

Chapter 2

Data-Driven Analysis of Collections of Big Datasets by the Bi-CoPaM Method Yields Field-Specific Novel Insights

**Basel Abu-Jamous, Chao Liu, David J. Roberts, Elvira Brattico
and Asoke K. Nandi**

Abstract Massive amounts of data have recently been, and are increasingly being, generated from various fields, such as bioinformatics, neuroscience and social networks. Many of these big datasets were generated to answer specific research questions, and were analysed accordingly. However, the scope of information contained in these datasets can usually answer much broader questions than what was originally intended. Moreover, many existing big datasets are related to each other but have different detailed specifications, and the mutual information that can be extracted from them collectively has been not commonly considered. To bridge this gap between the fast pace of data generation and the slower pace of data analysis, and to exploit the massive amounts of existing data, we suggest employing

Professor Nandi is a Distinguished Visiting Professor at Tongji University, Shanghai, China. This work was partly supported by the National Science Foundation of China grant number 61520106006 and the National Science Foundation of Shanghai grant number 16JC1401300.

B. Abu-Jamous · C. Liu · A.K. Nandi (✉)
Department of Electronic and Computer Engineering,
Brunel University London, London, UK
e-mail: asoke.nandi@brunel.ac.uk

C. Liu
e-mail: chao.liu@brunel.ac.uk

B. Abu-Jamous
Department of Plant Sciences, University of Oxford,
South Parks Road, Oxford OX1 3RB, UK
e-mail: basel.abujamous@plants.ox.ac.uk

D.J. Roberts
Radcliffe Department of Medicine, University of Oxford,
Headington, Oxford OX3 9DU, UK
e-mail: david.roberts@ndcls.ox.ac.uk

D.J. Roberts
National Health Service Blood and Transplant, John Radcliffe Hospital,
Headington, Oxford OX3 9BQ, UK

data-driven explorations to analyse collections of related big datasets. This approach aims at extracting field-specific novel findings which can be revealed from the data without being driven by specific questions or hypotheses. To realise this paradigm, we introduced the *binarisation of consensus partition matrices (Bi-CoPaM)* method, with the ability of analysing collections of heterogeneous big datasets to identify clusters of consistently correlated objects. We demonstrate the power of data-driven explorations by applying the Bi-CoPaM to two collections of big datasets from two distinct fields, namely bioinformatics and neuroscience. In the first application, the collective analysis of forty yeast gene expression datasets identified a novel cluster of genes and some new biological hypotheses regarding their function and regulation. In the other application, the analysis of 1,856 big fMRI datasets identified three functionally connected neural networks related to visual, reward and auditory systems during affective processing. These experiments reveal the broad applicability of this paradigm to various fields, and thus encourage exploring the large amounts of partially exploited existing datasets, preferably as collections of related datasets, with a similar approach.

Keywords Data-driven analysis • Tunable consensus clustering • Bi-CoPaM • Heterogeneous datasets • Gene expression • fMRI

1 Introduction

Advances in data collection and storage have generated huge amounts of big data in various research fields. For instance, tens of thousands of gene expression datasets have been generated during the last couple of decades. Each one of these datasets typically measures the expression of thousands to tens of thousands of genes over a number of samples which ranges from two to few hundreds [1]. In neuroscience, a typical magnetic resonance imaging (MRI) scan, covering the whole human brain, consists of millions of voxels, i.e. unitary three-dimensional sections, representing

E. Brattico
Center for Music in the Brain (MIB), Department of Clinical Medicine,
Aarhus University, Aarhus, Denmark
e-mail: elvira.brattico@clin.au.dk

E. Brattico
The Royal Academy of Music Aarhus/Aalborg, Aarhus, Denmark

E. Brattico
AMI Centre, School of Science, Aalto University, Espoo, Finland

A.K. Nandi
The Key Laboratory of Embedded Systems and Service Computing,
College of Electronic and Information Engineering, Tongji University,
Shanghai, People's Republic of China

different parts of the brain. Moreover, the functional MRI (fMRI) technology yields even larger amounts of data as it usually generates hundreds of 3D images along the dimension of time for the same imaging object, i.e. the brain, despite the fact that its individual images have relatively lower resolution than the MRI's. Many other fields yield high-throughput data including social networks [43], and economics [20], to name a few.

Information engineering techniques, such as machine learning and statistical methods, aim at identifying meaningful findings from data by computational analysis. Such techniques are required to analyse any single big dataset where it is not feasible to interpret it manually. This has been done very frequently in the literature when a research group produces a big dataset, analyses it using information engineering techniques, identifies novel conclusions, and publishes the conclusions alongside the dataset for the use of the community. Despite the numerous benefits of this approach, there are a few issues that have arisen over the years and decades. One issue is that many of these big datasets were generated by experiments designed to answer specific research questions or to test specific research hypotheses. Such datasets are commonly analysed in a relatively focused manner while being driven by these questions and hypotheses. Although the big dataset may well answer the question in hand, it usually has a much larger potential than that. Such partially explored datasets have been rapidly accumulating in public and private data repositories.

Moreover, answering a given research question in many cases requires the generation of a big dataset designed differently from any existing dataset. Nonetheless, it may well be related to some existing datasets which do not fully suffice to answer that question. Because of that, many existing datasets are found to belong to the same context but with differences in some details such as the conditions and the resolution of their samples, the used technology, the number of replicates and others. These collections of related datasets definitely carry mutual information which may not be inferred except by analysing them collectively. However, differences amongst such related datasets make them, in many cases, uncombinable in a straightforward concatenating manner. Information engineering techniques that are able to analyse multiple heterogeneous, yet related, datasets collectively have become a necessity to unearth the valuable mutual information buried in these groups of datasets.

In other instances, although a very big single dataset is generated, it semantically consists of many sub-datasets. In other words, it can be split into many big datasets where each one these datasets is semantically an independent dataset. Nonetheless, such big sub-datasets are indeed related to each other in some sense. This is therefore another example where methods that can analyse multiple datasets collectively are required.

In summary, it is evident that what we have learnt from the huge amounts of available big data does not match the potential that such data possess, as individual datasets or as collections of related datasets. We believe that revisiting existing data with novel and able sophisticated computational methods that have exploratory capabilities is crucial to try to bridge the gap between the rapid pace of big data

generation and the slower pace of data analysis and information inference. Such efforts should be in collaborations between information engineers and the researchers related to the specific fields of the datasets. Importantly, such efforts should ideally be exploratory and data-driven instead of guided hypothesis-driven or question-driven. In other words, let the data reveal the information that they contain while we explore them without being guided by specific and focused questions. Indeed, this requires computational methods with exploratory capabilities.

One class of exploratory methods is unsupervised clustering. However, existing clustering and consensus clustering methods, such as the k-means, hierarchical clustering (HC) and relabelling and voting, are generally applicable to a single dataset, and even tend to generate different results when applied to the same dataset. This leaves the exploration of multiple heterogeneous datasets together quite unresolved.

We have taken steps to realise the data-driven paradigm of research more efficiently by introducing a consensus clustering method, the binarisation of consensus partition matrices (Bi-CoPaM) [2]. This method not only considers the consensus of the applications of different individual clustering methods, which is what existing consensus clustering methods do, it also has the ability to analyse multiple heterogeneous datasets collectively with the objective of identifying the subsets of objects which are collectively correlated in all of the given datasets. The datasets that the Bi-CoPaM receives should measure some features for the same set of objects. However, remarkably, the types and numbers of features in these individual datasets do not have to be the same neither does the technology used have to be the same. We also present two case studies in which we demonstrate the ability of such exploratory approach in finding novel conclusions of interest to the specific research communities relevant to the analysed datasets.

The first case study presents the analysis of forty different yeast gene expression datasets from a broad range of conditions such as the yeast cell-cycle, metabolic cycle and stress conditions. However, what is common to all of these forty datasets is that they are from the same species, the budding yeast *Saccharomyces cerevisiae*. Therefore, this exploratory experiment aims at discovering findings that are generally applicable to this species regardless of the particular conditions which it experiences. Indeed, the selection of the collection of datasets had a major role in defining that aim. The experiment identified two clusters of genes, one of them is novel, and few hypotheses regarding the regulation and control of core biological processes in yeast, namely, ribosome biogenesis and stress response [4].

