Cluster Analysis of Heterogeneous Genomic Data

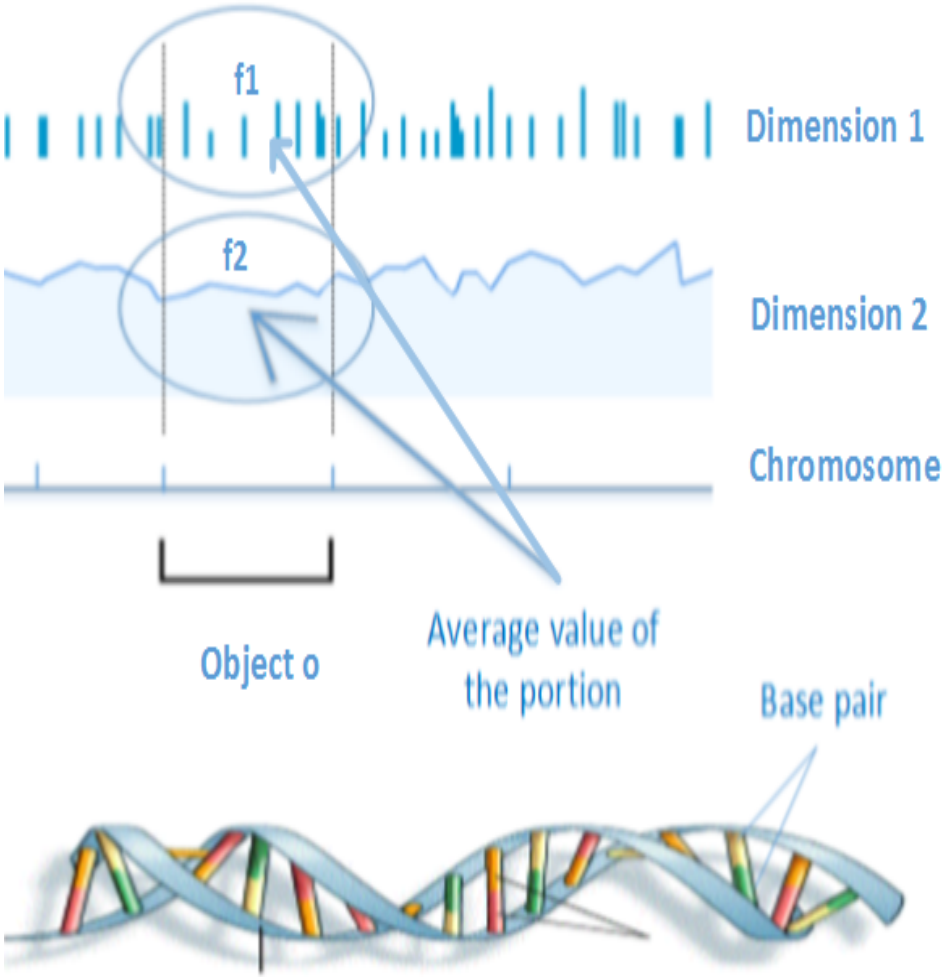
Phan Duc Thanh
supervised by
Christophe Rigotti (LIRIS, Beagle (EPC-INRIA), INSA Lyon)
in collaboration with
M. Leleu & J. Rougemont (BBCF, EPFL)

