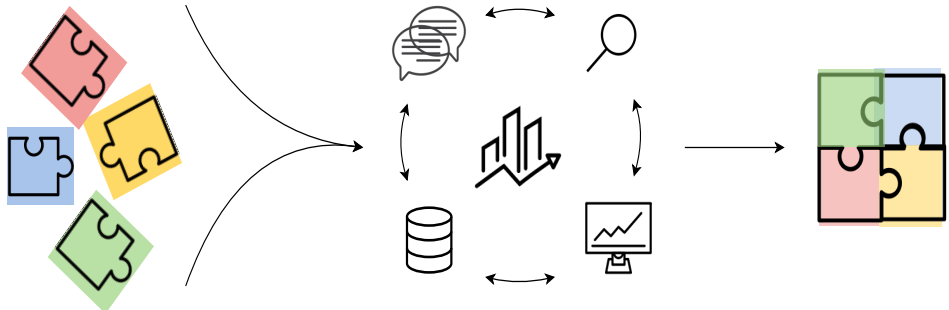
Multi–data learning

a multitude of ABC's

Marijke Van Moerbeke - Open Analytics

July 10 2019

Data integration



Illustration

Drug discovery

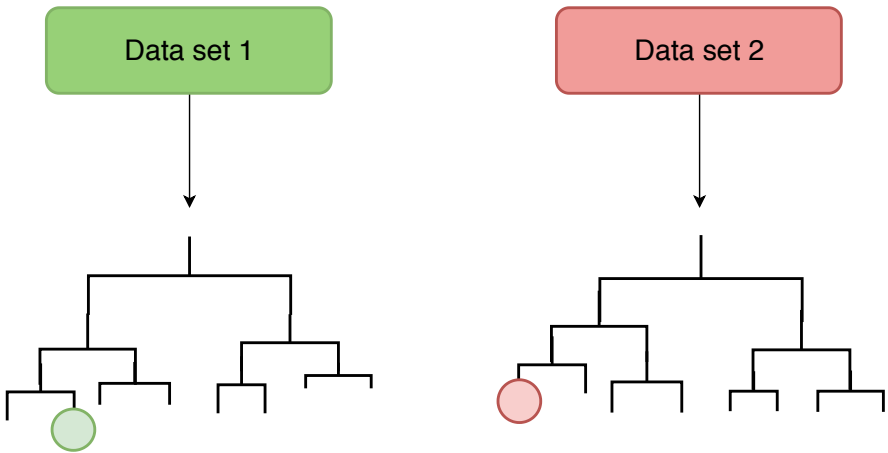
Prominent goal:

- Mechanism of Action (MoA)
- Desirable and undesirable effects

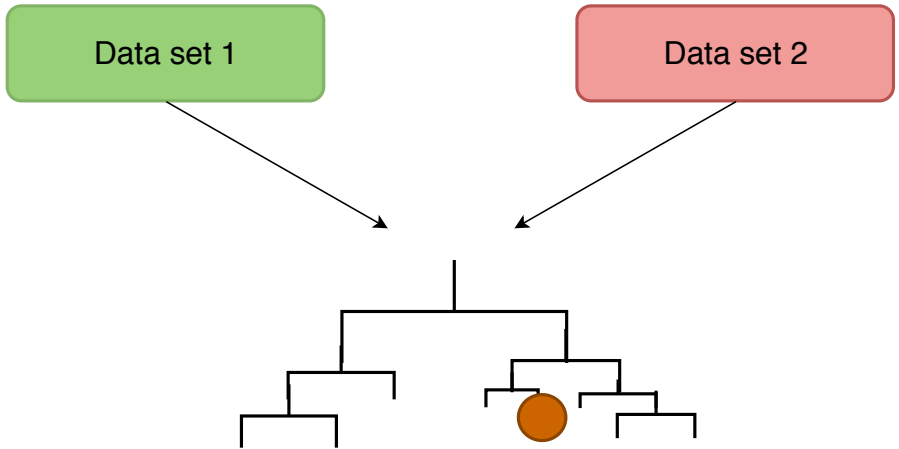
Multi-disciplinary field:

- Chemistry: molecular features
- Proteomics: target interaction
- Genomics: gene expression
- Network system biology: pathway analysis

A single data source is limited in its point of view . . .



