidt%20M%5BAuthor%5D&cauthor=true&cauthor_uid=21174458), [Lo FY](http://www.ncbi.nlm.nih.gov/pubmed?term=Lo%20FY%5BAuthor%5D&cauthor=true&cauthor_uid=21174458), [McBee J](http://www.ncbi.nlm.nih.gov/pubmed?term=McBee%20J%5BAuthor%5D&cauthor=true&cauthor_uid=21174458), [Baliga NS](http://www.ncbi.nlm.nih.gov/pubmed?term=Baliga%20NS%5BAuthor%5D&cauthor=true&cauthor_uid=21174458), [Siuzdak G](http://www.ncbi.nlm.nih.gov/pubmed?term=Siuzdak%20G%5BAuthor%5D&cauthor=true&cauthor_uid=21174458). metaXCMS: second-order analysis of untargeted metabolomics data. [Anal Chem.](http://www.ncbi.nlm.nih.gov/pubmed/21174458) 2011 Feb 1;83(3):696-700. doi: 10.1021/ac102980g. Epub 2010 Dec 21.

# [Xiao JF](http://www.ncbi.nlm.nih.gov/pubmed?term=Xiao%20JF%5BAuthor%5D&cauthor=true&cauthor_uid=22345829), [Zhou B](http://www.ncbi.nlm.nih.gov/pubmed?term=Zhou%20B%5BAuthor%5D&cauthor=true&cauthor_uid=22345829), [Ressom HW](http://www.ncbi.nlm.nih.gov/pubmed?term=Ressom%20HW%5BAuthor%5D&cauthor=true&cauthor_uid=22345829). Metabolite identification and quantitation in LC-MS/MS-based metabolomics. [Trends Analyt Chem.](http://www.ncbi.nlm.nih.gov/pubmed/22345829) 2012 Feb 1;32:1-14.

# <http://www.ncbi.nlm.nih.gov/pubmed/22345829>

# [Zhou B](http://www.ncbi.nlm.nih.gov/pubmed?term=Zhou%20B%5BAuthor%5D&cauthor=true&cauthor_uid=22768229), [Wang J](http://www.ncbi.nlm.nih.gov/pubmed?term=Wang%20J%5BAuthor%5D&cauthor=true&cauthor_uid=22768229), [Ressom HW](http://www.ncbi.nlm.nih.gov/pubmed?term=Ressom%20HW%5BAuthor%5D&cauthor=true&cauthor_uid=22768229). MetaboSearch: tool for mass-based metabolite identification using multiple databases. [PLoS One.](http://www.ncbi.nlm.nih.gov/pubmed/22768229) 2012;7(6):e40096. doi: 10.1371/journal.pone.0040096. Epub 2012 Jun 29.

# <http://www.ncbi.nlm.nih.gov/pubmed/22768229>

# [Tohge T](http://www.ncbi.nlm.nih.gov/pubmed?term=Tohge%20T%5BAuthor%5D&cauthor=true&cauthor_uid=19285697), [Fernie AR](http://www.ncbi.nlm.nih.gov/pubmed?term=Fernie%20AR%5BAuthor%5D&cauthor=true&cauthor_uid=19285697). Web-based resources for mass-spectrometry-based metabolomics: a user’s guide. [Phytochemistry.](http://www.ncbi.nlm.nih.gov/pubmed/19285697) 2009 Mar;70(4):450-6. doi: 10.1016/j.phytochem.2009.02.004. Epub 2009 Mar 13.