Master 2 Recherche en Intelligence Artificielle, Lyon

June 24, 2013

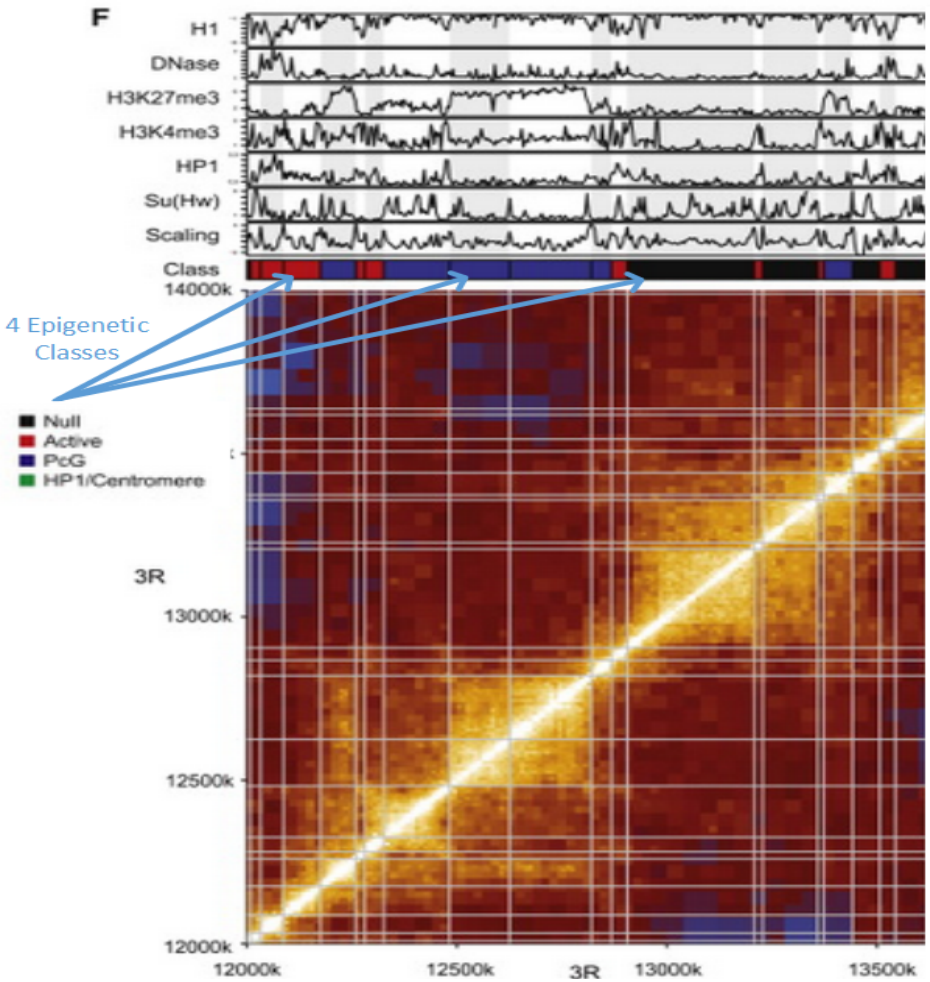
Introduction

Clustering of genomic data



- **Data :** a set of objects where
  - Objects : portions of chromosome
  - Dimensions : signals w.r.t some measure
  - Example : Object o is described by $\begin{pmatrix} f_1 \\ f_2 \end{pmatrix}$
- **Aims :** develop methods to find meaningful patterns for biologists
- **Advances in technologies :** availability of massive datasets of very high dimensionality

Meaningful clusterings : Sexton et al., Cell 2012



- **Sexton et al. 2012 :**
- partition of the genome into : 1169 *physical domains* using a statistical model and 3D contact map data
- clustering of the genome using these *physical domains* :
  - hierarchical with 315 dimensions (selected using chi-square tests)
  - k-means ($k = 4$) with 4 (manually selected) dimensions

Motivation

Observations :

- **Sexton et al. :** good result, interesting to biologists...
- heavily linked to initial partition of the genome (physical domains)
- require 3D contact maps : costly, not always available

Questions :

- finding meaningful cluterings without physical domains ?
- helping to select pertinent dimensions ? (among from a few dozens up to several hundreds) ?

Subspace clustering : extension of traditional clustering

Cluster analysis

- Unsupervised learning method of data exploration
- Similar objects into one group, dissimiliar objects into different groups

Definition of subspace clustering

- Finding clusters in *different relevant subspaces* of dimensions
- Given $(\mathbb{O}, \mathbb{S})$ the sets of objects and dimensions of the dataset, respectively
- Searching for set of *subspace clusters* $C = \{(O, D) | (O \subseteq \mathbb{O}, S \subseteq \mathbb{S}\}$
- *Clustering :* set of subspace clusters

Subspace clustering

Approaches :

- Grid-based : CLIQUE
- Density-based : SUBCLU
- Clustering-based : PROCLUS, SC-Kmeans

Advantages of subspace clustering

- ability to deal with high dimensionality
- might work when global dimensionality reduction/feature selection (PCA for instance) could not
- pertinent dimensions to "explain" the clusters found