The second case study presents the analysis of a large fMRI dataset which semantically consists of 1856 datasets, each one of them is a big dataset itself. Again, we demonstrate the exploratory power of our Bi-CoPaM method in extracting field-specific conclusions. Three functionally connected neural networks are identified related to visual, reward, and auditory systems during affective processing; this is one of the few studies that have observed these networks using a data-driven approach.

2 Methods

2.1 Bi-CoPaM

The Bi-CoPaM method is a consensus clustering method which can be applied to a set of datasets that quantify various sets of features for the same set of objects. The number and type of features (e.g. time points, conditions or stages) in these datasets does not have to be similar, as long as these datasets quantify some features for the same set of objects. The objective of the Bi-CoPaM is to identify the subsets (clusters) of objects which are consistently correlated with each other in all given datasets. Many objects may be left unassigned to any of the clusters if they do not meet the relevant external specification.

The Bi-CoPaM method consists of four main steps detailed in [2] and is summarised here:

1. Individual partitions generation

Various clustering methods are independently applied to each one of the datasets. For M different methods and L different datasets, this step generates $C = M \times L$ partitions (clustering results). All of these clustering methods should be applied to generate the same number of clusters (K).

2. Relabelling

Because clustering methods are unsupervised and their resulting clusters are unlabelled, the i th cluster of one of the C partitions may not necessarily correspond to the i th cluster of the other partitions. This issue is resolved by the relabelling step as it reorders the clusters in the partitions to be aligned. The min-max [2] and the min-min [4] techniques were proposed to realise the relabelling step, and the latter is what has been considered in the case studies in this paper.

3. Fuzzy consensus partition matrix (CoPaM) generation

The relabelled partitions are combined by membership averaging to produce a fuzzy consensus partition matrix (CoPaM). The CoPaM associates each object with each cluster with a fuzzy membership value between zero and unity, where a zero membership means the object does not belong to the cluster, a unity membership means it fully belongs to it, and a value between zero and one indicates a partial belongingness.

4. Binarisation

The fuzzy CoPaM is binarised by one of six techniques defined in [2] to assign objects to clusters with binary (crisp) values, i.e. an object either belongs or does not belong to any given cluster. Out of the six techniques, the technique which was adopted in all of the case studies here is the *difference threshold binarisation* (DTB) technique, which leaves the objects that do not significantly belong to a single cluster unassigned from all clusters. A tuning parameter $\delta \in [0, 1]$ controls the strictness of the DTB technique, which results in tighter but small clusters when increased, and results in larger but looser clusters when decreased.

2.2 *M-N Scatter Plots Technique*

As can be seen from the description of the Bi-CoPaM method above, two parameters need to be set, namely the number of clusters (K) and the tightness tuning parameter δ . Varying these parameters result in different results with clusters that vary widely in their size and tightness. The M-N scatter plots technique addresses this aspect by selecting the few clusters that best fulfill some quality criteria amongst all of those clusters generated by multiple applications of Bi-CoPaM to the same datasets with various values of K and δ .

The M-N scatter plot is a 2D plot on which each cluster is represented by a single point [5]. Its horizontal axis represents the dispersion (looseness) of the clusters measured by a modified version of the mean-squared error (MSE), while the vertical axis represents the size of the clusters measured by a scaled value of the logarithm of the number of objects included in the clusters.

The interest here is to select larger clusters with minimum dispersion, i.e. with greater tightness. Therefore, better clusters are those which maximise their value on the vertical axis while minimising it on the horizontal axis. In other words, better clusters are those which are closest to the top-left corner of the M-N scatter plot.

Consequently, the cluster whose point is closest to the top-left corner of the plot in Euclidean distance is selected first as the best cluster. Note that the clusters scattered on the plot are from various applications of the Bi-CoPaM over the same datasets, and thus many of them largely overlap. Therefore, all of the clusters which overlap with that first selected cluster are removed from the scatter plot. Then, out of the clusters remaining in the plot, the closest one to the top-left corner is selected as the second best cluster and all of those which overlap with it are removed from the plot. This iterative process is continued until no more clusters are left in the plots.

The result of this technique is a set of clusters listed in a decreasing order of their quality as measured by their distances from the top-left corner of the plot. In many applications, the first few clusters are considered meaningful while the rest of them are discarded. This can be decided based on the distribution of their distances and the specific requirements of the application.

3 Case Study I—Anti-correlated Gene Expression Profiles with Clusters of Yeast Genes

3.1 *Summary*

Eukaryotes are species with cells that have real nuclei and organelles, including all animals, plants, fungi and protists. Budding yeast (*S. cerevisiae*) is the most studied unicellular eukaryotic organism which has emerged as a model organism because of its easy manipulation and short cell-cycle, making it feasible to run rapid functional experiments. Many basic biological processes are shared by all eukaryotes; thus,

understanding the biology of yeast can help in understanding the biology of many other species, including humans.

To identify the core subsets of genes which are globally co-expressed (correlated in gene expression) in yeast, we applied the Bi-CoPaM to forty different gene expression datasets that span various conditions and were generated in different years and laboratories. This study, published in *BMC Bioinformatics* [4], identified two core clusters of genes which are consistently co-expressed over all forty datasets. More interestingly, the expression of those two globally co-expressed clusters were negatively correlated in all of the datasets. This observation strongly suggests that a common machinery may be regulating both clusters of genes globally but in opposite directions, i.e. while this regulatory machinery activates one of the two clusters of genes, it represses the genes in the other cluster. Furthermore, it was observed that the first cluster always goes up in expression under growth conditions and goes down under stress conditions, while the second cluster behaves oppositely. As expected, the first cluster was found to be significantly enriched with growth-related genes, particularly rRNA processing and ribosome biogenesis. However, the second cluster included many genes with unknown functions, which we predicted to have roles in stress response. Finally, we deduced hypotheses regarding the regulatory machinery which may be regulating both clusters in opposite directions.

The rest of this section describes the datasets and the experiments in more detail.

3.2 *Forty Genome-Wide Yeast Datasets from Various Conditions*

Table 1 lists the forty budding yeast datasets which were submitted to the Bi-CoPaM method alongside some description of them. It is clear that the datasets were generated in different years, by different authors, under different conditions, and with different numbers of features. The specific biological contexts to which these datasets belong include the cell-cycle, the metabolic cycle and various stresses. However, each one of these datasets has the expression profiles of the same 5,667 yeast genes, which represent the entire yeast genome as represented by the Affymetrix yeast genome 2.0 array.

3.3 *Two Globally Anti-correlated Clusters of Genes Identified in Yeast*

The Bi-CoPaM method identified two clusters of genes which were, respectively, named as C1 and C2, and which, respectively, include 257 and 47 genes out of the initial 5,667 genes. The average expression profiles of these genes in each one of the forty datasets are shown in Fig. 1.

Table 1 Budding yeast gene expression datasets

ID	GEO accession*	Year	N^	Description	References
D01	GSE8799	2008	15	Two mitotic cell-cycles (w/t)	Orlando et al. [41]
D02	GSE8799	2008	15	Two mitotic cell-cycles (mutated cyclins)	Orlando et al. [41]
D03	E-MTAB-643 ^S	2011	15	Response to an impulse of glucose	Dikicioglu et al. [18]
D04	E-MTAB-643 ^S	2011	15	Response to an impulse of ammonium	Dikicioglu et al. [18]
D05	GSE54951	2014	6	Response of <i>dal80Δ</i> mutant yeast to oxidative stress induced by linoleic acid hydroperoxide	- [#]
D06	GSE25002	2014	9	Osmotic stress response and treatment of transformants expressing the <i>C. albicans</i> Nik1 gene	- [#]
D07	GSE36298	2013	6	Mutations of OPI1, INO2 and INO4 under carbon-limited growth conditions	Chumnanpuen et al. [15]
D08	GSE50728	2013	8	120-h time-course during fermentation	- [#]
D09	GSE36599	2013	5	Stress adaptation and recovery	Xue-Franzén et al. [58]
D10	GSE47712	2013	6	Combinations of the yeast mediator complex's tail subunits mutations	Larsson et al. [32]
D11	GSE21870	2013	4	Combinations of mutations in DNUP60 and DADA2	- [#]
D12	GSE38848	2013	6	Various strains under aerobic or anaerobic growth	Liu et al. [35]
D13	GSE36954	2012	6	Response to mycotoxin type B trichothecenes	Suzuki and Iwahashi [51]
D14	GSE33276	2012	6	Response to heat stress for three different strains	Nakamura et al. [40]
D15	GSE40399	2012	7	Response to various perturbations (heat, myriocin treatment and lipid supplement)	Montefusco et al. [37]
D16	GSE31176	2012	6	W/t, <i>rlm1Δ</i> and <i>swi3Δ</i> cells with or without Congo Red exposure	Sanz et al. [46]
D17	GSE26923	2012	5	Varying levels of GCN5 F221A mutant expression	Lanza et al. [31]
D18	GSE30054	2012	31	CEN.PK122 oscillating for 2 h	- [#]
D19	GSE30051	2012	32	CEN.PL113-7D oscillating for 2 h	Chin et al. [14]
D20	GSE30052	2012	49	CEN.PL113-7D oscillating for 4 h	Chin et al. [14]
D21	GSE32974	2012	15	About 5 h of cell-cycle (w/t)	Kovacs et al. [30]