# <http://www.ncbi.nlm.nih.gov/pubmed/19285697>

# [Krumsiek J](http://www.ncbi.nlm.nih.gov/pubmed?term=Krumsiek%20J%5BAuthor%5D&cauthor=true&cauthor_uid=23093944), [Suhre K](http://www.ncbi.nlm.nih.gov/pubmed?term=Suhre%20K%5BAuthor%5D&cauthor=true&cauthor_uid=23093944), [Evans AM](http://www.ncbi.nlm.nih.gov/pubmed?term=Evans%20AM%5BAuthor%5D&cauthor=true&cauthor_uid=23093944), [Mitchell MW](http://www.ncbi.nlm.nih.gov/pubmed?term=Mitchell%20MW%5BAuthor%5D&cauthor=true&cauthor_uid=23093944), [Mohney RP](http://www.ncbi.nlm.nih.gov/pubmed?term=Mohney%20RP%5BAuthor%5D&cauthor=true&cauthor_uid=23093944), [Milburn MV](http://www.ncbi.nlm.nih.gov/pubmed?term=Milburn%20MV%5BAuthor%5D&cauthor=true&cauthor_uid=23093944), [Wägele B](http://www.ncbi.nlm.nih.gov/pubmed?term=W%C3%A4gele%20B%5BAuthor%5D&cauthor=true&cauthor_uid=23093944), [Römisch-Margl W](http://www.ncbi.nlm.nih.gov/pubmed?term=R%C3%B6misch-Margl%20W%5BAuthor%5D&cauthor=true&cauthor_uid=23093944), [Illig T](http://www.ncbi.nlm.nih.gov/pubmed?term=Illig%20T%5BAuthor%5D&cauthor=true&cauthor_uid=23093944), [Adamski J](http://www.ncbi.nlm.nih.gov/pubmed?term=Adamski%20J%5BAuthor%5D&cauthor=true&cauthor_uid=23093944), [Gieger C](http://www.ncbi.nlm.nih.gov/pubmed?term=Gieger%20C%5BAuthor%5D&cauthor=true&cauthor_uid=23093944), [Theis FJ](http://www.ncbi.nlm.nih.gov/pubmed?term=Theis%20FJ%5BAuthor%5D&cauthor=true&cauthor_uid=23093944), [Kastenmüller G](http://www.ncbi.nlm.nih.gov/pubmed?term=Kastenm%C3%BCller%20G%5BAuthor%5D&cauthor=true&cauthor_uid=23093944). Mining the unknown: a systems approach to metabolite identification combining genetic and metabolic information. [PLoS Genet.](http://www.ncbi.nlm.nih.gov/pubmed/23093944) 2012;8(10):e1003005. doi: 10.1371/journal.pgen.1003005. Epub 2012 Oct 18.

# <http://www.ncbi.nlm.nih.gov/pubmed/23093944>

# [Williams AJ](http://www.ncbi.nlm.nih.gov/pubmed?term=Williams%20AJ%5BAuthor%5D&cauthor=true&cauthor_uid=18428094). Public chemical compound databases. [Curr Opin Drug Discov Devel.](http://www.ncbi.nlm.nih.gov/pubmed/18428094) 2008 May;11(3):393-404.

# <http://www.ncbi.nlm.nih.gov/pubmed/18428094>

# [Stanstrup J](http://www.ncbi.nlm.nih.gov/pubmed?term=Stanstrup%20J%5BAuthor%5D&cauthor=true&cauthor_uid=23615935), [Gerlich M](http://www.ncbi.nlm.nih.gov/pubmed?term=Gerlich%20M%5BAuthor%5D&cauthor=true&cauthor_uid=23615935), [Dragsted LO](http://www.ncbi.nlm.nih.gov/pubmed?term=Dragsted%20LO%5BAuthor%5D&cauthor=true&cauthor_uid=23615935), [Neumann S](http://www.ncbi.nlm.nih.gov/pubmed?term=Neumann%20S%5BAuthor%5D&cauthor=true&cauthor_uid=23615935). Metabolite profiling and beyond: approaches for the rapid processing and annotation of human blood serum mass spectrometry data. [Anal Bioanal Chem.](http://www.ncbi.nlm.nih.gov/pubmed/?term=xcms+camera) 2013 Jun;405(15):5037-48. doi: 10.1007/s00216-013-6954-6. Epub 2013 Apr 25.

