International Journal of Biomedical Data Mining

Research Article Open Access

Biclustering Impact in Biomedical Sciences via Literature Mining

Haithem Aouabed[1,3*] Rodrigo Santamaria[2] and Mourad Elloumi[1]

[1]Laboratory of Technologies of Information and Communication and Electrical Engineering (LaTICE), National High School of Engineers of Tunis (ENSIT) University of Tunis, Tunis, Tunisia.
[2]Department of Computer and Automation, University of Salamanca, Salamanca, Spain.
[3]Faculty of Economic Sciences and Management of Sfax (FSEGS), University of Sfax, Sfax, Tunisia.

Abstract

Biclustering algorithms have matured from their initial applications in bioinformatics, evolving towards different approaches and bicluster definitions, which makes sometimes hard for the analyst to determine which one of the available algorithms best fits her problem. As a way of benchmarking these algorithms, several quality measures have been proposed in literature. Such measures cover numerical aspects related to the accuracy, the recovery power or the capability of retrieving previous biomedical knowledge. However, biclustering apparently remains as an uncommon option for biomedicine analysis.

Here we review the impact of biclustering algorithms in biomedicine and bioinformatics with the object of measuring and understanding non-numerical aspects of biclustering algorithms focusing on citation-based statistics that can be relevant for their application on the domain. In order to achieve this, we performed analyses of the citations impact of several clustering and biclustering algorithms, and propose a methodology that can cover this aspect of biclustering usage.

Keywords: Biclustering; Clustering; Citations impact; Algorithm usage; Biomedicine; Bioinformatics.

Introduction

It has been 17 years since the first application of biclustering algorithms to bioinformatics [1]. After an initial bloom of different approaches and concepts of what a bicluster is [2], new algorithms continue to be developed at a constant pace. Biclustering has two main theoretical advantages over traditional clustering for the application to biology. On one side, it provides bi-dimensionality, grouping both genes and conditions together, which is much closer to biology, since a group of genes can be co-regulated for a given condition but not for others. On the other side, it considers group overlaps, allowing genes to contribute to more than one activity. Although clustering does not cover either of these advantages, it seems that biclustering is far from replacing hierarchical or K-means clustering as a first analysis option.

A possible major reason is the lack of standards and benchmarks, along with the different interpretations of what a biclustering is. This has been addressed in the past years and a good deal of numerical quality measures have been developed, narrowing down the number of biclustering algorithms that fulfill recovery criteria. Additionally, the comparison of biclustering algorithms takes more and more interest in the literature. These studies have provided useful surveys of the biclustering landscape and also clarified the relative performance of several algorithms. Turner et al. [3] proposed an external measure in addition to a benchmark in order to assess biclustering algorithms. Prelić et al. [4] compared six algorithms on a synthetic dataset as well as on two real datasets. Results over synthetic data were evaluated by a measure called gene match score. For real data, the biclusters found were evaluated by gene ontology enrichment in addition to metabolic and protein-protein interaction networks. Bozda et al. [5] compared six biclustering algorithms which have the ability to find biclusters by means of shifting and scaling models. Impression of some parameters like bicluster size, noise and overlap was evaluated on artificial and real datasets. To estimate the exactness between found and implanted biclusters, the authors defined several external scores. Eren et al. [6] presented a comparative study between twelve biclustering algorithms. In order to evaluate their results, the authors used eight real datasets and some synthetic datasets that present six different bicluster models in addition to two external evaluation measures. Horta and Campello [7] focused on the definition of necessary properties that must be satisfied by a biclustering external evaluation measure. Their analysis is performed on fourteen measures where two of them, namely Clustering Error and Campello Soft Index, are recommended to use in such evaluation process. Padilha and Campello [8] presented a comparative study of seventeen biclustering algorithms. To achieve their tasks, the authors relied on three synthetic datasets and two real datasets. For synthetic data, five different experimental scenarios were studied based on noise, numbers of implanted biclusters, overlap levels and bicluster sizes, and the results were assessed with several external measures including those defined in [7]. Gene ontology enrichment and clustering reliability were used to assess results obtained from real data.

In spite of their important contributions, all these studies used only numerical measures to benchmark biclustering methods. However, there are non-numerical aspects of biclustering algorithms which have an impact on their success, especially for application fields such as biology or medicine, where a skilled bioinformatician is not always available. Among these aspects are: source code availability, easy to run executables or scripts, clear documentation, easy to parse and interpret results, connectivity to visualization tools, available packages for common bioinformatics languages such as R/BioConductor [9] or Biopython, etc.

An effective method to measure such a heterogeneous non-

***Corresponding author:** Haithem Aouabed, Laboratory of Technologies of Information and Communication and Electrical Engineering (LaTICE), National High School of Engineers of Tunis (ENSIT) University of Tunis, Tunisia, Tel: 216 94740843; E-mail: haithem.abdi@gmail.com

Received July 07, 2018; **Accepted** August 21, 2018; **Published** August 27, 2018

Citation: Aouabed H, Santamaria R, Elloumi M (2018) Biclustering Impact in Biomedical Sciences via Literature Mining. Int J Biomed Data Min 7: 134. doi: 10.4172/2090-4924.1000134

numerical bunch of factors can be to determine the final impact of the algorithm in the field of study, under the assumption that a citation of a biclustering algorithm in a biology or medicine paper might point to its actual usage or at least to its consideration or knowledge. On the other side, the citation of a biclustering algorithm in a bioinformatics paper might be for comparison, benchmarking or background.

These types of literature impact studies form a well-defined field of study [10], demonstrating their usefulness in different research domains. In fact, literature impact studies have been used to analyze bioinformatics, biology and medicine issues. Magana et al. [11] reviewed the state of integration of bioinformatics education into formal and informal educational settings. The selected publications were issued following a search in Google Scholar, Web of Science, ACM Digital Library, ERIC, and PubMed. Their search found 113 documents that were published from 1998 to 2013 and reported three types of scholarly publications: journal papers, conference proceedings, and magazine articles. The analysis process started by determining the frequencies of each type of article along with the year in which they were published. Then, the content of the abstracts in each of the categories was analyzed in order to extract the themes. Attia et al. [12] presented a systematic review on sexual transmission of HIV according to viral load and antiretroviral therapy. They searched the Medline and EMBASE databases in addition to conference abstracts from 1996 to 2009. Their search resulted on 305 publications, 56 of which were considered as conference abstracts and then statistically analyzed. Tseng et al. [13] investigated a review that focused on various biological purposes for microarray meta-analysis. They used two main sources: PubMed and manual collection. 333 out of 745 papers were found related to microarray meta-analysis. These papers were then classified by method type, meta-analysis type and purpose for further statistical analysis. In Duck et al. [14], the authors defined and investigated an evaluation method based on a literature review, measuring the rates of usage of databases as well as software in biological and medical domains. This comparison is defined over two axes: time and sub-disciplines of bioinformatics, biology and medicine domains. An invented dictionary as well as a rule-based resource recognition system named bioNerDS [15] was used to retrieve both old and new database and software names. The analysis process was applied on 25 articles using the Singular Value Decomposition (SVD) clustering method.