Contributions

Proposed framework

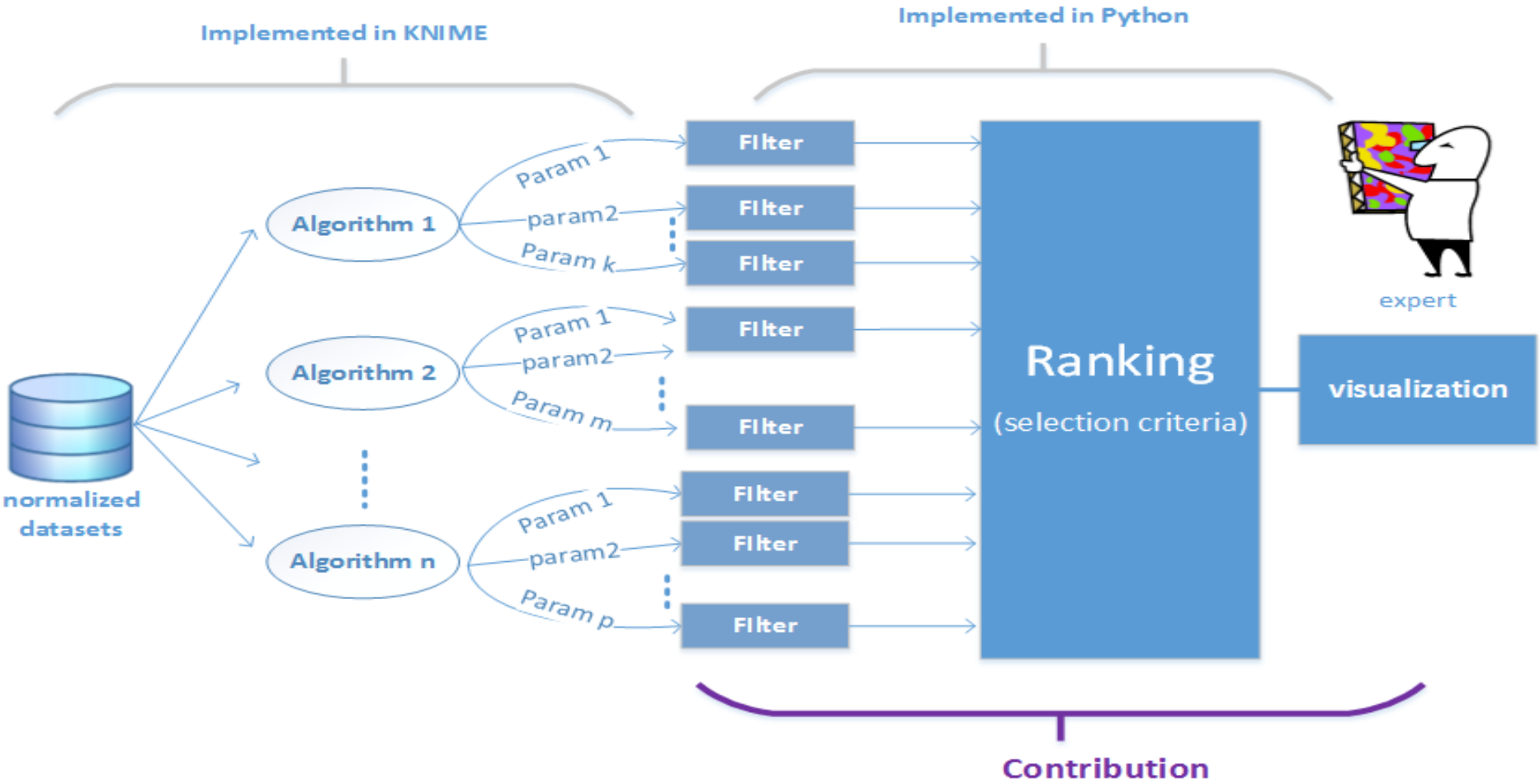


FIGURE : Proposed framework

Proposed clustering filtering process

Subspace clustering redundancies :

- overlapping allowed : "information overlapping-data coverage"
- highly undesirable :
  - little novel information
  - overwhelming to process
  - computationally-intensive
- many reasons : grid-based Apriori-like algorithms are redundancy-prone

Filtering process

- coverage filter : removal of clusters too small or too large
- adapted model from Gunnemann et al. 2011

Redundancy model

Ideas :

- cluster C_1 is redundant w.r.t cluster C_2 : $C_1 \prec_{f,red} C_2$ if :
  - C_1 is similar to C_2
  - the quality of C_1 is not as good as the quality of C_2