# [Draper J](http://www.ncbi.nlm.nih.gov/pubmed?term=Draper%20J%5BAuthor%5D&cauthor=true&cauthor_uid=19622150), [Enot DP](http://www.ncbi.nlm.nih.gov/pubmed?term=Enot%20DP%5BAuthor%5D&cauthor=true&cauthor_uid=19622150), [Parker D](http://www.ncbi.nlm.nih.gov/pubmed?term=Parker%20D%5BAuthor%5D&cauthor=true&cauthor_uid=19622150), [Beckmann M](http://www.ncbi.nlm.nih.gov/pubmed?term=Beckmann%20M%5BAuthor%5D&cauthor=true&cauthor_uid=19622150), [Snowdon S](http://www.ncbi.nlm.nih.gov/pubmed?term=Snowdon%20S%5BAuthor%5D&cauthor=true&cauthor_uid=19622150), [Lin W](http://www.ncbi.nlm.nih.gov/pubmed?term=Lin%20W%5BAuthor%5D&cauthor=true&cauthor_uid=19622150), [Zubair H](http://www.ncbi.nlm.nih.gov/pubmed?term=Zubair%20H%5BAuthor%5D&cauthor=true&cauthor_uid=19622150). Metabolite signal identification in accurate mass metabolomics data with MZedDB, an interactive m/z annotation tool utilising predicted ionisation behaviour ‘rules’. [BMC Bioinformatics.](http://www.ncbi.nlm.nih.gov/pubmed/19622150) 2009 Jul 21;10:227. doi: 10.1186/1471-2105-10-227.

# [Roux A](http://www.ncbi.nlm.nih.gov/pubmed?term=Roux%20A%5BAuthor%5D&cauthor=true&cauthor_uid=22770225), [Xu Y](http://www.ncbi.nlm.nih.gov/pubmed?term=Xu%20Y%5BAuthor%5D&cauthor=true&cauthor_uid=22770225), [Heilier JF](http://www.ncbi.nlm.nih.gov/pubmed?term=Heilier%20JF%5BAuthor%5D&cauthor=true&cauthor_uid=22770225), [Olivier MF](http://www.ncbi.nlm.nih.gov/pubmed?term=Olivier%20MF%5BAuthor%5D&cauthor=true&cauthor_uid=22770225), [Ezan E](http://www.ncbi.nlm.nih.gov/pubmed?term=Ezan%20E%5BAuthor%5D&cauthor=true&cauthor_uid=22770225), [Tabet JC](http://www.ncbi.nlm.nih.gov/pubmed?term=Tabet%20JC%5BAuthor%5D&cauthor=true&cauthor_uid=22770225), [Junot C](http://www.ncbi.nlm.nih.gov/pubmed?term=Junot%20C%5BAuthor%5D&cauthor=true&cauthor_uid=22770225). Annotation of the human adult urinary metabolome and metabolite identification using ultra high performance liquid chromatography coupled to a linear quadrupole ion trap-Orbitrap mass spectrometer. [Anal Chem.](http://www.ncbi.nlm.nih.gov/pubmed/22770225) 2012 Aug 7;84(15):6429-37. doi: 10.1021/ac300829f. Epub 2012 Jul 17.

# [Neumann S](http://www.ncbi.nlm.nih.gov/pubmed?term=Neumann%20S%5BAuthor%5D&cauthor=true&cauthor_uid=20936272), [Böcker S](http://www.ncbi.nlm.nih.gov/pubmed?term=B%C3%B6cker%20S%5BAuthor%5D&cauthor=true&cauthor_uid=20936272).Computational mass spectrometry for metabolomics: identification of metabolites and small molecules. [Anal Bioanal Chem.](http://www.ncbi.nlm.nih.gov/pubmed/20936272) 2010 Dec;398(7-8):2779-88. doi: 10.1007/s00216-010-4142-5. Epub 2010 Oct 9.