Recently, these types of analysis have been applied in new research areas such as Big Data or Cloud Computing and their impacts in biology/medicine fields have been evaluated. Hermon and Williams [16] proposed a review methodology which consists in finding different usage categories of big data in healthcare. The literature search yielded 40 articles identified from 16 academic journals, including JAMIA, Wiley Online and IEEE. These articles were then classified into five categories based on the diversity in context. In Baro et al. [17], the authors proposed a definition of big data in healthcare based on a systematic search of PubMed literature. The search method made use of online search amenities such as the Free PMC database, Google and Google Scholar, finding 196 papers of interest. These papers were then classified either as a paper describing a dataset (further divided into three subclasses), a dissertation or a review of the literature. Statistical analyses were made, focused on time evolution of the publications and the related datasets.

To our knowledge, none of these studies introduced the analysis of biclustering literature impact using measures such as the number of citations for each paper. In the following sections, we propose a method to evaluate the impact of biclustering and clustering papers in different fields of study, represented by a set of relevant journals. We reviewed several published clustering and biclustering algorithms from the point of view of the proposed method, in order to assess their impact in bioinformatics and biomedicine.

Methods

Selection of biclustering algorithms

We have chosen seventeen algorithms based on the availability and diversity in approaches to solve the biclustering problem, which correspond to the algorithms selected by a recent through comparative study [8]. The list of biclustering algorithms includes popular ones in literature, such as Cheng and Church (CCA) [1], Plaid [18], Spectral [19], ISA [20], Bimax [4], xMOTIFs [21], SAMBA [22], OPSM [23] and MSSRCC [24]. Furthermore, newer algorithms such as Bayesian Biclustering (BBC) [25], COALESCE [26], FABIA [27], CPB [28], QUBIC [29], LAS [30], BiBit [31] and DeBi [32] have recently proved their effectiveness to handle biclustering issues.

Selection of clustering algorithms

In order to compare biclustering impact with clustering impact, we also selected several clustering algorithms and performed the same analysis with them. These clustering algorithms were selected based on their implementation availability, their popularity (more than 5000 citations according to Google Scholar) and their ability to handle the specific needs of biological fields, such as microarray gene expression analysis. Therefore we focused our analysis on seven implementations of clustering algorithms: Eisen hierarchical clustering [33], Principal Component Analysis (PCA) [34], Support Vector Machines (SVM) [35], K-means [36], Model-Based [37], Self-Organizing Map (SOM) [38] and CAST [39].

Selection of application fields

The JCR (Journal Citation Report) directory has been used as a reference for the definition of fields of application. As a result, the papers citing these biclustering or clustering algorithms are classified depending on their application field, considering three of them:

- Applied biosciences (from now on, biomed): includes any journal found under different biology and medicine subjects in the Journal Citation Report (JCR) 2016. See Table 1 in supplementary information for a relation of the subjects and Table 2 in Supplementary Information for a relation of the journals included in such journals.
- Bioinformatics (from now on bioinfo): includes any journal found under the subject Mathematics and Computational Biology in JCR 2016. See Table 3 in Supplementary Information for a relation of the journals.
- Others: any other journal not in the former two fields. It also includes papers that are not in journals (papers in proceedings, technical reports, book chapters, etc.).

With this classification we expect to group most of the algorithms' applications to biology and medicine in biomed, and the algorithms 'performance tests, reviews, etc. in bioinfo. Setting hard separations on journals' themes will generate some misclassification (e.g. a biomed journal publishing a biclustering performance review), but we think that it is good enough for the aims of our study, as shown in the results. We also cover this issue by a manual curation of usage (see next section). It must also be noted that the biomed field is much

Algorithm	Citing papers (per year exc. self cit.)	Citing papers (total)	Usage		Usage (exc. self cit.)		Comparison papers	Review papers	Context	Total citations per year	Number of integrative tool or library
			Biomed	Bioinfo	Biomed	Bioinfo					
PCA	1.65	32	13	3	13	3	1	9	6	1038.23	2
Eisen	1.26	32	22	0	15	0	0	4	6	5869.68	6
Model-Based	1.12	20	3	0	3	0	2	7	8	474.68	1
SVM	1.60	30	3	3	3	1	2	7	15	2106.11	1
K-means	1.55	33	13	2	8	2	0	10	8	2039.50	4
SOM	1.83	36	10	2	5	2	1	7	16	3694.77	3
CAST	1.77	34	2	3	2	2	3	9	17	1250.31	1

Table 1: Summary of the studied clustering algorithms. Citation and usage measures for clustering algorithms. The detailed integrative tools and libraries for each clustering algorithm are presented in Table 6 of Supplementary Information.

Algorithm	Citing papers (per year exc. self cit.)	Citing papers (total)	Usage		Usage (exc. self cit.)		Comparison papers	Review papers	Context	Total Citations per year	Padilha and Campello performance	Number of integrative tool or library
			Biomed	Bioinfo	Biomed	Bioinfo						
Bimax	2.72	32	4	3	4	2	4	4	17	514.18	4	3
SAMBA	1.60	29	3	2	1	0	1	8	15	883.86	4	1
COALESCE	2.37	19	0	0	0	0	5	4	10	263.50	3	1
CCA	1.29	22	0	2	0	2	3	4	13	692.70	0	4
BiBit	1.00	7	1	0	1	0	3	0	3	12.83	4	1
ISA	2.35	35	5	1	3	0	6	2	18	777.94	2	3
BBC	2.22	20	1	0	1	0	3	3	13	73.33	0	0
FABIA	2.85	24	4	4	2	3	4	0	12	97.71	2	3
Plaid	1.46	22	0	0	0	0	1	5	16	630.26	2	2
Spectral	1.64	25	4	0	2	0	2	5	14	434.14	1	2
xMOTIFs	1.85	28	0	2	0	1	3	3	20	402.28	1	4
LAS	1.00	8	3	0	3	0	1	0	3	32.00	3	1
CPB	0.62	8	1	0	0	0	2	1	4	19.75	2	0
QUBIC	2.12	19	3	0	0	0	6	1	9	85.75	1	0
OPSM	1.85	26	0	2	0	2	5	2	17	465	0	1
MSSRCC	0.38	5	1	0	0	0	1	0	3	174.61	1	0
DeBi	2.33	14	0	1	0	1	1	1	11	28.50	0	0

Table 2: Summary of the 17 biclustering algorithms. Citation and usage measures for biclustering algorithms, as defined by the score metrics. The Padilha and Campello performance column shows the number of tests (out of 5) for which the algorithm has good results in [8]. The detailed integrative tools and libraries for each biclustering algorithm are presented in Table 7 of Supplementary Information.