. . . extend our knowledge by integrating multiple sources of information.



Multi-source clustering

Multi-source clustering

An ideal integrative clustering technique would discover:

1. similarity consistent across multiple data types
2. similarity specific to an individual data type
3. weak similarity yet consistent across data types emerging only due to the combination of levels of evidence

M-ABC: a multitude of ABC's

Multiple Aggregation of Bundles of Clusters

- For each data set $l = 1, \ldots, d$:
  - For r iterations
    - random subset of samples and features
    - cluster and divide in k clusters
    - incidence matrix $\mathbf{C_l}$

$$c_{ij} = \begin{cases} 0, \text{ if objects } i \text{ and } j \text{ belong to different clusters} \\ 1, \text{ if objects } i \text{ and } j \text{ belong to the same cluster} \end{cases}$$

M–ABC: extending your ABC's

- Sum all incidence matrices $\mathbf{C}_l$
- Divide by the number of times two objects are selected simultaneously
- Perform final clustering procedure

Data sets can be of any type

Any clustering algorithm can be applied

The IntClust R package

Multi-source clustering methods – I

Category	Method	R function	Reference
Direct Clustering	ADC	`ADC()`	Fodeh et al. [2013]
	ADEC	`ADEC()`	Fodeh et al. [2013]
Similarity-based approaches	Weighted	`WeightedClust()`	Perualila-Tan et al. [2016]
	SNF	`SNF()`	Wang et al. [2014]
Graph-based approaches	CSPA	`EnsembleClustering()`	Strehl and Gosh [2002]
	HGPA	`EnsembleClustering()`	Strehl and Gosh [2002]
	MCLA	`EnsembleClustering()`	Strehl and Gosh [2002]
	HBGF	`HBGF()`	Fern and Brodley [2004]
	Balls	`ClusteringAggregation()`	Gionis et al. [2007]
	Aggl.	`ClusteringAggregation()`	Gionis et al. [2007]
	Furthest	`ClusteringAggregation()`	Gionis et al. [2007]

Multi-source clustering methods – II

Category	Method	R function	Reference
Voting-based	CVAA	`CVAA()`	Saeed et al. [2012]
consensus	W-CVAA	`CVAA()`	Saeed et al. [2014]
approaches	IVC	`ConsensusClustering()`	Nguyen and Caruana [2007]
	IPVC	`ConsensusClustering()`	Nguyen and Caruana [2007]
	IPC	`ConsensusClustering()`	Nguyen and Caruana [2007]
	EA	`EvidenceAccumulation()`	Fred and Jain [2002]
	M-ABC	`M_ABC()`	Amaratunga et al. [2008]
	CTS	`LinkBasedClustering()`	Iam-on and Garrett [2010]
	SRS	`LinkBasedClustering()`	Iam-on and Garrett [2010]
	ASRS	`LinkBasedClustering()`	Iam-on and Garrett [2010]
	CEC	`CEC()`	Fodeh et al. [2013]
	WonM	`WonM()`	-
Hierarchy-based	EHC	`EHC()`	Hossain et al. [2012]
approaches	HEC	`HEC()`	Zheng et al. [2014]