(continued)

Table 1 (continued)

ID	GEO accession*	Year	N^	Description	References
D22	GSE32974	2012	15	About 4 h of cell-cycle (mutant lacking Cdk1 activity)	Kovacs et al. [30]
D23	GSE24888	2011	5	Untreated yeast versus yeasts treated with <i>E. arvense</i> herbs from the USE, China, Europe or India	Cook et al. [16]
D24	GSE19302	2011	6	Response to degron induction for w/t and nab2-td mutant	González-Aguilera et al. [24]
D25	GSE33427	2011	5	Untreated w/t, and wt/t, <i>yap1Δ</i> , <i>yap8Δ</i> and double mutant treated with AsV	Ferreira et al. [21]
D26	GSE17716	2011	7	Effect of overexpression and deletion of MSS11 and FLO8	Bester et al. [10]
D27	GSE31366	2011	4	Presence and absence of mutli-inhibitors for parental and tolerant strains	- [#]
D28	GSE26171	2011	4	Response to patulin and/or ascorbic acid	Suzuki and Iwahashi [50]
D29	GSE22270	2011	4	PY1 and Met30 strains in room temperature or 35 °C	- [#]
D30	GSE29273	2011	4	Time-series during yeast second fermentation	- [#]
D31	GSE29353	2011	5	Different haploid strains growing in low glucose medium	Parreiras et al. [42]
D32	GSE21571	2011	8	Different combinations of mutations in HTZ1, SWR1, SWC2 and SWC5	Morillo-Huesca et al. [39]
D33	GSE17364	2010	4	Untreated w/t and Slt2-deficient yeasts, or treated with sodium arsenate for 2 h	Matia-González and Rodríguez-Gabriel [36]
D34	GSE15352	2010	8	24-h time-course of yeast grown under a low temperature (10 °C)	Strassburg et al. [48]
D35	GSE15352	2010	8	24-h time-course of yeast grown under a normal temperature (28 °C)	Strassburg et al. [48]
D36	GSE15352	2010	8	24-h time-course of yeast grown under a high temperature (37 °C)	Strassburg et al. [48]
D37	GSE16799	2009	21	UC-V irradiation of w/t, <i>mig3Δ</i> , <i>SNF1Δ</i> , <i>RAD23Δ</i> , <i>RAD4Δ</i> and <i>snf1Δrad23Δ</i>	Wade et al. [56]
D38	GSE16346	2009	4	BY474 cells grown to mid-log under presence versus absence of L-carnitine and/or H ₂ O ₂	- [#]

(continued)

Table 1 (continued)

ID	GEO accession*	Year	N^	Description	References
D39	GSE14227	2009	10	Two hours of wild-type yeast growth	[23]
D40	GSE14227	2009	9	Two hours of <i>sch9Δ</i> mutant yeast growth	[23]

*GEO accession: the unique accession number of the dataset in the GEO repository. All of these datasets are publically available at this repository

^SThis particular accession is not a GEO accession number; it is rather an ArrayExpress accession number which can be used to access this dataset at the EMBL-EBI ArrayExpress repository

^NThe number of features (biological conditions or time points) which the dataset covers

[#]No references are available for some datasets. However, they can still be accessed on the GEO repository through their GEO accession

Interestingly, C1 and C2 show negative correlation consistently in each of the 40 datasets. Correlation values between the average expression profiles of C1 and C2 in each one of the datasets are shown in Fig. 2, where it can be seen that in 38 datasets the correlation value does not exceed the -0.70 , and that it is still a negative (but less significant) value in the two other datasets. This suggests that both clusters may be regulated by a common regulatory machinery.

3.4 C1 is a Ribosome Biogenesis (RiB) Cluster While C2 is Predicted to be a Stress Response Cluster

Gene ontology (GO) term analysis [47], at the time of our original study [4], showed that C1 is enriched with genes known to participate in ribosome biogenesis ($p\text{-value} = 10^{-129}$). Ribosome biogenesis is required to build the ribosomes which synthesise proteins, and are therefore largely needed under growth conditions. By scrutinising the conditions under which the expression of C1 goes up across all 40 datasets, it can be clearly seen that expression of genes in the C1 cluster does go up under growth conditions, as predicted from the gene ontology analysis.

On the other hand, C2, which goes down under growth conditions and up under stress conditions, included some stress response genes (only 6 out 47 with a $p\text{-value}$ of 10^{-3}), and many genes with unknown or unrelated functions. We labelled this cluster as “anti-phase with ribosome biogenesis” (APha-RiB). Given these observations, we predicted that many more genes in C2 may have roles in stress response.

It is worth pointing out that a comparison with other sets of genes identified by other studies as up-regulated under stress or down-regulated under growth demonstrated that C2 is largely a novel cluster of genes with many genes previously not identified as such [4]. This is despite the fact that C2 is much smaller than those

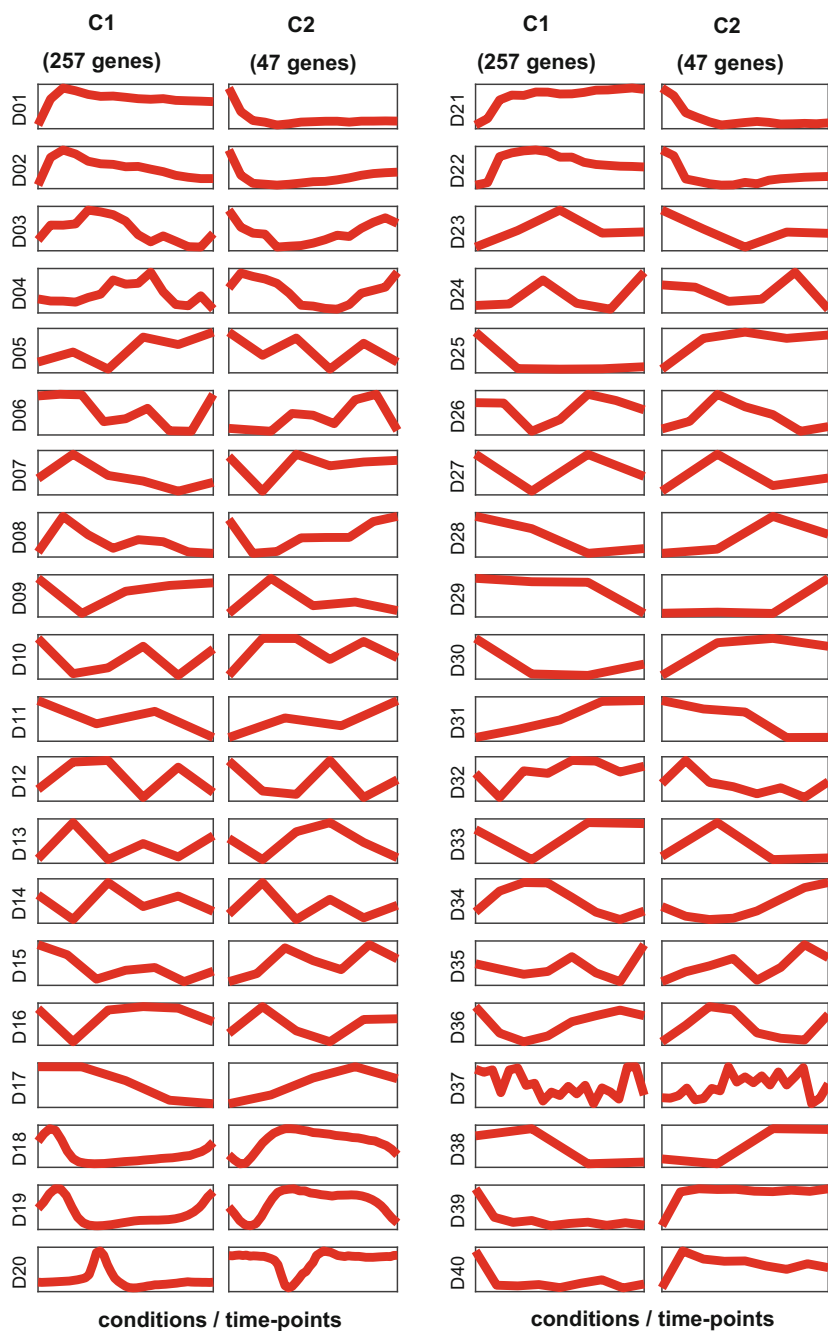


Fig. 1 The average expression profiles of the genes in clusters C1 and C2 over each one of the forty datasets D01 to D40