Biclustering	Comparison	Review	Usage	Usage (no s.c)	Context
Bioinfo (%)	15	7	9	6	69
Biomed (%)	15	20	21	12	44
TOTAL (%)	15	13	14	8	58
Clustering	**Comparison**	**Review**	**Usage**	**Usage (no s.c)**	**Context**
Bioinfo (%)	11	16	18	14	55
Biomed (%)	1	28	46	34	25
TOTAL (%)	4	24	36	27	35

Table 3: Percentage of biclustering and clustering papers classification

larger (1467 journals, compared to 56 in bioinfo) but also much more heterogeneous in applications and methods and sparser in biclustering citations. Given this and the characteristics of the results found, we consider these numbers acceptable for our objectives.

Definition of score metrics

Citations impact is a relevant indicator that helps to discover the penetration rate of a tool, a method or even an algorithm in a particular domain [40]. We will use this indicator in the study of the usage rate of clustering and biclustering algorithms in biomedicine and bioinformatics as follows:

Total number of citations= #citations in biomed + #citations in bioinfo.

#citations refers to the sum of the number of citations obtained by the most important papers citing a given algorithm (from now on, citing papers), as of January 2018, in each domain. We limit our analysis to the top 50 most cited citing papers, because of citations retrieval constrains and to be able to curate them (see below). In order to avoid favoring algorithms published long time ago, the number of citing papers and number of citations is averaged per year since the publication date.

Finally, in order to characterize citations, we curated the citing papers into four categories based on a manual classification process that take into consideration the citation semantics:

- Usage if the paper cites the algorithm among their methods or as producer of some of the paper's results.

- Context if the paper cites the algorithm in other sections, such as the introduction or background, but without actually using it.
- Comparison if the paper is a comparison among clustering or biclustering algorithms.
- Review if the paper is a review of clustering or biclustering algorithms.

These measures were corrected to remove self-citation or co-authored usage or citation counts.

Besides usage we also checked if the biclustering or clustering algorithms are implemented in broadly spread packages, libraries or integrative tools, such as Expander [41], Java TreeView [42], MCLUST [43], BicAT [44], BicOverlapper [45], Genesis [46], Bioconductor, HCE [47], biclust [48], biclustlib [8], MeV [49] and Sleipnir [50].

Citations retrieval

To retrieve the citations, we implemented bambu, a script based on scholar.py (https://github.com/ckreibich/scholar.py), a known Python approach to the Google Scholar web search engine (https://scholar.google.com). The developed tool is available at http://vis.usal.es/bambu and can be used for any paper title and list of journals, retrieving the top citing papers in the field, and measuring different parameters about its impact.

Results

Literature impact of clustering algorithms

Table 1 and Figure 1 show the literature impact for the selected clustering algorithms. Figure 1 shows that almost all of the clustering algorithms have at least one highly cited citing paper per year on biomed journals or, in the case of CAST, in bioinfo journals. About the relevance of the citing papers, Eisen algorithm seems to be the most relevant, with highly cited papers citing it.

Our curation process found that each clustering algorithm has at least two applications in the biomed domain within the topmost 50 citing papers (see Table 1 for a summary and Table 4 in Supplementary Information for a detailed description of the curation). All the methods have at least one citing paper per year since its publication. The usage ratio varies from methods with low penetration such as Model-Based (3 citing papers using it), SVM (6) or CAST (5); to more popular methods such as Eisen (22 citing papers, all in biomedicine), PCA or K-means (13 citing papers in biomedicine).

Most of the clustering algorithms have a larger number of applications in biomed than in bioinfo. For example, Eisen method has all of its 15 non-coauthored applications in biomed journals; in the case of SOM, there are 12 applications, 10 of them in biomed. These findings confirm that most of these clustering algorithms have a stronger impact in biomedicine, and that this one is a natural field of application for clustering.

Although most of the applications correspond to the usage of clustering algorithms in biomedical problems, some of them correspond to integrations or implementations of the algorithms in other platforms, especially in the bioinfo domain. For example, the two bioinfo applications of SOM consist on the implementation of the algorithm in a library [51] and its integration in a tool [46]. Something similar occurs with the three bioinfo applications of PCA, although it is not always the case SVM has two bioinfo applications that correspond to microarray data analyses (see Table 4 in Supplementary Information). The exclusion of self-citations reveal how dependent on the original authors the algorihtm is for actual usage, ranging from 33% of coauthored applications in Eisen or K-means to 0% in PCA, or Model Based algorithms (see Table 1). Only SOM has a coauthored application rate larger than one third (41.6%).

Regarding to biomed application details, most of them are used for classification of genomic data from different sources, usually gene expression data (see Table 4 in Supplementary Information, highlighted entries). For example, in the case of CAST, one of its applications is for clustering gene expression data to discover a subset of melanomas [52] while the second application is about clustering of DNA sequence motifs [53].

Literature impact of biclustering algorithms

Table 2 and Figure 2 show the impact and scores for the studied biclustering algorithms (see Table 5 in Supplementary Information for additional information about curated biclustering citations). Most of the analyzed biclustering algorithms have at least one citing paper per year, with a general preeminence of bioinfo papers. There is a large variation in the number of citations obtained by the citing papers, and the most cited ones (ISA and SAMBA) take the larger number of citations from biomed citing papers.