Case study I

MCF7 cell line

- 56 compounds
- 250 molecular features
- 477 target predictions

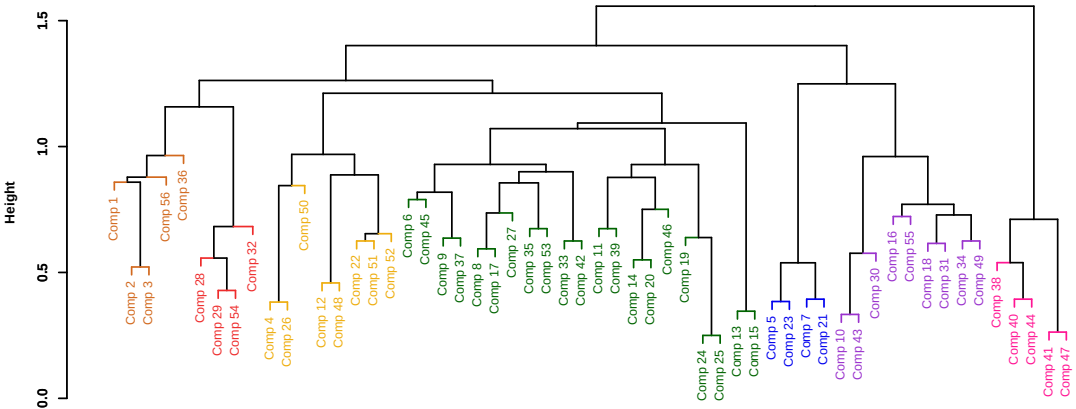
	FP157	FP158	FP159	FP160
metformin	0	0	0	1
phenformin	0	0	0	1
phenyl biguanide	1	0	0	0
estradiol	0	0	0	0
dexamethasone	0	0	0	0
verapamil	0	0	0	0

	TP12	TP13	TP14	TP15
metformin	0	0	0	0
phenformin	0	0	0	0
phenyl biguanide	0	0	0	0
estradiol	0	1	0	0
dexamethasone	0	1	0	0
verapamil	0	0	1	0

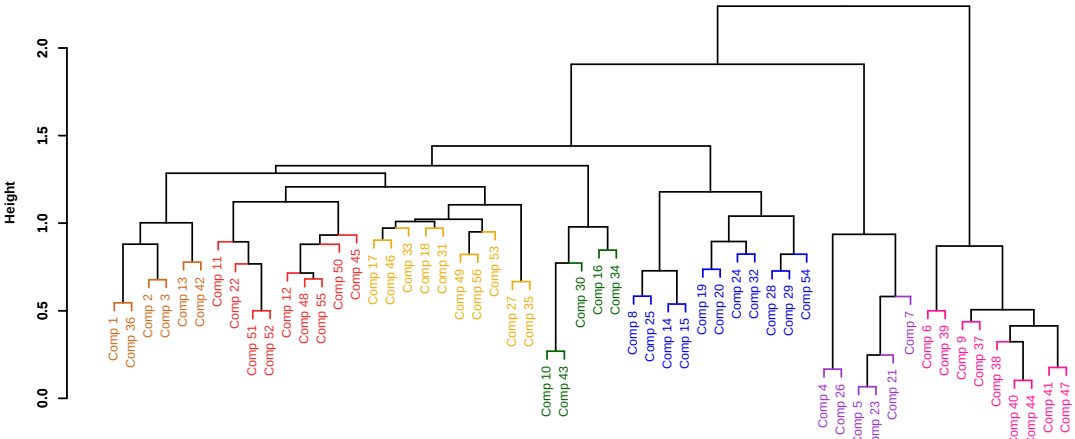
```
MCF7_F <- Cluster(Data=fingerprintMat,type="data",distmeasure="tanimoto")

ClusterPlot(Data1=MCF7_F,nrclusters=7, cols = Colours)
```

