Comparing the performance of biomedical clustering methods

Christian Wiwie[1], Jan Baumbach[1–4] & Richard Röttger[1,4]

Identifying groups of similar objects is a popular first step in biomedical data analysis, but it is error-prone and impossible to perform manually. Many computational methods have been developed to tackle this problem. Here we assessed 13 well-known methods using 24 data sets ranging from gene expression to protein domains. Performance was judged on the basis of 13 common cluster validity indices. We developed a clustering analysis platform, ClustEval (http://clusteval.mpi-inf.mpg.de), to promote streamlined evaluation, comparison and reproducibility of clustering results in the future. This allowed us to objectively evaluate the performance of all tools on all data sets with up to 1,000 different parameter sets each, resulting in a total of more than 4 million calculated cluster validity indices. We observed that there was no universal best performer, but on the basis of this wide-ranging comparison we were able to develop a short guideline for biomedical clustering tasks. ClustEval allows biomedical researchers to pick the appropriate tool for their data type and allows method developers to compare their tool to the state of the art.

In recent years, the amount of biomedical data produced in large-scale experiments has grown rapidly, leading to a need for efficient bioinformatics methods. A typical first step in data analysis is the identification of groups of objects that are likely to be functionally related or interacting. Clustering approaches can be used to identify such groups of similar objects and suggest functional classes and classification schemes. Such approaches are used in computational biology[1,2], information retrieval[3], computer linguistics[4], medical informatics[5] and many other fields. In biomedicine, clustering methods are applied extensively. Typical examples are cancer subtyping on the basis of gene expression levels[6,7], protein homology detection from amino acid sequences or structures[7,8] and the identification of protein complexes using protein-protein interactions[9,10]. Because of the variety of application areas, clustering quality has been defined ambiguously, and different formats for input and output have emerged, which impedes the interchangeability of existing cluster pipelines[11,12]. Many clustering methods have been developed for the optimization of different criteria. Consequently, the output of different tools used to analyze the same data set may vary substantially. Existing tools are classified into partitioning (*k*-means), hierarchical, density-based, model-based and graph-based approaches[13].

Despite the wealth of tools and research, four key problems remain.

1. Tool picking. There is no general best-performing approach, as the choice of method is highly dependent on the data set at hand[13]. The plethora of existing tools and their characteristics might overwhelm researchers who are not experts in data clustering.
2. Parameter optimization. Clustering methods can be adjusted by at least one parameter, which influences the number and size of the resulting clusters. Parameter choice also depends on the data set and question at hand (for instance, whether a set of protein sequences is to be clustered into protein families (many small clusters) or superfamilies (a few big clusters)). Often parameters are guessed at, and researchers then select the one leading to the most 'promising' result[14].
3. Quality measures. The output (i.e., the clusters) are validated using commonly agreed upon cluster quality measures (also called cluster validity indices). Each of the indices rewards and penalizes different criteria such that several cluster validity indices might lead to contradictory results. For instance, the silhouette value for a non-convex data set might not agree well with an external index such as the F1 score. Consequently, a scientist may struggle to select the most appropriate measure for a given real-world question[13].
4. Standardized evaluation. All of the aforementioned steps are usually carried out manually, which renders cluster analyses error-prone, time-consuming and hard to reproduce. A lack of standardized data formats and evaluation protocols leads to poor comparability of clustering results[15]. New methods are usually evaluated on only a few selected data sets. Often they are compared against only some other tools, and usually not on the same data sets.

We developed ClustEval, a publicly available integrative clustering evaluation framework for comprehensive and objective evaluation

[1]Department of Mathematics and Computer Science, University of Southern Denmark, Odense, Denmark. [2]Computational Systems Biology, Max Planck Institute for Informatics, Saarbrücken, Germany. [3]Cluster of Excellence for Multimodal Computing and Interaction, Saarland University, Saarbrücken, Germany. [4]These authors jointly supervised this work. Correspondence should be addressed to J.B. (jbaumbac@mpi-inf.mpg.de).

RECEIVED 12 MARCH; ACCEPTED 24 JULY; PUBLISHED ONLINE 21 SEPTEMBER 2015; DOI:10.1038/NMETH.3583

of the currently available wealth of clustering tools using different data sets of different types and varying parameters and quality measures. We used it to examine different clustering methods with various data sets simultaneously in an automated and standardized way. In particular, ClustEval helped us automatically determine optimal parameters for the different clustering methods.