We investigated the semantics of citations in biomed and bioinfo journals (Table 2, biomed and bioinfo usage). Bimax and FABIA have the highest numbers of application papers (5 or more citations each), pointing to a stronger impact in biomedicine and bioinformatics. In the case of COALESCE and Plaid model, although their high number of citing papers (19 and 22 respectively), none of these papers either in biomed or bioinfo journals is an application of the algorithm, although for COALESCE two of its citing papers are biclustering comparisons in which it has good results [54,55]. Notice that these biclustering comparisons are both published in biomed journals (Genome Biology and Nucleic Acids Research). In the case of Bimax, 7 out of 32 citing papers in biomed and bioinfo journals are actual applications. Among these applications, 4 of them represent actual applications of the algorithm in biology. Thereby, Bimax has a prominent role for identifying groups of genes with similar expression profiles, using it as the preferred classification method for; A. thaliana gene expression [56], cancer data [57], co-regulated genes with drugs [58] and Adverse Drug Events in the United States Food and Drug Administration's (FDA) Spontaneous Reporting System [59]. In the case of SAMBA, with a considerable number of citations of its journals in the two fields (biomed and bioinfo), 5 out of the 29 citing papers are applications of the algorithm. Focusing on the biomed domain, there are 3 papers out of the 5 which are actual applications of SAMBA in biology. The first application uses a modified version of the algorithm to bicluster gene expression data [60]. However, this is a paper where two of the original authors of SAMBA appear as co-authors, so it can be considered as proof of the usage in the biology field but not as a third-party corroboration of successful usage. The second application consists in biclustering usage on a gene set matrix constructed from an association of lincRNAs data with their functional gene sets [61] while the third one, another co-authored paper, uses the algorithm as a principal method of biclustering from the Expander tool [62]. For CCA biclustering algorithm, only 2 out of its 22 citing papers can be considered as actual applications [44,63]. Note that these applications are in bioinfo domain. In the case of ISA, 6 out of 35 citing papers are applications in biomed and bioinfo domains. Note that there are 3 other papers for ISA in bioinfo that represent new versions of the algorithm; USA [64], EDISA [65] and PISA [66]. Moreover, in the case of LAS and

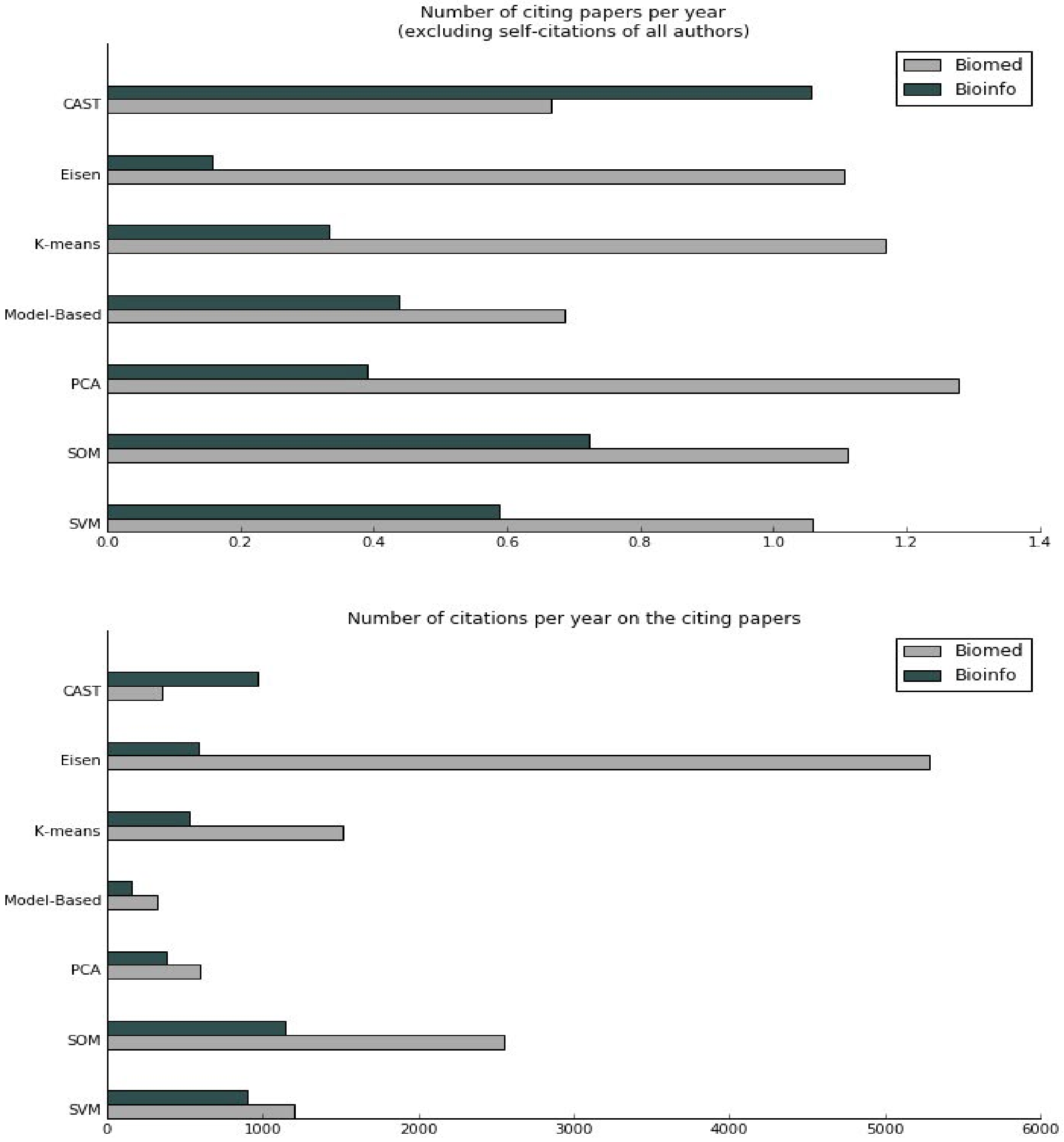

Figure 1: Literature impact of selected clustering algorithms in each domain

despite its low number of citing papers, 4 of its 8 papers are considered as applications of the algorithm. As for ISA, there is a paper in bioinfo where a new version of LAS was proposed based on a preprocessing technique named localization [67]. In the case of FABIA, 8 out of its 24 citing papers are considered as actual applications (3 co-authored). Among these 8 papers, there are 4 of them that can be considered as applications of the FABIA algorithm on biology domain. The first two ones used FABIA as the preferable method of identification. The first paper used the biclustering algorithm to identify very short Identity by descent (IBD) segments characterized by rare variants [68] while the second one used FABIA to identify transcriptional modules [69]. We mention that these 2 applications are co-authored. The third FABIA application used the algorithm for biclustering of microbiome data [70] while the last biology application used FABIA to classify gene expression profiles [71]. About the remaining biclustering algorithms, BBC, BiBit and DeBi have one third-party citing paper using the algorithm [72-74]. MSSRCC, QUBIC and CPB also have papers using them, but linked to their original authors.