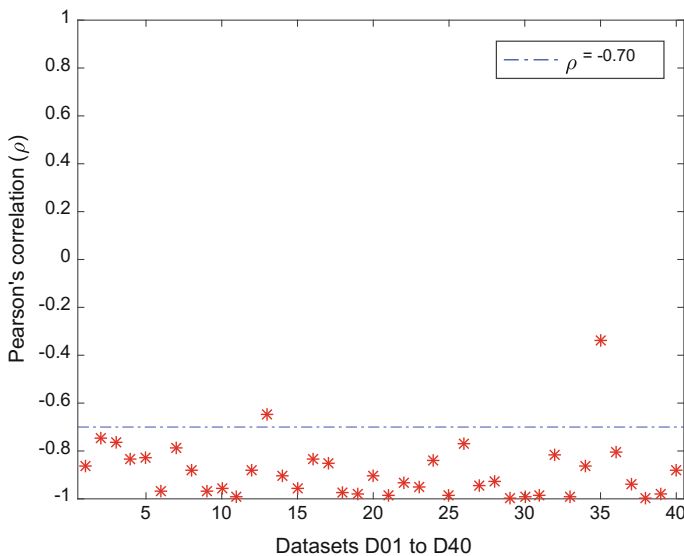


Fig. 2 Pearson's correlation values between the average profiles of C1 and C2 in each one of the forty datasets. The *horizontal dotted line* marks the correlation value of -0.70

other sets which include between 280 and 550 genes. Indeed, identifying a focused cluster of genes which is consistently anti-correlated with growth in forty datasets from broad conditions is more significant than identifying a large cluster of genes which are anti-correlated with growth in a single study. Many genes in the latter might well show such behaviour out of coincidence, while those in the former are highly expected to be genuinely controlled by growth and stress conditions. This demonstrates the power of our approach in identifying clusters of genes from collections of datasets.

3.5 Predicted Regulatory Network Controlling C1 (RiB) and C2 (APha-RiB)

We used the MEME [8] and the TOMTOM [25] tools to identify the regulatory elements (DNA motifs) and the regulators that may control C1 (RiB) and C2 (APha-RiB). Figure 3 summarises our findings, together with previous observation from the literature, in the form of a regulatory network.

Many of the regulatory interactions that we predicted agreed with what has been found in the literature. However, some other interactions were not recognised before our study and we hypothesised that they may exist. Notably, the transcription factor (TF) Stb3 was shown to repress RiB genes under stress, and we

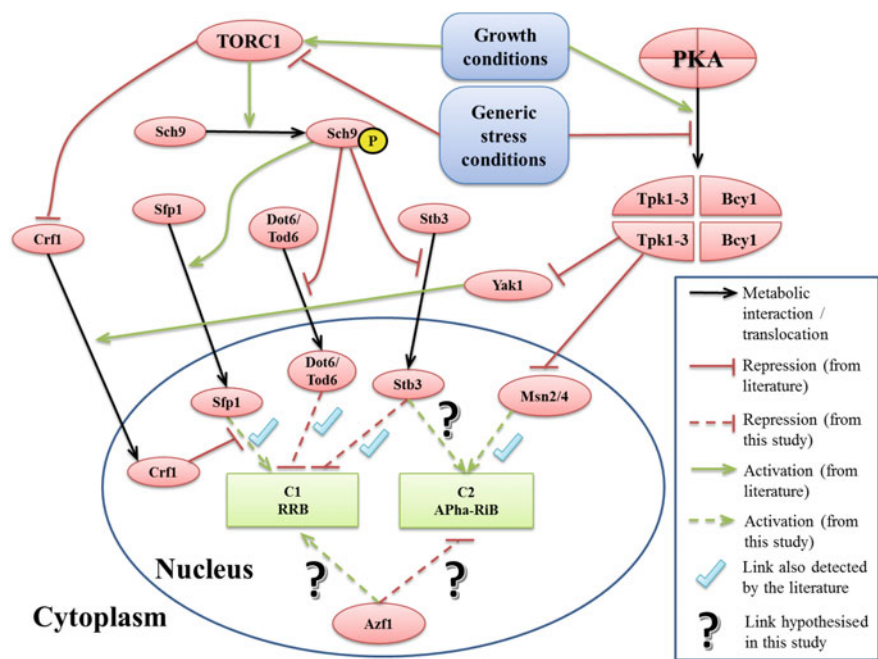


Fig. 3 Predicted regulatory network that controls both C1 and C2 while showing the regulatory interactions predicted in this study and the ones which were derived from the literature

predicted that it may have a role in activating APha-RiB genes at the same time. Also, we predicted that the regulator Azf1 may have a role in activating RiB genes while repressing APha-RiB genes under growth conditions.

4 Case Study II—Brain Functional Network Involved in Affective Processing

4.1 Background

Blood oxygen level dependent (BOLD) signals measured by fMRI datasets indirectly reflect the neural activities during cognitive tasks [22, 38] or during resting-state [33]. Obtaining these brain data is the first and fundamental step to capture the brain status. To further study the brain functions, information conveyed by brain data needs to be extracted and quantitatively analysed. Up to now, various fMRI data analysis methods have been developed such as statistical parametric mapping (SPM) [22], seed-based connectivity analysis, independent component analysis (ICA), multivariate pattern analysis (MVPA) [17] as well as clustering analysis [55]. It is difficult to state which method outperforms others, rather the choice of method depends on the design of fMRI experiment paradigm.

Traditional SPM method has been used for decades. The modelling of BOLD signal suits the fMRI experiments with stimuli having explicit timing and duration but when it comes to the resting-state, i.e. no stimuli are given during scanning, or very complex experiment paradigm, then modelling of these BOLD signals becomes very difficult or impossible. Recently, a study reported that many modern software packages incorporating SPM method for analysing neuroimaging data suffer from “bugs” that could invalidate many statistical results [19]. Seed-based connectivity analysis searches the voxels showing similar BOLD fluctuations to the predefined seed area/voxel, which limits its application to regional analysis rather than the whole brain scale connectivity analysis.

ICA can decompose the observed BOLD signals into a set of independent sources/components, which could be mixed to reconstitute the observed signals. However, the underlying assumption that the components are mutually independent is debatable. Also the components yielded by the analysis are not easy to interpret and still less validated. MVPA is based on supervised classification that detects differences between conditions with higher sensitivity than conventional univariate analysis (e.g. SPM) by focusing on the analysis and comparison of distributed patterns of activity (multivariate), but it also only suits the fMRI experiments with known and well separated stimuli due to the supervised classification nature. Clustering approach, which belongs to unsupervised classification, attempts to divide the data objects into clusters where objects in the same cluster are more similar to each other than objects from different clusters. The unsupervised feature makes clustering a very flexible and powerful tool of exploring and analysing the neuroimaging data. It has gained a lot of popularity in connectivity analysis and brain parcellation [34, 52].

Clustering itself is an ill-posed problem where the solution violates at least one of the common assumptions regarding cluster consistency, richness and scale invariance [26]. Almost every clustering algorithm explicitly or implicitly assumes a certain data model based on which the data is clustered [49]. Without a priori knowledge, each clustering algorithm could produce plausible results based on its underlying assumption of the data structures that might not be satisfied by the input data. To match the assumed data model and the true input data structure, it is crucial to obtain the correct a priori information to obtain good clustering results; however, such a priori information is often hard to obtain even from experts. A recent main criticism from the neuroscience field relies on the wide variety of analysis strategies, combined with small sample sizes, used to investigate regional brain activity measured with fMRI that frequently lead to inconsistent findings [13].