Here we present an objective comparison of 13 popular bioinformatics clustering methods (**Table 1**) using 24 data sets (**Table 2**), including gene expression levels, protein-sequence similarities and synthetic data. Our results summarize hundreds of thousands of clustering outputs on varying parameters evaluated in comparison with gold standard. We further evaluated the robustness of each method to noise that we artificially introduced to the data sets. We used 13 different validity indices (**Table 3**) that could generally be separated into two different classes: internal and external indices. Internal quality measures judge a clustering on the basis of certain intrinsic statistical properties of the clustering itself, whereas external indices compare the clustering to a user-given gold-standard clustering (the 'ground truth'). Each validity index allows for ranking of the tools for each data set. Furthermore, we used a selected set of five data sets to assess the robustness of the methods. We generated distorted versions of each original data set and analyzed the performance stability of the methods. All ClustEval results are stored in a standard format and can be reproduced and inspected on the ClustEval website. Thus any future results could be seamlessly compared to the existing results.

RESULTS

Performance comparison

Here we discuss the results on the basis of only the F1 score and the silhouette value as representatives for one external and one internal measure, respectively (Online Methods). The results for all indices can be inspected interactively online at the ClustEval website.

No single method performed best in all settings. Nevertheless, some methods were among the top three performers more often than others were. When we assessed performance on the basis of the F1 score, Clusterdp (CDP) (14 times), hierarchical clustering (HC) (13 times), density-based spatial clustering of applications with noise (DBS) (13 times) and transitivity clustering (TC) (11 times) were the top-performing methods (**Fig. 1** and **Supplementary Fig. 1**). When we used the internal silhouette value instead, the top methods were TC (20 times), HC (14 times), partitioning around medoids (PAM) (14 times) and CDP (11 times) (**Fig. 1**).

The intuitive classification of the synthetic data sets into easy, medium and hard (**Table 4** and Online Methods) was reflected well by the external index but not by the internal one. For all methods, the hard data sets achieved average F1 scores of 0.66, whereas medium and easy ones achieved higher values (0.78 and 0.84, respectively). This trend was generally not reflected by the mean silhouette values (easy, 0.33; medium, 0.07; hard, 0.25).

At first glance, the tool rankings suggested by the external and internal measures do not agree, regardless of the type of data set used. In a real-world setting (usually lacking a gold standard), a researcher performing a clustering task on a new data set is generally left with internal validity indices. We therefore investigated the correlation among all internal and external indices (**Fig. 2**). The silhouette value correlated best (i.e., most coefficients were >0.7) with the F1 score, F2 score, FM index, Jaccard index and V-measure (**Fig. 2a**). We then focused on biomedical data sets and aimed to use the silhouette value to retrieve an optimal parameter set for each clustering method. We tested whether the clustering methods achieved (1) high F1 scores in general and (2) high F1 scores for the parameter set that yielded the highest silhouette value. TC, HC, CDP and PAM performed best (median F1 scores of >0.85) on biomedical data sets (**Fig. 2b**). When the parameter sets were deduced using the silhouette value, TC, HC and PAM still showed superior performance (median F1 scores of >0.82), but CDP showed high variance (**Fig. 2c**).

Thus, we carefully suggest the following guideline for researchers working with a new biomedical data set but no gold

Table 1 | The clustering methods included in this study

Abbreviation	Name	Optimized parameter(s)	Software version and package	Classification
AP	Affinity propagation[22,*]	Preference $p \in [\wedge,\vee]$	16/02/2007	G
CDP	Clusterdp[23,*]	Kernel radius $dc \in [\wedge,\vee]$ Number of clusters $k \in [2, n]$	21/08/2014	D
CL1	clusterONE[10,*]	Size threshold $s \in [1, 0.1n]$ Density threshold $d \in [0, 1]$	1.0	G
DBS	DBSCAN[24]	Epsilon neighborhood eps $\in [\wedge,\vee]$ Density region size minPts $\in [1, n]$	R (fpc), 2.1-7	D
F	Fanny[25]	Number of clusters $k \in [2, n]$ Membership exponent me $\in (1.1, 2, 5)$	R (cluster), 1.14.2	K
HC	Hierarchical clustering[26]	Number of clusters $k \in [2, n]$	R (stats), 2.15.1	H
KM	*K*-means[26]	Number of clusters $k \in [2, n-1]$	R (stats), 2.15.1	K
MC	Markov clustering[27]	Inflation $I \in [1.1, 10]$	12-068	G
MO	MCODE[28]	Fluffing $F \in (0, 1)$, haircut $H \in (0, 1)$, similarity threshold $T \in [\wedge,\vee]$, vertex weight percentage $v \in [0, 0.9]$	1.3.2	G
PAM	Partitioning around medoids[25]	Number of clusters $k \in [2, n-1]$	R (cluster), 1.14.2	K
SOM	Self-organizing maps[29]	Grid size $x \in [2, n]$, $y \in [1, n]$	R (kohonen), 2.0.14	M
SC	Spectral clustering[30]	Number of clusters $k \in [2, n]$	R (kernlab), 0.9.14	H
TC	Transitivity clustering[31,*]	Similarity threshold $T \in [\wedge,\vee]$	1.0	G

$\wedge$ and $\vee$ refer respectively to the minimal and maximal value in a given similarity matrix S, which assigns each pair of data objects a similarity value. n denotes the number of objects in a data set. The classification into clustering-strategy categories was extracted from ref. 13. K, *k*-means; H, hierarchical; D, density-based; M, model-based; G, graph-based. Tools marked with an asterisk were published too recently to be considered in ref. 13.