# [Bowen BP](http://www.ncbi.nlm.nih.gov/pubmed?term=Bowen%20BP%5BAuthor%5D&cauthor=true&cauthor_uid=20452782), [Northen TR](http://www.ncbi.nlm.nih.gov/pubmed?term=Northen%20TR%5BAuthor%5D&cauthor=true&cauthor_uid=20452782). Dealing with the unknown: metabolomics and metabolite atlases. [J Am Soc Mass Spectrom.](http://www.ncbi.nlm.nih.gov/pubmed/20452782) 2010 Sep;21(9):1471-6. doi: 10.1016/j.jasms.2010.04.003. Epub 2010 Apr 12.

# [Gerlich M](http://www.ncbi.nlm.nih.gov/pubmed?term=Gerlich%20M%5BAuthor%5D&cauthor=true&cauthor_uid=23494783), [Neumann S](http://www.ncbi.nlm.nih.gov/pubmed?term=Neumann%20S%5BAuthor%5D&cauthor=true&cauthor_uid=23494783). MetFusion: integration of compound identification strategies. [J Mass Spectrom.](http://www.ncbi.nlm.nih.gov/pubmed/23494783) 2013 Mar;48(3):291-8. doi: 10.1002/jms.3123.

# [Horai H](http://www.ncbi.nlm.nih.gov/pubmed?term=Horai%20H%5BAuthor%5D&cauthor=true&cauthor_uid=20623627), [Arita M](http://www.ncbi.nlm.nih.gov/pubmed?term=Arita%20M%5BAuthor%5D&cauthor=true&cauthor_uid=20623627), [Kanaya S](http://www.ncbi.nlm.nih.gov/pubmed?term=Kanaya%20S%5BAuthor%5D&cauthor=true&cauthor_uid=20623627), [Nihei Y](http://www.ncbi.nlm.nih.gov/pubmed?term=Nihei%20Y%5BAuthor%5D&cauthor=true&cauthor_uid=20623627), [Ikeda T](http://www.ncbi.nlm.nih.gov/pubmed?term=Ikeda%20T%5BAuthor%5D&cauthor=true&cauthor_uid=20623627), [Suwa K](http://www.ncbi.nlm.nih.gov/pubmed?term=Suwa%20K%5BAuthor%5D&cauthor=true&cauthor_uid=20623627), [Ojima Y](http://www.ncbi.nlm.nih.gov/pubmed?term=Ojima%20Y%5BAuthor%5D&cauthor=true&cauthor_uid=20623627), [Tanaka K](http://www.ncbi.nlm.nih.gov/pubmed?term=Tanaka%20K%5BAuthor%5D&cauthor=true&cauthor_uid=20623627), [Tanaka S](http://www.ncbi.nlm.nih.gov/pubmed?term=Tanaka%20S%5BAuthor%5D&cauthor=true&cauthor_uid=20623627), [Aoshima K](http://www.ncbi.nlm.nih.gov/pubmed?term=Aoshima%20K%5BAuthor%5D&cauthor=true&cauthor_uid=20623627), [Oda Y](http://www.ncbi.nlm.nih.gov/pubmed?term=Oda%20Y%5BAuthor%5D&cauthor=true&cauthor_uid=20623627), [Kakazu Y](http://www.ncbi.nlm.nih.gov/pubmed?term=Kakazu%20Y%5BAuthor%5D&cauthor=true&cauthor_uid=20623627), [Kusano M](http://www.ncbi.nlm.nih.gov/pubmed?term=Kusano%20M%5BAuthor%5D&cauthor=true&cauthor_uid=20623627), [Tohge T](http://www.ncbi.nlm.nih.gov/pubmed?term=Tohge%20T%5BAuthor%5D&cauthor=true&cauthor_uid=20623627), [Matsuda F](http://www.ncbi.nlm.nih.gov/pubmed?term=Matsuda%20F%5BAuthor%5D&cauthor=true&cauthor_uid=20623627), [Sawada