Discussion

Clustering and biclustering impact comparison in the biomed domain

Based on our compiled results (Table 3), algorithm usage is less frequent on citing papers for biclustering (14%) than for clustering (36%). The difference is threefold if we remove self-citations (8% vs

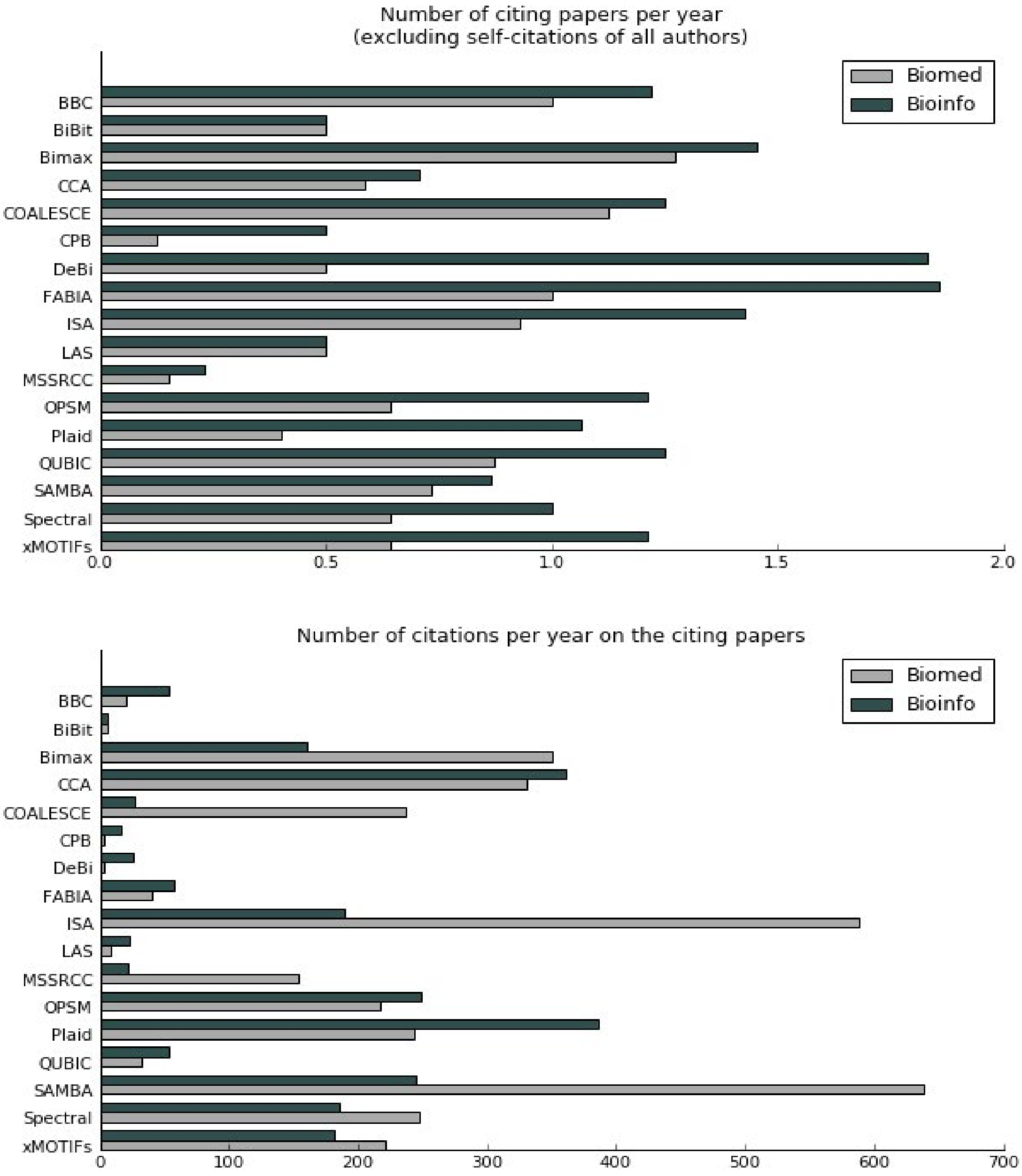

Figure 2: Literature impact of selected biclustering algorithms in each domain

27%). Biclustering is mostly cited as context (58%), while clustering is less cited as context, especially in biomed journals (25%). This may suggest that biclustering is known in the field but is not so frequently used as clustering. The penetration of biomed comparisons in biclustering is also remarkable (14% of total biomed biclustering citing papers) compared to clustering (1%), which might also point to a stage of competition among methods in performance or quality.

A comparison of algorithm usage respect to algorithm citation (corrected per year) on both clustering and biclustering shows that clustering usage spikes faster than biclustering usage with the increase of citations (Figure 3, linear regression slope of 0.68 vs. 0.18).

Although the first biclustering algorithm dates back to 1972 [75], this paradigm only started to draw the biomedical community's attention after its first application [1]. Most biclustering algorithms, measures or methodologies have been developed since then. At the same time, in 2000, the clustering literature was rather mature, with decades of research and improvements [76]. Many clustering algorithms, evaluation measures and benchmarks were available. So, we believe that this could be one of the main reasons that clustering methods are still more popular than biclustering methods in biosciences and there is a clear timing advantage in favor of traditional clustering algorithms due to a previous consolidation in statistics, rather than to a previous publishing of the methods themselves.