To answer these common problems in clustering analysis, we explored the consensus clustering analysis (Bi-CoPaM) of fMRI data and its benefits compared to single clustering methods such as K-means, hierarchical clustering and self-organising map (SOM). The validation results show that there are obvious differences among the clustering solutions by different single clustering algorithms as well as the different combination of single algorithms. Bi-CoPaM paradigm can not only find the objects that consistently belong to same cluster across many clustering experiments but also is able to extract the clusters according to their

tightness (quality) through tuning the consensus level. These selected non-trivial clusters have very strong inner similarity as well as relatively large number of members, making it more reliable to draw neuroscientific conclusions.

4.2 *fMRI Experiment*

The fMRI data comes from an experiment studying human brain responses to music with different preference and emotions. All the music stimuli are divided into four categories that are liked happy (LH) [this represents the case of when a participant considers the music is “happy” and “likes” it], liked sad (LS) [this represents the case of when a participant considers the music is “sad” and “likes” it], disliked happy (DH) [this represents the case of when a participant considers the music is “happy” and “dislikes” it], and disliked sad (DS) [this represents the case of when a participant considers the music is “sad” and “dislikes” it]. The participants were from two groups that are musicians, who have received systematically music performance training, and non-musicians. A total of 13 musicians and 16 non-musicians participated in the fMRI scanning sessions. The whole fMRI experiment for one participant consists of 450 whole brain scans ($TR = 3$ s) while listening to one of the 32 music categories with each one repeated twice. Each scan covering the whole brain contains 228,453 voxels after pre-processing. In total, there are 1856 fMRI data segments corresponding to $1856 (= 29 \times 32 \times 2)$ listening trials. The study was approved by the ethical committee of the Helsinki University Central Hospital and complied with the Helsinki Declaration. The dataset is a subset of a larger data collection, parts of which have been published in [11, 44].

4.3 *Clustering Experiment*

The K-means, Hierarchical and SOM were applied on each of 1856 excerpts data (normalised to 0 mean and unit variance) with K equals to 10, 25, 50 and 100. These clustering results generated by different algorithms with four different cluster numbers were combined using the Bi-CoPaM paradigm for each possible combination of three methods, followed by cluster ranking and selection by the M-N scatter plot technique, yielding the consensus clustering results. In total, seven sets of final clustering results were obtained, i.e. clustering results from K-means, hierarchical clustering, SOM, K-means and hierarchical clustering, K-means and SOM, hierarchical clustering and SOM, and a combination of all the three methods. The brain regions within clusters were extracted by using automated anatomical labelling (AAL) atlas [54] and names of these regions were used to identify the neuroscience background of clusters.

4.4 Comparing Clustering Results Among Different Combination Scenarios

We chose the clusters covering three important areas in human brain, namely the auditory system, reward system and visual system, to evaluate the accuracy and completeness of clustering results. To measure quantitatively the similarity between two clusters, Jaccard index is used. Table 2 shows the Jaccard index between different method combinations for the cluster covering visual area. Table 3 shows the Jaccard index between different method combinations for the cluster covering reward system and Table 4 shows the Jaccard index between different method combinations for the cluster covering auditory system. KM represents K-means and HC represents hierarchical clustering. Figures 4, 5 and 6 are the visualisations of different clustering results.

Note that for those methods or method combinations having all zero Jaccard index with other clusters means that this method or method combination was not able to detect the corresponding cluster. For example, in Table 4, SOM method did not group the voxels within the auditory system into a cluster; neither did the combination of KM and HC. The combination of all the three methods never failed to detect any cluster.

4.5 Robustness Test Against Subject Functional Data Variability

Considering that the brain activity to a certain stimulus varied among different individuals, it raises the problem whether the algorithm could perform well regarding the functional data variation among participants. Since it has been demonstrated that Bi-CoPaM with three methods provides the most complete set of clusters, we carried out the test to investigate the robustness of Bi-CoPaM with three clustering methods against data variability on afore-mentioned three clusters. We generated two groups of subsets for test. One was created by randomly selecting 75% of the musicians (10 out of 13) and non-musicians (12 out of 16) as well as 75% of the excerpts for each participant, which yields a subset consisting approximately 56% of all the data from the fMRI experiment. The above random selection was repeated 10 times and these 10 subsets formed group A. Similarly, we chose a different ratio of 90% of the musicians (12 out of 13) and non-musicians (14 out of 16) as well as 90% of the excerpts for each participant, when the data were randomly selected. Repeating the selection with new ratio for 10 times formed group B consisting approximately 80% of all the data. Then we applied Bi-CoPaM with three clustering algorithms on these subsets and recorded the clustering results.

From Tables 5 and 6, we could see, in most of the trials, the three important clusters (Visual, Reward and Auditory) could be identified despite different subsets being used. Meanwhile, with the increasing proportion of data used, the results

Table 2 Jaccard index for the cluster covering visual area

	KM	HC	SOM	KM + HC	KM + SOM	HC + SOM	KM + HC + SOM
KM		0.48	0.70	0.70	0.78	0.73	0.40
HC	0.48		0.38	0.57	0.53	0.56	0.26
SOM	0.70	0.38		0.58	0.66	0.62	0.50
KM + HC	0.70	0.57	0.58		0.76	0.84	0.32
KM + SOM	0.78	0.53	0.66	0.76		0.82	0.37
HC + SOM	0.73	0.56	0.62	0.84	0.82		0.35
KM + HC + SOM	0.40	0.26	0.50	0.32	0.37	0.35	

Table 3 Jaccard index for the cluster covering reward system

	KM	HC	SOM	KM + HC	KM + SOM	HC + SOM	KM + HC + SOM
KM		0.00	0.56	0.41	0.53	0.52	0.53
HC	0.00		0.00	0.00	0.00	0.00	0.00
SOM	0.56	0.00		0.43	0.57	0.55	0.57
KM + HC	0.41	0.00	0.43		0.33	0.33	0.37
KM + SOM	0.53	0.00	0.57	0.33		0.67	0.61
HC + SOM	0.52	0.00	0.55	0.33	0.67		0.60
KM + HC + SOM	0.53	0.00	0.57	0.37	0.61	0.60	

Table 4 Jaccard index for the cluster covering auditory system

	KM	HC	SOM	KM + HC	KM + SOM	HC + SOM	KM + HC + SOM
KM		0.49	0.00	0.00	0.41	0.71	0.62
HC	0.49		0.00	0.00	0.35	0.51	0.56
SOM	0.00	0.00			0.00	0.00	0.00
KM + HC	0.00	0.00			0.00	0.00	0.00
KM + SOM	0.41	0.35	0.00	0.00		0.48	0.40
HC + SOM	0.71	0.51	0.00	0.00	0.48		0.65
KM + HC + SOM	0.62	0.56	0.00	0.00	0.40	0.65	

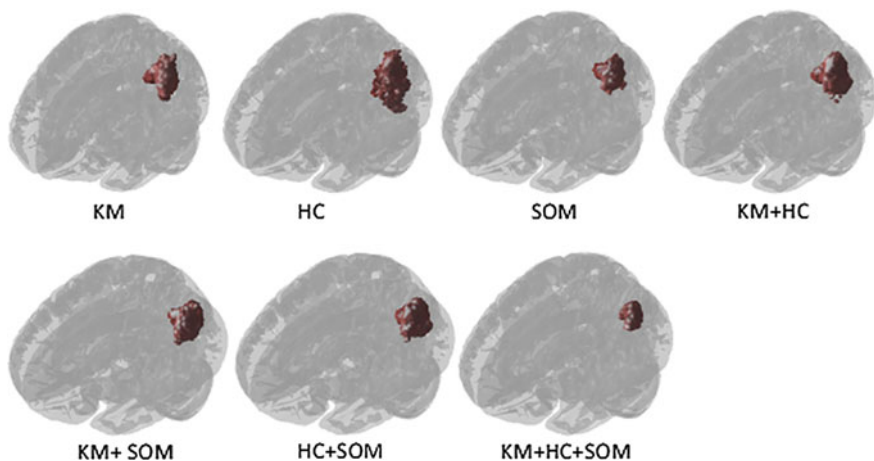


Fig. 4 3D illustrations of clusters covering visual area detected by seven different method combinations

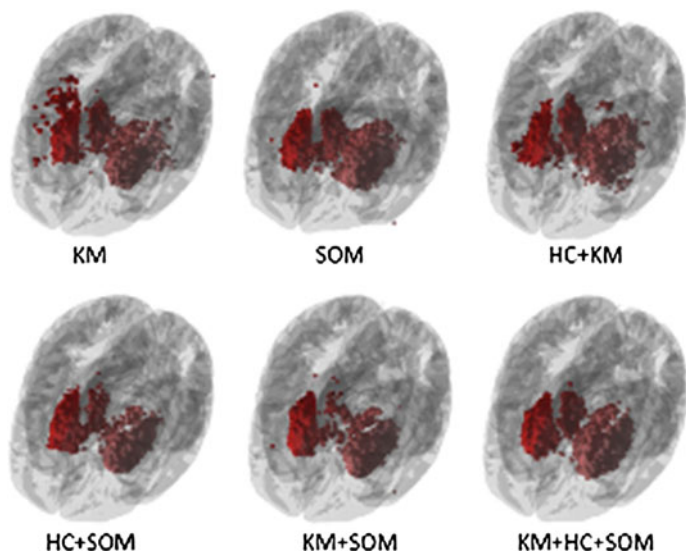


Fig. 5 3D illustrations of clusters covering reward system detected by six different method combinations

become more stable. For example, the Reward cluster was missed three times in the first experiment reported in Table 5, but not missed at all in the experiment reported in Table 6.