Y](http://www.ncbi.nlm.nih.gov/pubmed?term=Sawada%20Y%5BAuthor%5D&cauthor=true&cauthor_uid=20623627), [Hirai MY](http://www.ncbi.nlm.nih.gov/pubmed?term=Hirai%20MY%5BAuthor%5D&cauthor=true&cauthor_uid=20623627), [Nakanishi H](http://www.ncbi.nlm.nih.gov/pubmed?term=Nakanishi%20H%5BAuthor%5D&cauthor=true&cauthor_uid=20623627), [Ikeda K](http://www.ncbi.nlm.nih.gov/pubmed?term=Ikeda%20K%5BAuthor%5D&cauthor=true&cauthor_uid=20623627), [Akimoto N](http://www.ncbi.nlm.nih.gov/pubmed?term=Akimoto%20N%5BAuthor%5D&cauthor=true&cauthor_uid=20623627), [Maoka T](http://www.ncbi.nlm.nih.gov/pubmed?term=Maoka%20T%5BAuthor%5D&cauthor=true&cauthor_uid=20623627), [Takahashi H](http://www.ncbi.nlm.nih.gov/pubmed?term=Takahashi%20H%5BAuthor%5D&cauthor=true&cauthor_uid=20623627), [Ara T](http://www.ncbi.nlm.nih.gov/pubmed?term=Ara%20T%5BAuthor%5D&cauthor=true&cauthor_uid=20623627), [Sakurai N](http://www.ncbi.nlm.nih.gov/pubmed?term=Sakurai%20N%5BAuthor%5D&cauthor=true&cauthor_uid=20623627), [Suzuki H](http://www.ncbi.nlm.nih.gov/pubmed?term=Suzuki%20H%5BAuthor%5D&cauthor=true&cauthor_uid=20623627), [Shibata D](http://www.ncbi.nlm.nih.gov/pubmed?term=Shibata%20D%5BAuthor%5D&cauthor=true&cauthor_uid=20623627), [Neumann S](http://www.ncbi.nlm.nih.gov/pubmed?term=Neumann%20S%5BAuthor%5D&cauthor=true&cauthor_uid=20623627), [Iida T](http://www.ncbi.nlm.nih.gov/pubmed?term=Iida%20T%5BAuthor%5D&cauthor=true&cauthor_uid=20623627), [Tanaka K](http://www.ncbi.nlm.nih.gov/pubmed?term=Tanaka%20K%5BAuthor%5D&cauthor=true&cauthor_uid=20623627), [Funatsu K](http://www.ncbi.nlm.nih.gov/pubmed?term=Funatsu%20K%5BAuthor%5D&cauthor=true&cauthor_uid=20623627), [Matsuura F](http://www.ncbi.nlm.nih.gov/pubmed?term=Matsuura%20F%5BAuthor%5D&cauthor=true&cauthor_uid=20623627), [Soga T](http://www.ncbi.nlm.nih.gov/pubmed?term=Soga%20T%5BAuthor%5D&cauthor=true&cauthor_uid=20623627), [Taguchi R](http://www.ncbi.nlm.nih.gov/pubmed?term=Taguchi%20R%5BAuthor%5D&cauthor=true&cauthor_uid=20623627), [Saito K](http://www.ncbi.nlm.nih.gov/pubmed?term=Saito%20K%5BAuthor%5D&cauthor=true&cauthor_uid=20623627), [Nishioka T](http://www.ncbi.nlm.nih.gov/pubmed?term=Nishioka%20T%5BAuthor%5D&cauthor=true&cauthor_uid=20623627). MassBank: a public repository for sharing mass spectral data for life sciences. [J Mass Spectrom.](http://www.ncbi.nlm.nih.gov/pubmed/20623627) 2010 Jul;45(7):703-14. doi: 10.1002/jms.1777.