$\rightarrow$ we can remove C_1 without "losing" too much *information.*

Redundancy exemple

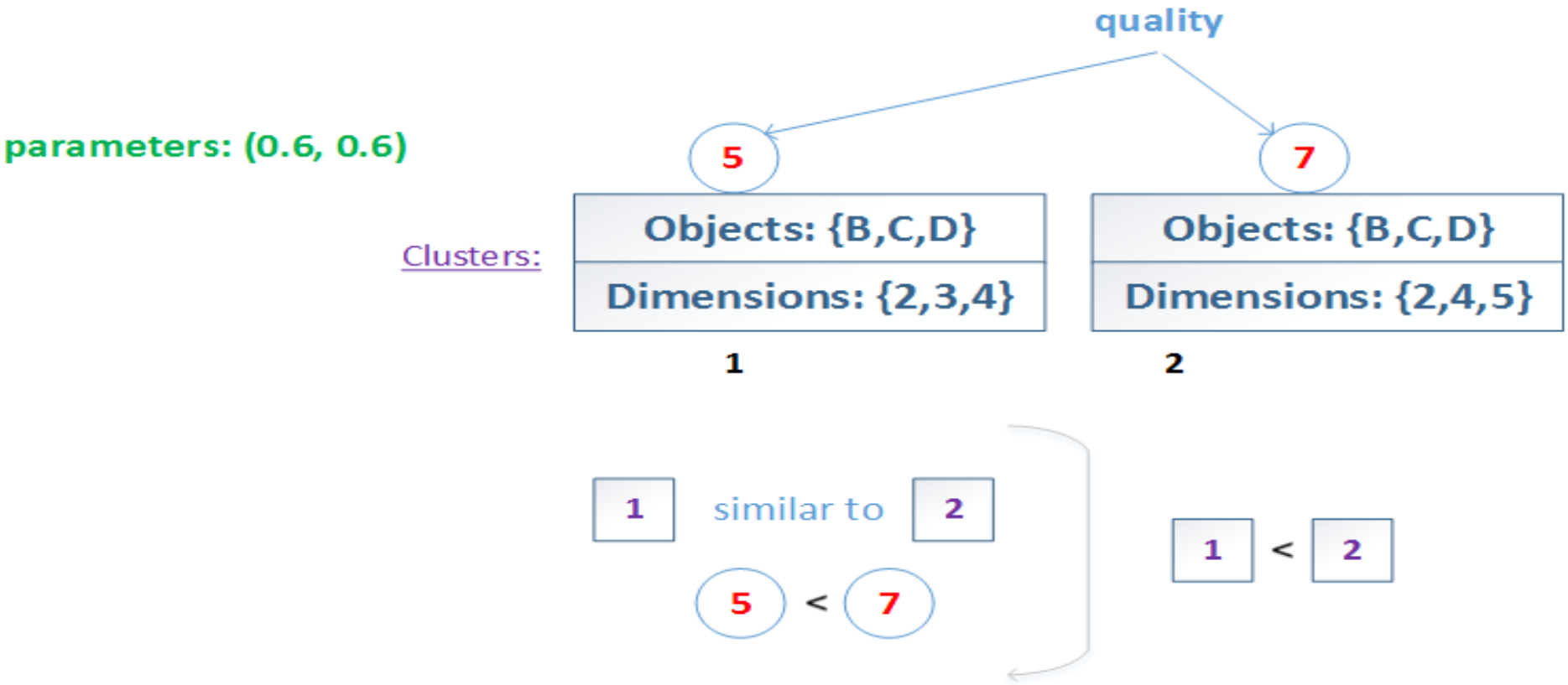


FIGURE : Cluster 1 is redundant w.r.t Cluster 2

Proposed measures for the ranking process

FIGURE : 1 clustering with 8 objects, 3 clusters (red, blue, yellow) : 4 color changes

- **Intuition :** In a good clustering, adjacent objects are likely to belong to the same cluster (same color) $\rightarrow$ only a few color changes