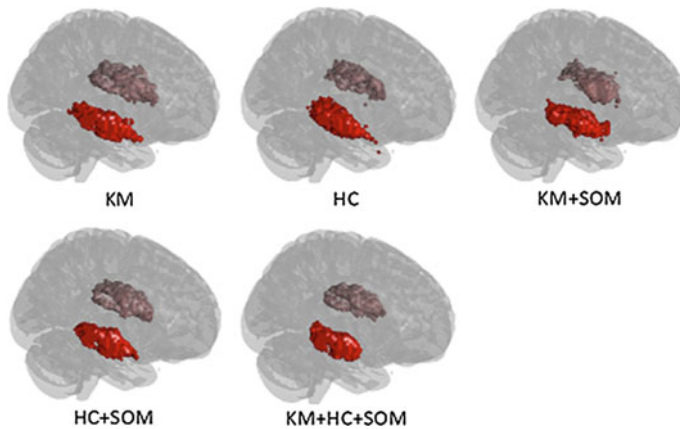


Fig. 6 3D illustrations of clusters covering auditory system detected by five different method combinations

4.6 Topology of Final Clusters

From Sects. 4.4 and 4.5, it can be concluded that the consensus scenario with all the three clustering algorithms gives relatively the best solution among all the possibilities of method combination. Figure 7 illustrates the clusters from combining all the three clustering algorithms in details such as their topologies within anatomical brain structures and the area information within each cluster.

Cluster A comprises bilateral visual areas, namely the calcarine fissure and the cuneus. Cluster B comprises bilateral neural structures of the reward system, namely the ventral striatum—extending to the globus pallidus—the thalamus, the amygdala, the orbitofrontal cortex and the left insula. Cluster C comprises the auditory areas, namely the bilateral superior temporal gyrus, Heschl’s gyrus, the left middle temporal gyrus, as well as one region of the somatosensory cortex, namely the right Rolandic operculum. These areas correspond well with the model-based literature [11].

4.7 Summary of Case Study II

In this study, Bi-CoPaM was used to combine the clustering results on a real fMRI data with multiple clustering algorithms to improve the consistency and stability performance of clustering in neuroimaging data analysis. The results show that the proposed method has great capability and offer much potential to help address the consistency issue in clustering analysis of neuroimaging data.

Table 5 Results of test group A (56% of the full data). Final size is the obtained cluster size using the whole participants' data. Trial size is the cluster size obtained in different trials. Intersection is the size of the part that the clusters in each trial intersect with the clusters obtained using the whole participants' data

Clusters		Trial										Mean
		1	2	3	4	5	6	7	8	9	10	
Visual	Final size	410	410	410	410	410	410	410	410	410	410	
	Trial size	551	1031	961	900	924	1838	713	877	2027	651	
	Intersection	357	410	410	410	405	410	396	409	410	393	
	Dice Coeff.	0.74	0.57	0.60	0.63	0.61	0.36	0.71	0.64	0.34	0.74	0.594
Reward	Final size	4293	4293	4293	4293	4293	4293	4293	4293	4293	4293	
	Trial size	2857	2591	4646	0	3263	1028	0	1620	0	4168	
	Intersection	1479	2046	3218	0	2278	915	0	1403	0	3303	
	Dice Coeff.	0.41	0.59	0.72	0	0.60	0.34	0	0.47	0	0.78	0.391
Auditory	Final size	2123	2123	2123	2123	2123	2123	2123	2123	2123	2123	
	Trial size	2381	3477	1452	2980	1576	2191	1801	2161	856	2006	
	Intersection	1738	2021	16	1995	439	1581	1459	1343	797	1673	
	Dice Coeff.	0.77	0.72	0	0.78	0.24	0.73	0.74	0.63	0.54	0.81	0.596

Table 6 Results of test group B (80% of the full data)

Clusters		Trial										Mean
		1	2	3	4	5	6	7	8	9	10	
Visual	Final size	410	410	410	410	410	410	410	410	410	410	
	Trial size	591	991	385	821	803	1119	784	1216	821	803	
	Intersection	402	410	326	406	406	410	407	410	406	406	
	Dice Coeff.	0.80	0.59	0.82	0.66	0.67	0.54	0.68	0.50	0.66	0.67	0.66
Reward	Final size	4293	4293	4293	4293	4293	4293	4293	4293	4293	4293	
	Trial size	3028	2867	3118	4381	1813	4390	4426	2362	4381	1813	
	Intersection	1273	1027	2467	3308	1519	3157	3461	1891	3308	1519	
	Dice Coeff.	0.35	0.29	0.67	0.76	0.50	0.73	0.79	0.57	0.76	0.50	0.592
Auditory	Final size	2123	2123	2123	2123	2123	2123	2123	2123	2123	2123	
	Trial size	476	1948	2910	1348	2496	3158	3036	771	1348	2496	
	Intersection	458	1548	1997	338	1855	2051	1704	578	338	1855	
	Dice Coeff.	0.35	0.76	0.79	0.19	0.8	0.78	0.66	0.40	0.19	0.8	0.572

From Tables 2, 3 and 4, which show the mutual differences between any two method combination scenarios, and the three-dimensional demonstration of all the clusters covering three important neural systems in human brain, we can infer the obvious differences among different clustering schemes. For the cluster covering the visual system, although all of the method combinations were able to detect this area, details of the regions such as the size and shape vary among each other. For the cluster covering the reward system and another cluster covering the auditory system, not only the results are different from each combination of clustering algorithms but also some single methods failed to detect the important brain structures based on their BOLD response time profiles. However, when it comes to the results obtained from any combined results from different single methods, these two important clusters could be identified, and only one scenario (the combination of KM and HC) failed to form the auditory cluster. This demonstrates the power of integrating the results from multiple clustering methods. As for the individual functional data variability test, Tables 5 and 6 demonstrate that Bi-CoPaM can detect the three important clusters in most of the trials consisting only part of all the data. On one hand, this demonstrates that the Bi-CoPaM is robust to variability of participants' data and thus generates reproducible results; on the other hand, it also shows the benefits of using a large number of subjects for more reliable results in data-driven analysis of functional brain imaging data.

From the neuroscientific aspect, the innovative use of Bi-CoPaM allows us to find clusters including functionally and anatomically related neural networks responding to emotional music. After the cluster generation and selection, emotion-related brain structures responsible for rewarding and pleasurable sensations such as the basal ganglia, thalamus, insula [45] and other areas involved with processing of auditory features such as the Heschl's gyrus, the Rolandic operculum and the superior temporal gyrus [7] were grouped into corresponding clusters separately. One of the most important findings of this study is that, without any predetermined model assigning a value to each stimulus, the Bi-CoPaM was able to obtain a single cluster including the anatomically connected sub-cortical and cortical structures of the reward circuit, responding to music with various emotions and preferences. This is one of the few studies obtaining such finding with a data-driven method. For example, a recent study using a data-driven network science method to study affective music processing [57], did not observe any reward circuit activity. As a matter of fact, neural structures of the reward circuit have also been found to be more or less connected only in other functional connectivity analysis studies such as one studying attention—deficit/hyperactivity disorder (ADHD) [53]. Our study confirms findings on the neural structures related to musical emotions obtained with model-based approaches [9, 11, 12, 27–29, 45].