# [Tautenhahn R](http://www.ncbi.nlm.nih.gov/pubmed?term=Tautenhahn%20R%5BAuthor%5D&cauthor=true&cauthor_uid=22965049), [Cho K](http://www.ncbi.nlm.nih.gov/pubmed?term=Cho%20K%5BAuthor%5D&cauthor=true&cauthor_uid=22965049), [Uritboonthai W](http://www.ncbi.nlm.nih.gov/pubmed?term=Uritboonthai%20W%5BAuthor%5D&cauthor=true&cauthor_uid=22965049), [Zhu Z](http://www.ncbi.nlm.nih.gov/pubmed?term=Zhu%20Z%5BAuthor%5D&cauthor=true&cauthor_uid=22965049), [Patti GJ](http://www.ncbi.nlm.nih.gov/pubmed?term=Patti%20GJ%5BAuthor%5D&cauthor=true&cauthor_uid=22965049), [Siuzdak G](http://www.ncbi.nlm.nih.gov/pubmed?term=Siuzdak%20G%5BAuthor%5D&cauthor=true&cauthor_uid=22965049). An accelerated workflow for untargeted metabolomics using the METLIN database. [Nat Biotechnol.](http://www.ncbi.nlm.nih.gov/pubmed/22965049) 2012 Sep;30(9):826-8. doi: 10.1038/nbt.2348.

# [Biswas A](http://www.ncbi.nlm.nih.gov/pubmed?term=Biswas%20A%5BAuthor%5D&cauthor=true&cauthor_uid=20702401), [Mynampati KC](http://www.ncbi.nlm.nih.gov/pubmed?term=Mynampati%20KC%5BAuthor%5D&cauthor=true&cauthor_uid=20702401), [Umashankar S](http://www.ncbi.nlm.nih.gov/pubmed?term=Umashankar%20S%5BAuthor%5D&cauthor=true&cauthor_uid=20702401), [Reuben S](http://www.ncbi.nlm.nih.gov/pubmed?term=Reuben%20S%5BAuthor%5D&cauthor=true&cauthor_uid=20702401), [Parab G](http://www.ncbi.nlm.nih.gov/pubmed?term=Parab%20G%5BAuthor%5D&cauthor=true&cauthor_uid=20702401), [Rao R](http://www.ncbi.nlm.nih.gov/pubmed?term=Rao%20R%5BAuthor%5D&cauthor=true&cauthor_uid=20702401), [Kannan VS](http://www.ncbi.nlm.nih.gov/pubmed?term=Kannan%20VS%5BAuthor%5D&cauthor=true&cauthor_uid=20702401), [Swarup S](http://www.ncbi.nlm.nih.gov/pubmed?term=Swarup%20S%5BAuthor%5D&cauthor=true&cauthor_uid=20702401). MetDAT: a modular and workflow-based free online pipeline for mass spectrometry data processing, analysis and interpretation. [Bioinformatics.](http://www.ncbi.nlm.nih.gov/pubmed/20702401) 2010 Oct 15;26(20):2639-40. doi: 10.1093/bioinformatics/btq436. Epub 2010 Aug 11.

# [Sugimoto M](http://www.ncbi.nlm.nih.gov/pubmed?term=Sugimoto%20M%5BAuthor%5D&cauthor=true&cauthor_uid=22438836), [Kawakami M](http://www.ncbi.nlm.nih.gov/pubmed?term=Kawakami%20M%5BAuthor%5D&cauthor=true&cauthor_uid=22438836), [Robert M](http://www.ncbi.nlm.nih.gov/pubmed?term=Robert%20M%5BAuthor%5D&cauthor=true&cauthor_uid=22438836), [Soga T](http://www.ncbi.nlm.nih.gov/pubmed?term=Soga%20T%5BAuthor%5D&cauthor=true&cauthor_uid=22438836), [Tomita M](http://www.ncbi.nlm.nih.gov/pubmed?term=Tomita%20M%5BAuthor%5D&cauthor=true&cauthor_uid=22438836). Bioinformatics Tools for Mass Spectroscopy-Based Metabolomic Data Processing and Analysis. [Curr Bioinform.](http://www.ncbi.nlm.nih.gov/pubmed/22438836) 2012 Mar;7(1):96-108.

# R Development Core Team. *R: a language and environment for statistical computing.* *Vienna, Austria*: *R Foundaton for Statistical Computing*; *2008*. *Available at* [*http://www.R-project.org*](http://www.R-project.org).