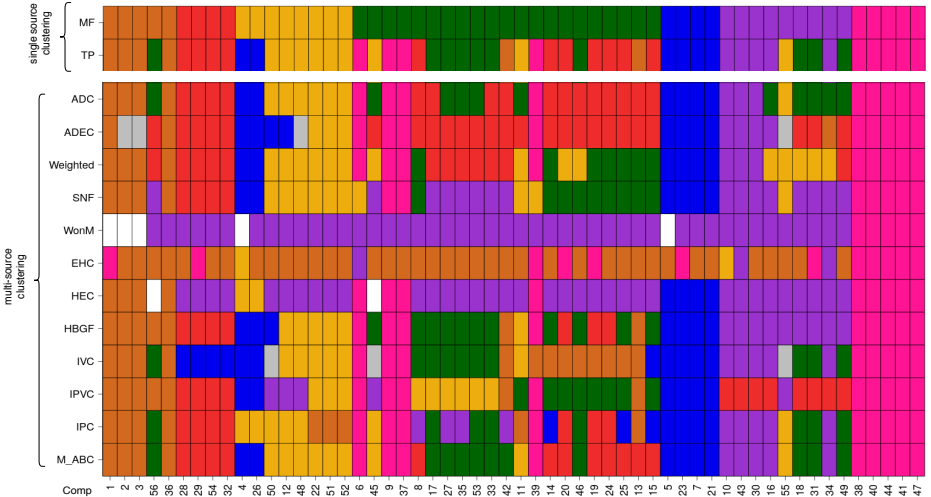
Molecular features



Target predictions

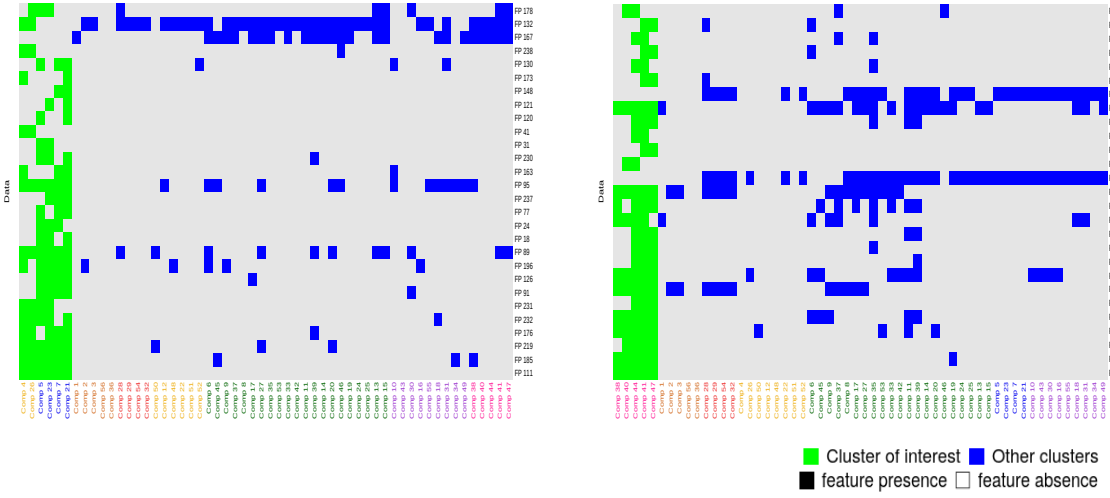


Multi-source clustering



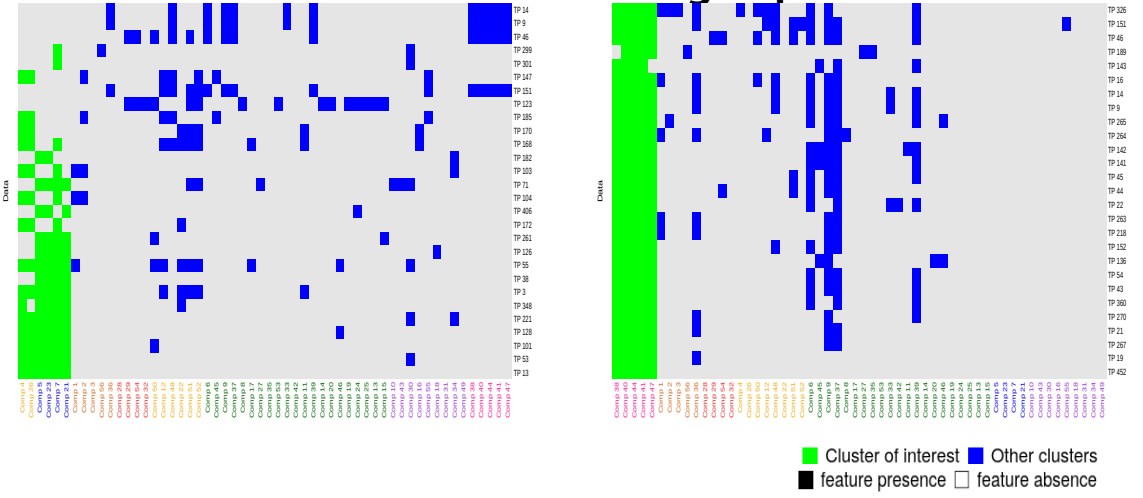
`ComparePlot()`

Characteristic features – molecular features



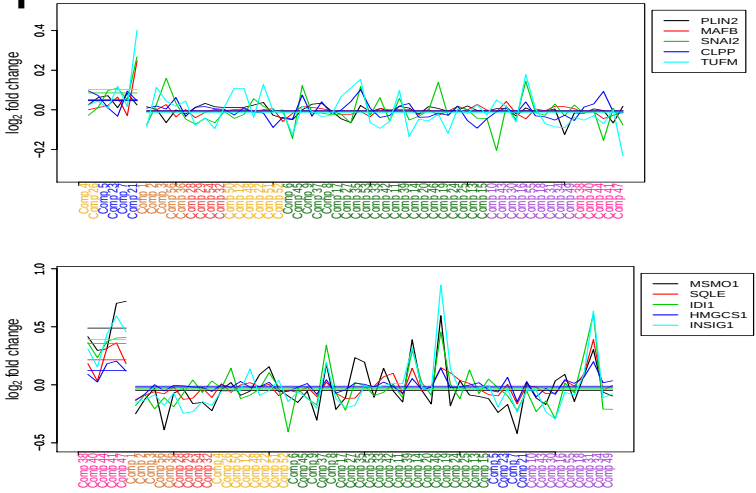

```
CharacteristicFeatures()
BinFeaturesPlot()
```