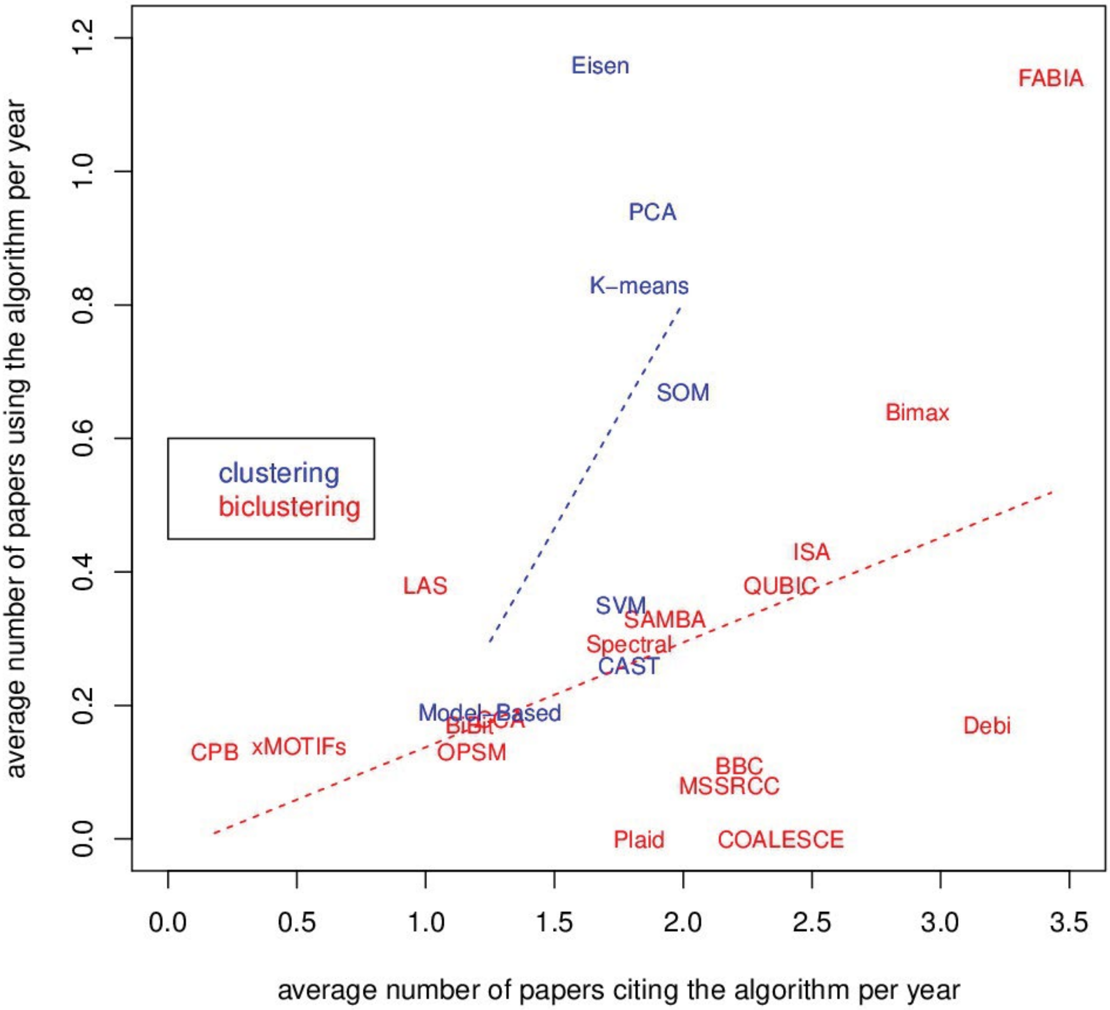

Figure 3: Citation vs. usage of biclustering and clustering algorithms in biomed domain (based on data from Table 3).

Another possible explanation of clustering dominance is that clustering is easier to understand (no overlaps, only one dimension) and also fits well with very rooted biological concepts such as taxonomical trees, as in the case of hierarchical clustering.

Possible catalyzers of biclustering impact

Only two biclustering algorithms have more than 0.5 citing papers using them per year on average: Bimax and FABIA, although a total of 9 biclustering algorithms have more than 2 citing papers per year on average. Although all the biclustering algorithms have a publicly available implementation, some of them are not included in easy to use libraries or integrative tools, such as BBC, MSSRCC, QUBIC and Debi; all of them with more than 2 citing papers per year. On the other side, Bimax has at least three implementations (R package biclust and visual integrative tools BicAT and BicOverlapper) and FABIA has its own BioConductor package. This availability could be one source of success for biclustering algorithm usage. Another possible indicator of this is the fact that self- citation rate in usage is larger in biclustering (43%) than in clustering (25%), suggesting that the biclustering tools are less approachable without an expert involved in the algorithm co-working on its application to a given problem.

Taking into account the recent and exhaustive biclustering comparison by Padilha and Campello [8], good performance in benchmark tests is another candidate pre-condition for biclustering usage. The average usage ratio per year of all the algorithms without superior results in any of Padilha tests is below 0.2. Bimax, with a usage ratio above 0.6 fulfills 4 out of 5 Padilha's tests (Table 2). However, other algorithms with good performance like BiBit or COALESCE have low usage impacts (Table 2 and Figure 1 in Supplementary Information) pointing that a moderately good performance may be a necessary but not sufficient pre-condition for usage.

Finally, another interesting fact is that one of the most successfully used algorithm according to this analysis is Bimax, a biclustering algorithm with a very simple definition of what a bicluster is (constant biclusters based on data binarization), which might encourage researchers to use and interpret its results, rather than more complex definitions (scale and shift, coherent evolution biclusters).

Conclusion

Biclustering algorithms have been available in bioinformatics analysis since 2000, almost from the first genome-wide expression analyses [33]. In this time, in the light of our review, they have become known in biomedical domains, some of them appearing in highly cited papers and reviews on journals such as Nucleic Acids Research, BMC Biology, Nature Genetics or Cancer Research. However, their penetration into this domain as a popular tool to be used by non-experts is still far from the penetration of clustering algorithms.

The success of clustering over biclustering can be due to, in one hand, the relative novelty of biclustering algorithms, not in bioinformatics but in statistics, respect to clustering. In other hand, clustering might be favored by a more consistent presence in the field of statistics and its easier interpretation of results, maybe combined with the familiarity of concepts such as hierarchical classification. However, there is room for biclustering in biomedical applications, as shown for some algorithms that present biomedical penetration in reviews and some approaches, especially if they succeed at the issues of availability, benchmarking and easiness of usage and interpretation.

It is possibly of interest for the field to consider these aspects, in order to make biclustering a more useful and used approach for one of the major research fields it is intended to.

Acknowledgements

This work has received funding within the CHIST-ERA programme under the following national grant agreement PCIN-2017-064 (MINECO Spain).