Proposed measures for the ranking process

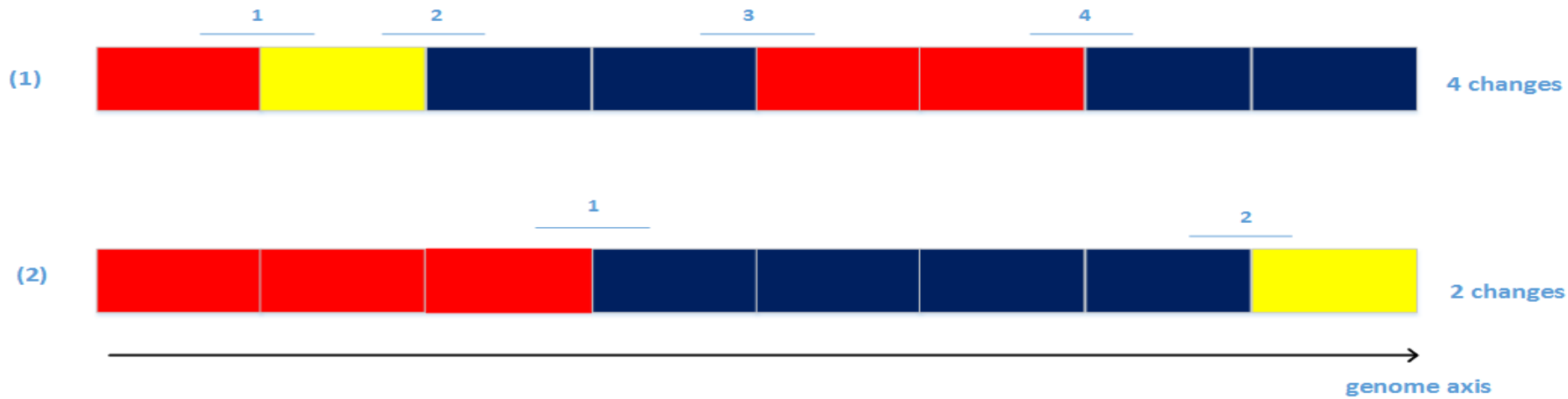


FIGURE : 2 clusterings with : 8 objects, 3 clusters (red, blue, yellow)

- **Intuition :** In good clustering, adjacent objects are likely to belong to the same cluster (same color) $\rightarrow$ **clustering 2 better then clustering 1**

Proposed measures for the ranking process

Spatial coherence measure : ratio version

$sc_{ratio} = \frac{nb}{E(nb)}$ where nb the random variable representing the number of color changes and $E(nb)$ its expected value under the assumptions that objects are randomly assigned to clusters.

Spatial coherence measure : difference version

$sc_{diff} = \frac{nb - E(nb)}{max(nb - E(nb))} = \frac{nb - E(nb)}{n - E(nb)}$, n : the total number of objects. $E(sc_{diff}) = 0$

Experiments

Data preparation and pre-processing

Datasets

- $9, 14$ dimensions, supersets of the 4 selected by Sexton et al., 2012
- 2000 objects of length 1000 base pair each, located on the portion from base pair $12e6$ to base pair $14e6$ of the Drosophila 3R chromosome

Data sources

- Fillion et al, 2009, retrieved from Gene Expression Omnibus platform
- Vital-IT, EPFL
- different formats

Pre-processing

- cleaning, convertion of retrieved datasets
- discretization of the genome portion into 2000 bin of size 1000 base pairs.
- normalization by log-quantiles : $x_{norm} = -2log(1 - \frac{\text{rank}(x_{raw})}{\max(\text{rank}(x_{raw}))})$