Characteristic features – target predictions




```
CharacteristicFeatures()
BinFeaturesPlot()
```

Gene expression




```
DiffGenes()
ProfilePlot()
```

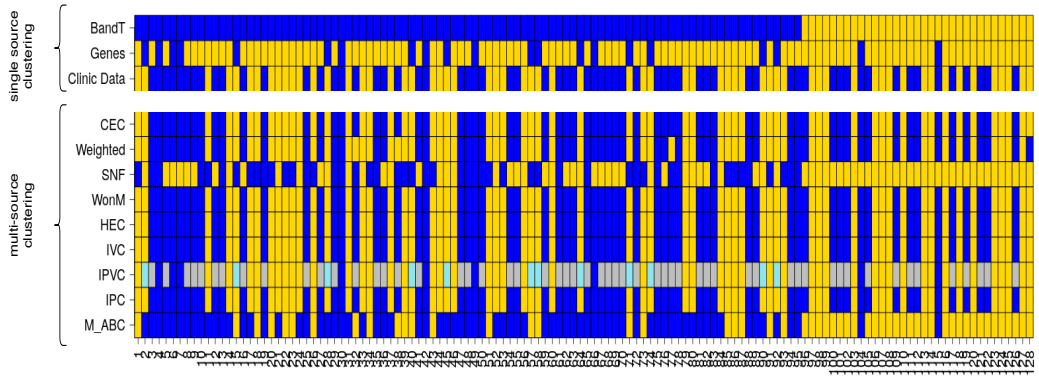
Case study II

The ALL leukemia data set

- 128 leukemia patients
- gene expression of 12,625 genes
- clinical data

Can we retrieve the B-ALL and T-ALL classes in the cluster compositions?