References

1. Cheng Y, Church GM (2000) Biclustering of expression data. Proceedings Int Conf Intell Syst Mol Biol.
2. Madeira SC, Oliveira AL (2004) Biclustering algorithms for biological data analysis: a survey. IEEE Trans Comput Biol Bioinforma1:24-45.
3. Turner H, Bailey T, Krzanowski W (2005) Improved biclustering of microarray data demonstrated through systematic performance tests. Comput Stat Data Anal 48:235-254.
4. Prelic A, Bleuler S, Zimmermann P, Wille A, Buhlmann P, et al. (2006) A systematic comparison and evaluation of biclustering methods for gene expression data. Bioinformatics 22:1122-1129.
5. Bozdag D, Kumar AS, Catalyurek UV (2010) Comparative analysis of biclustering algorithms. In: Proceedings of the First ACM International Conference on Bioinformatics and Computational Biology - BCB 10. New York, New York, USA: ACM Press 265.
6. Eren K, Deveci M, Kucuktunc O, Çatalyurek U V (2013) A comparative analysis of biclustering algorithms for gene expression data. Brief Bioinform.
7. Horta D, Campello RJGB (2014) Similarity Measures for Comparing Biclusterings. IEEE/ACM Trans Comput Biol Bioinforma 11:942-954.
8. Padilha VA, Campello RJGB (2017) A systematic comparative evaluation of biclustering techniques. BMC Bioinformatics.
9. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, et al. (2004) Bioconductor: open software development for computational biology and bioinformatics. Genome Biol Evol.
10. Kumar MJ (2009) Evaluating Scientists: Citations, Impact Factor, h-Index, Online Page Hits and What Else? Iete Tech Rev 26:165.
11. Magana AJ, Taleyarkhan M, Alvarado DR, Kane M, Springer J, et al. (2014) A survey of scholarly literature describing the field of bioinformatics education and bioinformatics educational research. CBE Life Sci Educ 13:607-623.
12. Attia S, Egger M, Muller M, Zwahlen M, Low N (2009) Sexual transmission of HIV according to viral load and antiretroviral therapy: systematic review and meta-analysis. AIDS 23:1397-1404.
13. Tseng GC, Ghosh D, Feingold E (2012) Comprehensive literature review and statistical considerations for microarray meta-analysis. Nucleic Acids Res 40:3785-3799.
14. Duck G, Nenadic G, Filannino M, Brass A, Robertson DL, et al. (2016) A survey of bioinformatics database and software usage through mining the literature. PLoS One 11:1-25.
15. Duck G, Nenadic G, Brass A, Robertson DL, Stevens R (2013) bioNerDS: exploring bioinformatics database and software use through literature mining. BMC Bioinformatics.
16. Hermon R, Williams P (2014) Big data in healthcare: What is it used for? Proc 3rd Aust eHealth Informatics Secur Conf 40-49.
17. Baro E, Degoul S, Beuscart R, Chazard E (2015) Toward a literature-driven definition of big data in healthcare. Biomed Res Int.
18. Lazzeroni L, Owen A (2000) Plaid Models for Gene Expression Data. CEUR Work Proc 1542:33-36.
19. Kluger Y, Basri R, Chang JT, Gerstein M (2003) Spectral Biclustering of Microarray Data: Coclustering Genes and Conditions 703-716.
20. Bergmann S, Ihmels J, Barkai N (2003) Iterative signature algorithm for the analysis of large-scale gene expression data. Phys Rev E - Stat Nonlinear Soft Matter Phys 67:1-18.
21. Murali TM, Kasif S (2003) Extracting conserved gene expression motifs from gene expression data. Pacific Symp Biocompu 88:77-88.
22. Tanay A, Sharan R, Shamir R (2002) Discovering statistically significant biclusters in gene expression data. Bioinformatics 18 Suppl 1:S136-144.
23. Ben DA, Chor B, Karp R, Yakhini Z (2002) Discovering local structure in gene expression data. Proc sixth Annual International Conference Computational Biology RECOMB 2:49-57.
24. Cho H, Dhillon IS, Guan Y, Sra S (2004) Minimum Sum-Squared Residue Co-clustering of Gene Expression Data.
25. Gu J, Liu JS (2008) Bayesian biclustering of gene expression data. BMC Genomics 9 Suppl 1:S4.
26. Huttenhower C, Tsheko MK, Indik N, Yang W, Schroeder M, et al. (2009) Detailing regulatory networks through large scale data integration. Bioinformatics 25:3267-3274.
27. Hochreiter S, Bodenhofer U, Heusel M, Mayr A, Mitterecker A, et al. (2010) FABIA: Factor analysis for bicluster acquisition. Bioinformatics 26:1520-1527.
28. Bozdag D, Parvin JD, Catalyurek UV (2009) A biclustering method to discover co-regulated genes using diverse gene expression datasets. Lect Notes Comput Sci 5462 LNBI: 151-163.
29. Li G, Ma Q, Tang H, Paterson AH, Xu Y (2009) QUBIC: A qualitative biclustering algorithm for analyses of gene expression data. Nucleic Acids Res: 37.
30. Shabalin AA, Weigman VJ, Perou CM, Nobel AB (2009) Finding large average submatrices in high dimensional data.
31. Rodriguez BDS, Perez PAJ, Aguilar RJS (2011) A biclustering algorithm for extracting bit-patterns from binary datasets. Bioinformatics 27:2738-2745.
32. Serin A, Vingron M (2011) DeBi: Discovering Differentially Expressed Biclusters using a Frequent Itemset Approach. Algorithms Mol Biol 6:18.
33. Eisen MB, Spellman PT, Brown PO, Botstein D (1998) Cluster analysis and display of genome-wide expression patterns. Proc Natl Acad Sci 95:1-6.
34. Raychaudhuri S, Stuart JM, Altman RB (2000) Principal components analysis to summarize microarray experiments: application to sporulation time series. Pac Symp Biocomput 455-466.
35. Brown MP, Grundy WN, Lin D, Cristianini N, Sugnet CW, et al. (2000) Knowledge-based analysis of microarray gene expression data by using support vector machines. Proc Natl Acad Sci USA 97:262-267.
36. Tavazoie S, Hughes JD, Campbell MJ, Cho RJ, Church GM (1999) Systematic determination of genetic network architecture. Nature Genet 22:281-285.
37. Yeung KY, Fraley C, Murua A, Raftery AE, Ruzzo WL (2001) Model-based clustering and data transformations for gene expression data. Bioinformatics 17:977-987.
38. Tamayo P, Slonim D, Mesirov J, Zhu Q, Kitareewan S, et al. (1999) Interpreting patterns of gene expression with self-organizing maps: methods and application to hematopoietic differentiation. Proc Natl Acad Sci USA 96:2907-2912.
39. Ben DA, Shamir R, Yakhini Z (1998) Clustering gene expression patterns. Journal of Computational Biolgy 6:281-297.
40. Waltman L (2016) A review of the literature on citation impact indicators. J. Informetr 10:365-391.