Clustering using KNIME components

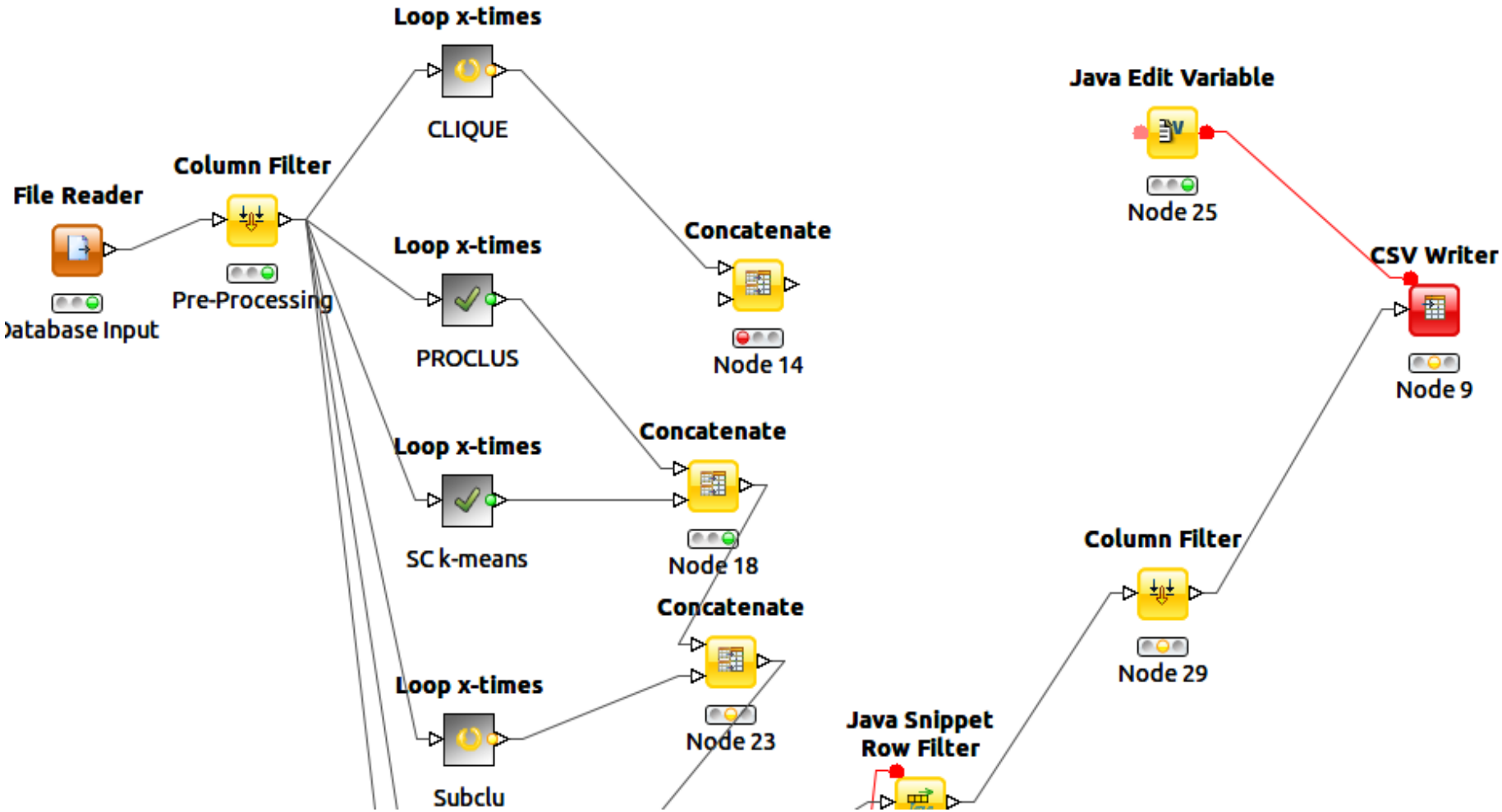


FIGURE : Subspace clustering algorithms employed

Clustering results

FIGURE : Visualization of some of the obtained clusterings, 9 dimensions

External quality measures

- Spatial coherence measure quality : comparison of our top-rank clusterings with a good reference one
- Taking results found from Sexton et al., Cell 2012 as reference clustering
- External measures :
  - **Objects only** : Entropy, F1-score (object purity and coverage)
  - **Objects and dimensions** : RNIA, CE

External quality measures

Rank for our Spatial coherence	Ranks for different external measures			
	RNIA	F1	CE	Entropy
1	9	5	58	4
2	5	15	35	7
3	49	40	273	26
4	3	4	3	6
5	9	2	15	3
6	2	6	2	8
7	4	8	5	2
8	18	22	49	14
9	93	17	127	28
10	23	1	28	1

TABLE : Top 10 out of 1552 total clusterings w.r.t spatial coherence (difference version) on 3R, along with the ranks w.r.t other measures

- Our top performers are close to the reference clustering (good rank and good score for external measures)