# [Go EP](http://www.ncbi.nlm.nih.gov/pubmed?term=Go%20EP%5BAuthor%5D&cauthor=true&cauthor_uid=19418229). Database resources in metabolomics: an overview. [J Neuroimmune Pharmacol.](http://www.ncbi.nlm.nih.gov/pubmed/19418229) 2010 Mar;5(1):18-30. doi: 10.1007/s11481-009-9157-3. Epub 2009 May 7.

# [Wishart DS](http://www.ncbi.nlm.nih.gov/pubmed?term=Wishart%20DS%5BAuthor%5D&cauthor=true&cauthor_uid=21083105). Computational strategies for metabolite identification in metabolomics. [Bioanalysis.](http://www.ncbi.nlm.nih.gov/pubmed/21083105) 2009 Dec;1(9):1579-96. doi: 10.4155/bio.09.138.

# <http://www.ncbi.nlm.nih.gov/pubmed/21083105>

# [Knox C](http://www.ncbi.nlm.nih.gov/pubmed?term=Knox%20C%5BAuthor%5D&cauthor=true&cauthor_uid=21059682), [Law V](http://www.ncbi.nlm.nih.gov/pubmed?term=Law%20V%5BAuthor%5D&cauthor=true&cauthor_uid=21059682), [Jewison T](http://www.ncbi.nlm.nih.gov/pubmed?term=Jewison%20T%5BAuthor%5D&cauthor=true&cauthor_uid=21059682), [Liu P](http://www.ncbi.nlm.nih.gov/pubmed?term=Liu%20P%5BAuthor%5D&cauthor=true&cauthor_uid=21059682), [Ly S](http://www.ncbi.nlm.nih.gov/pubmed?term=Ly%20S%5BAuthor%5D&cauthor=true&cauthor_uid=21059682), [Frolkis A](http://www.ncbi.nlm.nih.gov/pubmed?term=Frolkis%20A%5BAuthor%5D&cauthor=true&cauthor_uid=21059682), [Pon A](http://www.ncbi.nlm.nih.gov/pubmed?term=Pon%20A%5BAuthor%5D&cauthor=true&cauthor_uid=21059682), [Banco K](http://www.ncbi.nlm.nih.gov/pubmed?term=Banco%20K%5BAuthor%5D&cauthor=true&cauthor_uid=21059682), [Mak C](http://www.ncbi.nlm.nih.gov/pubmed?term=Mak%20C%5BAuthor%5D&cauthor=true&cauthor_uid=21059682), [Neveu V](http://www.ncbi.nlm.nih.gov/pubmed?term=Neveu%20V%5BAuthor%5D&cauthor=true&cauthor_uid=21059682), [Djoumbou Y](http://www.ncbi.nlm.nih.gov/pubmed?term=Djoumbou%20Y%5BAuthor%5D&cauthor=true&cauthor_uid=21059682), [Eisner R](http://www.ncbi.nlm.nih.gov/pubmed?term=Eisner%20R%5BAuthor%5D&cauthor=true&cauthor_uid=21059682), [Guo AC](http://www.ncbi.nlm.nih.gov/pubmed?term=Guo%20AC%5BAuthor%5D&cauthor=true&cauthor_uid=21059682), [Wishart DS](http://www.ncbi.nlm.nih.gov/pubmed?term=Wishart%20DS%5BAuthor%5D&cauthor=true&cauthor_uid=21059682). DrugBank 3.0: a comprehensive resource for ‘omics’ research on drugs. [Nucleic Acids Res.](http://www.ncbi.nlm.nih.gov/pubmed/21059682) 2011 Jan;39(Database issue):D1035-41. doi: 10.1093/nar/gkq1126. Epub 2010 Nov 8.

# <http://www.ncbi.nlm.nih.gov/pubmed/21059682>

# Pence HE, Williams AJ. ChemSpider: an online chemical information resource. J. Chem. Educ. 2010;87(11):1123–1124. doi: 10.1021/ed100697w. [[Cross Ref](http://dx.doi.org/10.1021%2Fed100697w" \t "_blank)]

# ChemSpider ChemSpider. 2009 [http://www.chemspider.com](http://www.chemspider.com" \t "_blank). Ref Type: Electronic Citation.