Multi-source clustering results



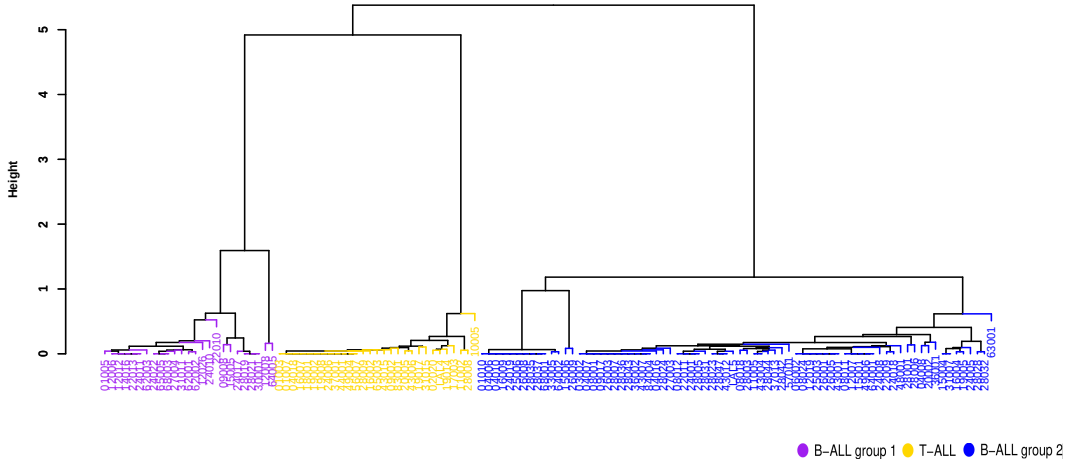

```
ComparePlot()
```

Misclassification error

Method	MCE	Method	MCE
Gene Expression	0.39	HEC	0.42
Clinical Data	0.52	CSPA	0.39
CECa	0.42	HGPA	0.49
CECb	0.39	MCLA	0.42
CECc	0.48	HBGF	0.48
Weighted	0.50	cts	0.39
SNF	0.35	srs	0.39
WonM	0.39	asrs	0.39