41. Shamir R, Maron KA, Tanay A, Linhart C, Steinfeld I, et al. (2005) EXPANDER an integrative program suite for microarray data analysis. BMC Bioinformatics 6:232.

42. Saldanha AJ (2004) Java Treeview Extensible visualization of microarray data. Bioinformatics 20:3246-3248.

43. Fraley C, Raftery AE (2002) MCLUST: Software for Model-Based Clustering, Density Estimation and Discriminant Analysis.

44. Barkow S, Bleuler S, Prelic A, Zimmermann P, Zitzler E (2006) BicAT: a biclustering analysis toolbox. Bioinformatics 22:1282-1283.

45. Santamaria R, Theron R, Quintales L (2014) BicOverlapper 2.0: visual analysis for gene expression. Bioinformatics 30:1785-1786.

46. Sturn A, Quackenbush J, Trajanoski Z (2002) Genesis: cluster analysis of microarray data. Bioinformatics 18:207-208.

47. Seo J, Shneiderman B (2002) Interactively exploring hierarchical clustering results. Comput (Long Beach Calif) 35.

48. Kaiser S, Santamaria R, Khamiakova T, Sill M, Theron R, et al. (2013) biclust: BiCluster Algorithms. R package version 1.0.2.

49. Dudoit S, Gentleman RC, Quackenbush J (2003) Open Source Software for the Analysis of Microarray Data 45-51.

50. Huttenhower C, Schroeder M, Chikina MD, Troyanskaya OG (2008) The Sleipnir library for computational functional genomics. Bioinforma Appl Note 24:1559-1561.

51. De HMJL, Imoto S, Nolan J, Miyano S (2004) Open source clustering software. Bioinformatics 20:1453-1454.

52. Bittner M, Meltzer P, Chen Y, Jiang Y, Seftor E, et al. (2000) Molecular classification of cutaneous malignant melanoma by gene expression profiling. Nature 406: 536-540.

53. Bussemaker HJ, Li H, Siggia ED (2001) Regulatory element detection using correlation with expression 27:167-171.

54. Waltman P, Kacmarczyk T, Bate AR, Kearns DB, Reiss DJ, et al. (2010) Multi-species integrative biclustering. Genome Biology 11:96.

55. Reiss DJ, Plaisier CL, Wu WJ, Baliga NS (2015) cMonkey2: Automated, systematic, integrated detection of co-regulated gene modules for any organism. Nucleic Acids Research 43:e87.

56. Fode B, Siemsen T, Thurow C, Weigel R, Gatz C (2008) The Arabidopsis GRAS protein SCL14 interacts with class II TGA transcription factors and is essential for the activation of stress-inducible promoters. Plant Cell. ASPB 20:3122-3135.

57. Navin N, Kendall J, Troge J, Andrews P, Rodgers L, et al. (2011) Tumour evolution inferred by single-cell sequencing. Nature 472:90-94.

58. Iskar M, Zeller G, Blattmann P, Campillos M, Kuhn M, et al. (2013) Characterization of drug-induced transcriptional modules: Towards drug repositioning and functional understanding. Mol Syst Biol 9.

59. Harpaz R, Perez H, Chase HS, Rabadan R, Hripcsak G, et al. (2011) Biclustering of Adverse Drug Events in the FDA's Spontaneous Reporting System. Clin Pharmacol Ther 89:243-250.

60. Dudley AM, Janse DM, Tanay A, Shamir R, Church GM (2005) A global view of pleiotropy and phenotypically derived gene function in yeast. Mol Syst Biol Mar 1:E1-E11.

61. Guttman M, Amit I, Garber M, French C, Lin MF, et al. (2009) Chromatin signature reveals over a thousand highly conserved large non-coding RNAs in mammals. Nature 458:223-227.

62. Ulitsky I, Maron KA, Shavit S, Sagir D, Linhart C, et al. (2009) Expander: from expression microarrays to networks and functions. Nat Protoc 5:303-322.

63. Ihmels J, Bergmann S, Barkai N (2004) Defining transcription modules using large-scale gene expression data. Bioinformatics 20:1993-2003.

64. Kim H, Hu W, Kluger Y (2006) Unraveling condition specific gene transcriptional regulatory networks in Saccharomyces cerevisiae. BMC Bioinformatics 7:1-17.

65. Supper J, Strauch M, Wanke D, Harter K, Zell A (2007) EDISA: Extracting biclusters from multiple time-series of gene expression profiles. BMC Bioinformatics 8:1-14.

66. Kloster M, Tang C, Wingreen NS (2005) Finding regulatory modules through large-scale gene-expression data analysis. Bioinformatics 21:1172-1179.

67. Erten C, Sozdinler M (2010) Improving performances of suboptimal greedy iterative biclustering heuristics via localization. Bioinformatics 26:2594-2600.

68. Hochreiter S (2017) Hap FABIA: Identification of very short segments of identity by descent characterized by rare variants in large sequencing data. Nucleic Acids Research.

69. Verbist B, Klambauer G, Vervoort L, Talloen W, Shkedy Z, et al. (2015) Using transcriptomics to guide lead optimization in drug discovery projects: Lessons learned from the QSTAR project. Drug Discov Today 20:505-513.

70. Falcony G, Joossens M, Vieira SS, Wang J, Darzi Y, et al. (2016) Population-level analysis of gut microbiome variation. Science 352:560-564.

71. Quan Y, Li B, Sun YM, Zhang HY (2015) Elucidating pharmacological mechanisms of natural medicines by biclustering analysis of the gene expression profile: A case study on curcumin and Si-Wu-Tang. Int J Mol Sci 16:510-520.

72. Gupta M, Cheung CL, Hsu YH, Demissie S, Cupples LA, et al. (2011) Identification of homogeneous genetic architecture of multiple genetically correlated traits by block clustering of genome-wide associations. Journal of bone and mineral research 26:1261-1271.

73. Lopez FH, Santos HM, Capelo JL, Fdez RF, Glez PD, et al. (2015) Mass-Up: an all-in-one open software application for MALDI-TOF mass spectrometry knowledge discovery. BMC Bioinformatics 16:318.

74. Amar D, Yekutieli D, Maron-Katz A, Hendler T, Shamir R (2015) A hierarchical Bayesian model for flexible module discovery in three-way time-series data. Bioinformatics; 31:17-26.

75. Hartingan JA (1972) Direct Clustering of a Data Matrix. J Am Stat Soc 67:123-129

76. Jain AK, Murty MN, Flynn PJ (1999) Data clustering: a review. ACM Computing Surveys 31:264-323.