# <http://www.ncbi.nlm.nih.gov/pmc/articles/PMC3598674/>

# [Uppal K](http://www.ncbi.nlm.nih.gov/pubmed?term=Uppal%20K%5BAuthor%5D&cauthor=true&cauthor_uid=23323971), [Soltow QA](http://www.ncbi.nlm.nih.gov/pubmed?term=Soltow%20QA%5BAuthor%5D&cauthor=true&cauthor_uid=23323971), [Strobel FH](http://www.ncbi.nlm.nih.gov/pubmed?term=Strobel%20FH%5BAuthor%5D&cauthor=true&cauthor_uid=23323971), [Pittard WS](http://www.ncbi.nlm.nih.gov/pubmed?term=Pittard%20WS%5BAuthor%5D&cauthor=true&cauthor_uid=23323971), [Gernert KM](http://www.ncbi.nlm.nih.gov/pubmed?term=Gernert%20KM%5BAuthor%5D&cauthor=true&cauthor_uid=23323971), [Yu T](http://www.ncbi.nlm.nih.gov/pubmed?term=Yu%20T%5BAuthor%5D&cauthor=true&cauthor_uid=23323971), [Jones DP](http://www.ncbi.nlm.nih.gov/pubmed?term=Jones%20DP%5BAuthor%5D&cauthor=true&cauthor_uid=23323971). xMSanalyzer: automated pipeline for improved feature detection and downstream analysis of large-scale, non-targeted metabolomics data. [BMC Bioinformatics.](http://www.ncbi.nlm.nih.gov/pubmed/23323971) 2013 Jan 16;14:15. doi: 10.1186/1471-2105-14-15.

Yu T, Park Y, Johnson JM, Jones DP: **apLCMS–adaptive processing of high-resolution LC/MS data.**  Bioinformatics 2009, **25**(15)**:**1930-1936.

Smith CA, Want EJ, O’Maille G, Abagyan R, Siuzdak G: **XCMS: processing mass spectrometry data for metabolite profiling using nonlinear peak alignment, matching, and identification.**  Anal Chem 2006, **78**(3)**:**779-787.

Kelder T, van Iersel MP, Hanspers K, Kutmon M, Conklin BR, Evelo C, Pico AR. (2011) WikiPathways: building research communities on biological pathways. NAR

Lange E, Tautenhahn R, Neumann S, Gropl C: **Critical assessment of alignment procedures for LC-MS proteomics and metabolomics measurements.** BMC Bioinformatics 2008, **9:**375.

Soltow QA, Strobel FH, Mansfield KG, Wachtman L, Park Y, Jones DP: **High-performance metabolic profiling with dual chromatography-Fourier-transform mass spectrometry (DC-FTMS) for study of the exposome.**  Metabolomics 2011.

Smith CA, O’Maille G, Want EJ, Qin C, Trauger SA, Brandon TR, Custodio DE, Abagyan R, Siuzdak G: **METLIN: a metabolite mass spectral database.**  Ther Drug Monit 2005, **27**(6)**:**747-751.

# [Xiao JF](http://www.ncbi.nlm.nih.gov/pubmed?term=Xiao%20JF%5BAuthor%5D&cauthor=true&cauthor_uid=22345829), [Zhou B](http://www.ncbi.nlm.nih.gov/pubmed?term=Zhou%20B%5BAuthor%5D&cauthor=true&cauthor_uid=22345829), [Ressom HW](http://www.ncbi.nlm.nih.gov/pubmed?term=Ressom%20HW%5BAuthor%5D&cauthor=true&cauthor_uid=22345829). Metabolite identification and quantitation in LC-MS/MS-based metabolomics. [Trends Analyt Chem.](http://www.ncbi.nlm.nih.gov/pubmed/22345829) 2012 Feb 1;32:1-14.

1. \*To whom correspondence should be addressed. [↑](#footnote-ref-2)