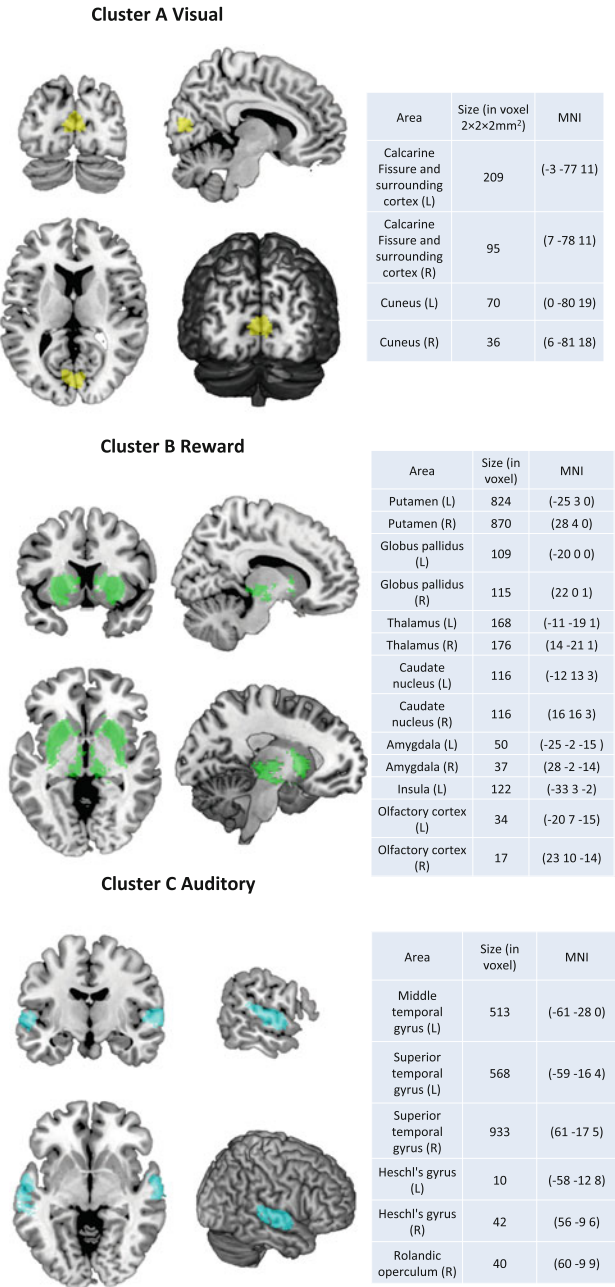


Fig. 7 The 3D illustrations of clusters and the size of each sub-cluster with voxels falling within a known anatomical brain structure, identified with the AAL atlas

5 Discussion

Data-driven exploration of multiple datasets collectively has the ability to find novel results that are not straightforwardly accessible through the analysis of individual datasets and serves in the exploitation of the huge amounts of existing partially exploited datasets. The Bi-CoPaM paradigm [2] serves as a powerful example of a method which can be used to achieve such goals as demonstrated by the two case studies presented.

The two chosen case studies were selected from two distinct fields of research with datasets of different types, sizes and dynamic ranges. This implies that the scope of the method's applicability is much broader than the contexts of bioinformatics data or fMRI data; it is rather, expected to be applicable to many other fields, as long as their data were arranged in the proper format.

It may be argued that the methods which are only applicable to a single dataset can be applied to collections of datasets if the datasets themselves were combined, by concatenation for example, to form a single bigger dataset. Although this approach may work in some cases, it is not applicable when the datasets have key differences in their format, the technology used, or the dynamic ranges of their values. In some cases, certain normalisation and standardisation techniques can help in transforming the individual datasets into a unified format with dynamic ranges that make it valid to combine them, but in many other cases this cannot be done. For instance, gene expression datasets generated using one-colour microarrays, two-colour microarrays and next-generation sequencing cannot be trivially combined. Moreover, merging many datasets into a single dataset may result in an even larger single dataset that may not be feasibly analysed using the available computational power. In addition to that, the integrity of each dataset, as an independent dataset, sometimes is lost once it has been merged with other datasets. Taken together, the analysis of multiple datasets collectively can indeed be fruitfully achieved using the Bi-CoPaM paradigm even when they are heterogeneous and not trivially combinable.

Available datasets in the fields of bioinformatics and neuroscience are many more than what has been analysed by the Bi-CoPaM. Few other studies by the Bi-CoPaM have been already been published, such as the analysis of five *E. coli* bacterial datasets [6] and the analysis of two yeast cell-cycle datasets [3]. However, revisiting the collections of datasets from many other contexts is expected to yield new insights in biology and neuroscience as well as many other fields.

To conclude, we recommend that the large amounts of existing and emerging big datasets should be revisited with data-driven explorations to accelerate the pace of information inference. The Bi-CoPaM is indeed a method that has been shown to be able to achieve that goal in many applications and can be used widely to explore many more collections of datasets. Furthermore, efforts to design other sophisticated methods that can handle large amounts of heterogeneous datasets from other points of view, other than the Bi-CoPaM's, are required.

References

1. B. Abu-Jamous, R. Fa, A.K. Nandi, *Integrative cluster analysis in bioinformatics*, 1st edn. (Wiley, s.l., 2015)
2. B. Abu-Jamous, R. Fa, D.J. Roberts, A.K. Nandi, Paradigm of tunable clustering using binarization of consensus partition matrices (Bi-CoPaM) for gene discovery. *PLoS ONE* **8**(2), e56432 (2013)
3. B. Abu-Jamous, R. Fa, D.J. Roberts, A.K. Nandi, Yeast gene CMR1/YDL156W is consistently co-expressed with genes participating in DNA-metabolic processes in a variety of stringent clustering experiments. *J. R. Soc. Interface* **10**(81), 20120990 (2013)
4. B. Abu-Jamous, R. Fa, D.J. Roberts, A.K. Nandi, Comprehensive analysis of forty yeast microarray datasets reveals a novel subset of genes (APha-RiB) consistently negatively associated with ribosome biogenesis. *BMC Bioinform.* **15**, 322 (2014)
5. B. Abu-Jamous, R. Fa, D.J. Roberts, A.K. Nandi, UNCLES: method for the identification of genes differentially consistently co-expressed in a specific subset of datasets. *BMC Bioinform.* **16**, 184 (2015)
6. B. Abu-Jamous, R. Fa, D.J. Roberts, A.K. Nandi, Application of the Bi-CoPaM method to five *Escherichia coli* datasets generated under various biological conditions. *J. Signal Process. Syst.* **79**(2), 159–166 (2015)
7. V. Alluri et al., Large-scale brain networks emerge from dynamic processing of musical timbre, key and rhythm. *NeuroImage* **59**(4), 3677–3689 (2012)
8. T.L. Bailey, C. Elkan, *Fitting a mixture model by expectation maximization to discover motifs in biopolymers* (AAAI Press, Menlo Park, CA, s.n., 1994), pp. 28–36
9. L. Barrett, T. Wager, The structure of emotion evidence from neuroimaging studies. *Curr. Dir. Psychol. Sci.* **15**(2), 79–83 (2006)
10. M.C. Bester, D. Jacobson, F.F. Bauer, Many *Saccharomyces cerevisiae* cell wall protein encoding genes are coregulated by Mss11, but cellular adhesion phenotypes appear only Flo protein dependent. *G3 (Bethesda)* **2**(1), 131–141 (2012)
11. E. Brattico et al., A functional MRI study of happy and sad emotions in music with and without lyrics. *Front. Psychol.* **2**(December), 308 (2011)
12. E. Brattico, B. Bogert, T. Jacobsen, Toward a neural chronometry for the aesthetic experience of music. *Front. Psychol.* **4**, 206 (2013)
13. K.S. Button et al., Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**(5), 365–376 (2013)
14. S.L. Chin, I.M. Marcus, R.R. Klevecz, C.M. Li, Dynamics of oscillatory phenotypes in *Saccharomyces cerevisiae* reveal a network of genome-wide transcriptional oscillators. *FEBS J.* **279**(6), 1119–1130 (2012)
15. P. Chumnanpuen, I. Nookaew, J. Nielsen, Integrated analysis, transcriptome-lipidome, reveals the effects of INO-level (INO2 and INO4) on lipid metabolism in yeast. *BMC Syst. Biol.* **7** (Suppl 3), S7 (2013)
16. R. Cook et al., The *Saccharomyces cerevisiae* transcriptome as a mirror of phytochemical variation in complex extracts of *Equisetum arvense* from America, China, Europe and India. *BMC Genomics* **14**, 445 (2013)
17. F. De Martino et al., Combining multivariate voxel selection and support vector machines for mapping and classification of fMRI spatial patterns. *NeuroImage* **43**(1), 44–58 (2008)
18. D. Dikicioglu et al., How yeast re-programmes its transcriptional profile in response to different nutrient impulses. *BMC Syst. Biol.* **5**, 148–163 (2011)
19. A. Eklund, T. Nichols, H. Knutsson, Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *Proc. Natl. Acad. Sci.* **113**(28), 7900–7905 (2016)
20. G. Elliott, A. Timmermann, Forecasting in economics and finance. *Annu. Rev. Econ.* **8**, 81–110 (2016)
21. R.T. Ferreira et al., Arsenic stress elicits cytosolic Ca(2+) bursts and Crz1 activation in *Saccharomyces cerevisiae*. *Microbiology* **158**(Pt 9), 2293–2302 (2012)