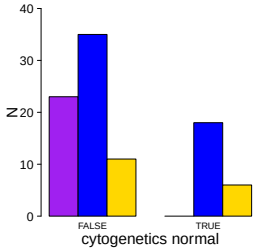
M−ABC : 0.20

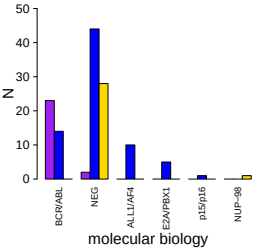
Clustering results

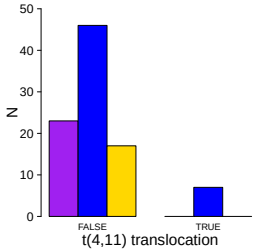


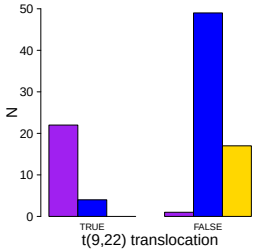

```
M_ABC()
ClusterPlot()
```

Clinical information



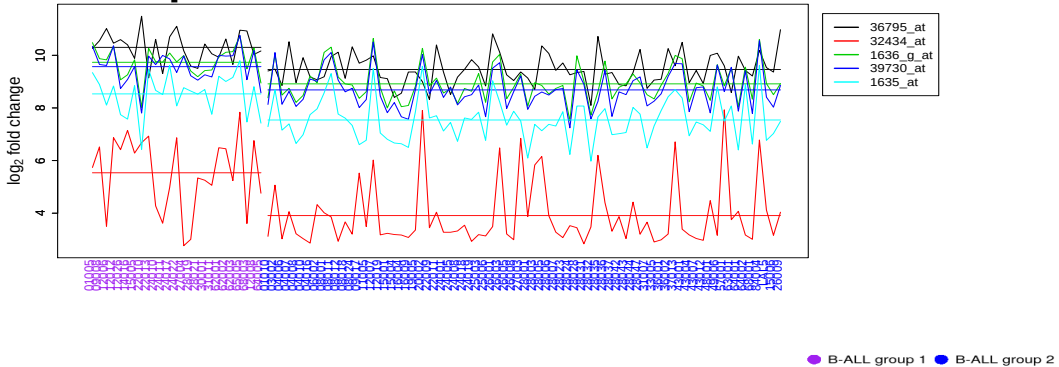






B−ALL group 1 T−ALL B−ALL group 2

Gene expression



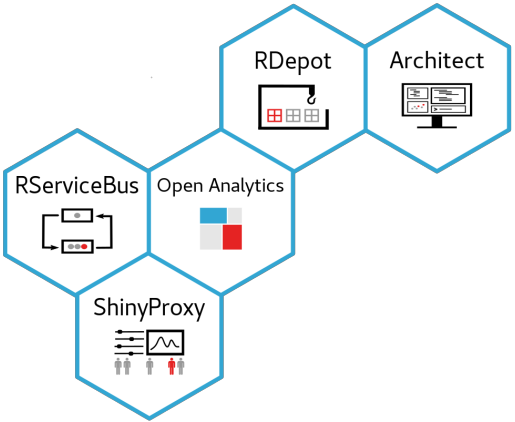

```
DiffGenes()
ProfilePlot()
```

Concluding Remarks

Concluding remarks

- Integration of multiple data sources
  - multiple aspects of underlying biology
  - Save resources by exploring the MoA *in silico*
- M-ABC
  - promising multi-source clustering method
- R package : **IntClust**

Thank you



https://cran.r-project.org/web/packages/IntClust/index.html

marijke.vanmoerbeke@openanalytics.eu

D. Amaratunga, J. Cabrera, and V. Kovtun. Microarray learning with abc. *Biostatistics*, 9:128–136, 2008.

X. Z. Fern and C. E. Brodley. Solving cluster ensemble problems by bipartite graph partitioning. In *Proceedings of the 21th International Conference on Machine Learning*, 2004.

J. Fodeh, C. Brandt, B. T. Luong, A. Haddad, M. Schultz, T. Murphy, and M. Krauthammer. Complementary ensemble clustering of biomedical data. *Journal of Biomedical Informatics*, 46(3):436–443, 2013.

A. L. N. Fred and A. K. Jain. Data clustering using evidence accumulation. *International Conference on Pattern Recognition*, 16:276–280 vol.4, 2002.

A. Gionis, H. Mannila, and P. Tsaparas. Clustering aggregation. *ACM Transactions on Knowledge Discovery from Data*, 1(1):4, 2007.

M. Hossain, S. M. Bridges, Y. Wang, and J. E. Hodges. An effective ensemble method for hierarchical clustering. In *Proceedings of the Fifth International C* Conference on Computer Science and Software Engineering*, pages 18–26, 2012.

N. Iam-on and S. Garrett. Linkclue: A matlab package for link-based cluster ensembles. *Journal of Statistical Software*, 36(9):1–36, 2010.

N. Nguyen and R. Caruana. Consensus clusterings. In *Proceedings - IEEE International Conference on Data Mining, ICDM*, pages 607–612, 2007.

N. Perualila-Tan, Z. Shkedy, W. Talloen, H. W. H. Goehlmann, QSTAR Consortium, M. Van Moerbeke, and A. Kasim. Weighted-similarity based clustering of chemical structure and bioactivity data in early drug discovery. *Journal of Bioinformatics and Computational Biology*, 14(4):1650018, 2016.

F. Saeed, N. Salim, and A. Abdo. Voting-based consensus clustering for combining multiple clustering of chemical structures. *Journal of Cheminformatics*, 4:37, 2012.

F. Saeed, A. Ahmed, and M. S. Shamsir. Weighted voting-based consensus clustering for chemical structure databases. *Jounral of computer-aided molecular design*, 28:675–684, 2014.

A. Strehl and J. Gosh. Cluster ensembles - a knowledge reuse framework for combining multiple partitions. *Journal of Machine Learning Research*, 3: 583–617, 2002.

B. Wang, M. A. Mezlini, F. Demir, M. Fiume, Z. Tu, M. Brudno, B. Haibe-Kains, and A. Goldenberg. Similarity network fusion for aggregrating data types on a genomic scale. *Nature*, 11(3):333–337, 2014.

L. Zheng, T. Li, and C. Ding. A framework for hierarchical ensemble clustering. *ACM Transactions on Knowledge Discovery from Data*, 9(2):9:1–9:23, 2014.