Spearman's rank correlation coefficients

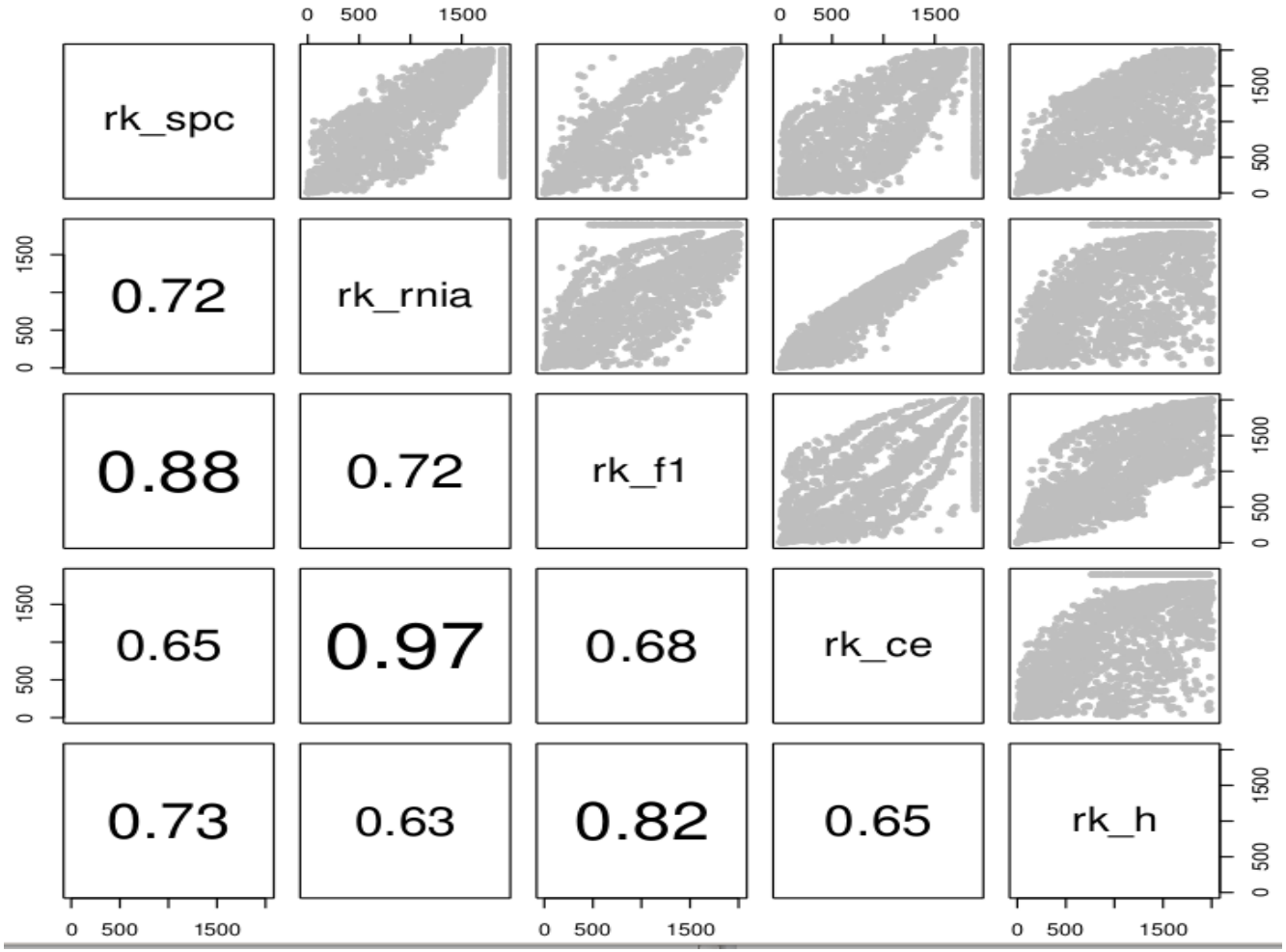


FIGURE : Scatter plot matrix between measures, 9 dimensions

Conclusions

- **Questions re-stated :**
- can we find meaningful cluterings without having to rely on "physical domain" information (expensive 3D contact map data) ?
- **Our results indicate :**
- it seems indeed possible to select interesting clusterings using our proposed framework and the measure of spatial coherence
- **Perspectives :**
- execution on other portions of the genome at different scales
- take into consideration other measures (V-Measure for external, p-value for spatial coherence)

References

- Sexton et al. 2003, "Three-dimensional folding and functional organization principles of the Drosophila genome." Cell 148.3 (2012) : 458.
- Muller et al 2009, . "Evaluating clustering in subspace projections of high dimensional data." Proceedings of the VLDB Endowment 2.1 (2009) : 1270-1281.
- Kriegel et al. "Clustering high-dimensional data : A survey on subspace clustering, pattern-based clustering, and correlation clustering." ACM Transactions on Knowledge Discovery from Data (TKDD) 3.1 (2009) : 1.
- Gunnemann et al. 2011, "DB-CSC : a density-based approach for subspace clustering in graphs with feature vectors." Machine Learning and Knowledge Discovery in Databases. Springer Berlin Heidelberg, 2011. 565-580.

Questions