22. K.J. Friston et al., Statistical parametric maps in functional imaging: a general linear approach. *Hum. Brain Mapp.* **2**(4), 189–210 (1995)
23. H. Ge et al., Comparative analyses of time-course gene expression profiles of the long-lived sch9Delta mutant. *Nucleic Acids Res.* **38**(1), 143–158 (2010)
24. C. González-Aguilera et al., Nab2 functions in the metabolism of RNA driven by polymerases II and III. *Mol. Biol. Cell* **22**(15), 2729–2740 (2011)
25. S. Gupta, J.A. Stamatoyannopoulos, T. Bailey, W.S. Noble, Quantifying similarity between motifs. *Genome Biol.* **8** (2007)
26. J. Kleinberg, An impossibility theorem for clustering. *Adv. Neural Inf. Process. Syst.* 446–453 (2002)
27. S. Koelsch, Towards a neural basis of music-evoked emotions. *Trends Cogn. Sci.* **14**(3), 131–137 (2010)
28. S. Koelsch, Brain correlates of music-evoked emotions. *Nat. Rev. Neurosci.* **15**(3), 170–180 (2014)
29. S. Koelsch et al., Investigating emotion with music: an fMRI study. *Hum. Brain Mapp.* **27**(3), 239–250 (2006)
30. L.A.S. Kovacs et al., Cyclin-dependent kinases are regulators and effectors of oscillations driven by a transcription factor network. *Mol. Cell* **45**(5), 669–679 (2012)
31. A.M. Lanza, J.J. Blazeck, N.C. Crook, H.S. Alper, Linking yeast Gcn5p catalytic function and gene regulation using a quantitative, graded dominant mutant approach. *PLoS ONE* **7**(4), e36193 (2012)
32. M. Larsson et al., Functional studies of the yeast med5, med15 and med16 mediator tail subunits. *PLoS ONE* **8**(8), e73137 (2013)
33. M. Lee, C. Smyser, J. Shimony, Resting-state fMRI: a review of methods and clinical applications. *Am. J. Neuroradiol.* **34**, 1866–1872 (2013)
34. C. Liu, B. Abu-Jamous, E. Brattico, A.K. Nandi, Towards tunable consensus clustering for studying functional brain connectivity during affective processing. *Int. J. Neural Syst.* **27**(2), 1650042 (2017)
35. Z. Liu et al., Anaerobic α -amylase production and secretion with fumarate as the final electron acceptor in *Saccharomyces cerevisiae*. *Appl. Environ. Microbiol.* **79**(9), 2962–2967 (2013)
36. A.M. Matia-González, M.A. Rodríguez-Gabriel, Slt2 MAPK pathway is essential for cell integrity in the presence of arsenate. *Yeast* **28**(1), 9–17 (2011)
37. D.J. Montefusco et al., Distinct signaling roles of ceramide species in yeast revealed through systematic perturbation and systems biology analyses. *Sci. Signal.* **6**(299), rs14 (2013)
38. M. Monti, Statistical analysis of fMRI time-series: a critical review of the GLM approach. *Front. Hum. Neurosci.* **5**(609), 28 (2011)
39. M. Morillo-Huesca, M. Clemente-Ruiz, E. Andújar, F. Prado, The SWR1 histone replacement complex causes genetic instability and genome-wide transcription misregulation in the absence of H2A.Z. *PLoS ONE* **5**(8), e12143 (2010)
40. T. Nakamura et al., Identification of a gene, FMP21, whose expression levels are involved in thermotolerance in *Saccharomyces cerevisiae*. *AMB Express* **4**, 67 (2014)
41. D.A. Orlando et al., Global control of cell-cycle transcription by coupled CDK and network oscillators. *Nature* **453**, 944–947 (2008)
42. L.S. Parreiras, L.M. Kohn, J.B. Anderson, Cellular effects and epistasis among three determinants of adaptation in experimental populations of *Saccharomyces cerevisiae*. *Eukaryot. Cell* **10**(10), 1348–1356 (2011)
43. E.D. Raj, L.D. Babu, An enhanced trust prediction strategy for online social networks using probabilistic reputation features. *Neurocomputing* **219**, 412–421 (2017)
44. S. Saarikallio et al., Dance moves reflect current affective state illustrative of approach–avoidance motivation. *Psychol. Aesthet. Creat. Arts* **7**(3), 296–305 (2013)
45. V. Salimpoor, R. Zatorre, Neural interactions that give rise to musical pleasure. *Psychol. Aesth. Creat. Arts* **7**(1), 62–75 (2013)

46. A.B. Sanz et al., Chromatin remodeling by the SWI/SNF complex is essential for transcription mediated by the yeast cell wall integrity MAPK pathway. *Mol. Biol. Cell* **23**(14), 2805–2817 (2012)
47. SGD, Term Finder tool (2014). <http://www.yeastgenome.org/cgi-bin/GO/goTermFinder.pl>
48. K. Strassburg et al., Dynamic transcriptional and metabolic responses in yeast adapting to temperature stress. *OMICS* **14**(3), 249–259 (2010)
49. A. Strehl, J. Ghosh, Cluster ensembles—a knowledge reuse framework for combining multiple partitions. *J. Mach. Learn. Res.* **3**, 583–617 (2002)
50. T. Suzuki, Y. Iwahashi, Gene expression profiles of yeast *Saccharomyces cerevisiae* sod1 caused by patulin toxicity and evaluation of recovery potential of ascorbic acid. *J. Agric. Food Chem.* **59**(13), 7145–7154 (2011)
51. T. Suzuki, Y. Iwahashi, Comprehensive gene expression analysis of type B trichothecenes. *J. Agric. Food Chem.* **60**(37), 9519–9527 (2012)
52. B. Thirion, G. Varoquaux, E. Dohmatob, J.-B. Poline, Which fMRI clustering gives good brain parcellations? *Front. Neurosci.* **8**, 1–13 (2014)
53. D. Tomasi, N. Volkow, Abnormal functional connectivity in children with attention-deficit/hyperactivity disorder. *Biol. Psychiatry* **71**(5), 443–450 (2012)
54. N. Tzourio-Mazoyer et al., Automated anatomical labeling of activations in SPM using a macroscopic anatomical parcellation of the MNI MRI single-subject brain. *NeuroImage* **15**(1), 273–289 (2002)
55. M. van den Heuvel, H. Pol, Exploring the brain network: a review on resting-state fMRI functional connectivity. *Eur. Neuropsychopharmacol.* **20**(8), 519–534 (2010)
56. S.L. Wade, K. Poorey, S. Bekiranov, D.T. Auble, The Snf1 kinase and proteasome-associated Rad23 regulate UV-responsive gene expression. *EMBO J.* **28**(19), 2919–2931 (2009)
57. R.W. Wilkins et al., Network science and the effects of music preference on functional brain connectivity: from Beethoven to Eminem. *Sci. Rep.* **4**, 6130 (2014)
58. Y. Xue-Franzén, J. Henriksson, T.R. Bürglin, A.P. Wright, Distinct roles of the Gcn5 histone acetyltransferase revealed during transient stress-induced reprogramming of the genome. *BMC Genom.* **14**, 479 (2013)

Frontiers in Electronic Technologies

Trends and Challenges

Prabaharan, S.R.S.; Thalmann, N.M.; Kanchana

Bhaaskaran, V.S. (Eds.)

2017, XV, 161 p. 88 illus., 64 illus. in color., Hardcover

ISBN: 978-981-10-4234-8