Table 2 | Overview of the data sets

Type	Name	Size ((*n*, *d*) or number of similarities)
Gene expression	bone_marrow[32]	38, 999
	tcga[33]	293
PPI	ppi_mips[34]	1,562
Protein-sequence similarity	astral1_161[35]	507
	astral_40_seqsim_beh[2,31]	1,047
	brown[36]	232
Protein-structure similarity	astral_40_strsim[37]	1,048
Social network	zachary[38]	34
Synthetic	chang_pathbased[39]	300, 2
	chang_spiral[39]	312, 2
	fraenti_s3[40]	5,000, 2
	fu_flame[41]	240, 2
	gionis_aggregation[42]	788, 2
	veenman_r15[43]	600, 2
	zahn_compound[44]	399, 2
	twonorm_50d[45,#]	200, 50
	twonorm_100d[45,#]	200, 100
	synthetic_cassini[45,#]	250, 2
	synthetic_cuboid[45,#]	250, 3
	synthetic_spirals[45,#]	250, 2
Word-sense disambiguation	coli_find[46]	420
	coli_need[46]	105
	coli_state[46]	190
	coli_time[46]	511

n, number of objects; *d*, number of features (dimensionality). The gold standards for the nonsynthetic data sets were derived from the cited studies. The synthetic data had gold standards by design. The synthetic data sets marked with "#" were generated using the R library mlbench; all others were extracted from the cited papers. An interactive visualization of all results for all data sets can be found online (http://clusteval.mpi-inf.mpg.de).

Table 3 | Cluster validity indices

	Name	Best if . . .	Range
Internal	Davies-Bouldin index[47]	Low	[−∞, +∞]
	Dunn index[48]	High	[−∞, +∞]
	Silhouette value[49]	High	[−1, +1]
External	F_β score[50]	High	[0, 1]
	False discovery rate[51]	Low	[0, 1]
	False positive rate[52]	Low	[0, 1]
	Fowlkes-Mallows index[53]	High	[0, 1]
	Jaccard index[54]	High	[0, 1]
	Rand index[55]	High	[0, 1]
	Sensitivity (recall)[52]	High	[0, 1]
	Specificity[52]	High	[0, 1]
	V-measure[56]	High	[0, 1]

Formal definitions are given in the **Supplementary Note**. Note that in the text we discuss the F_β score only for $\beta = 1$ and refer to it as the F1 score.

standard: (1) Use TC, HC or PAM. (2) Compute the silhouette values for clustering results using a broad range of parameter-set variations. (3) Pick the result for the parameter set yielding the highest silhouette value. One can look for a similar data set at the ClustEval website to determine a meaningful starting point. We emphasize that this guideline is based solely on analysis of the biomedical data sets presented in this study. The best approach generally is to rely on external measures calculated using ground-truth annotations for a subset of the data set to be clustered.

Robustness analysis

We distorted a selection of five data sets in two different ways to varying degrees (**Supplementary Fig. 2**). First, we randomly removed objects from the data sets (density reduction). Second, we added random objects to introduce background noise (noise addition).

For two biomedical data sets, we tested noise levels of 5% and 10%. For the synthetic data, we increased the noise levels to 20% and 40%. In general, the data sets are not very susceptible to density reduction. Previous studies suggest that the remaining objects and their similarities maintain the original cluster structure. For example, when clustering protein sequences, using only 20% of the similarity values may allow for a clustering performance with F-measures of at least 0.8 when compared to a clustering without density reduction[16]. For the "bone_marrow" gene expression data set, all methods but two achieved very good F1 scores of >0.8. Most tools were very insensitive to noise; only the F1 scores determined by clusterONE (CL1), HC, *k*-means (KM), PAM and spectral clustering (SC) dropped slightly (15–20%). In contrast, the "astral1_161" data set was rather difficult to cluster, and the methods yielded greatly varying F1 scores between 0.25 and 0.8. Accordingly, even the addition of a small amount of noise (5%) to this data set resulted in a substantial drop in performance across all methods (except for those that performed poorly even on the undistorted data). The synthetic data sets "gionis_aggregation," "chang_pathbased" and "synthetic_cassini," in contrast, had more prominent cluster structures. It would have been reasonable to expect no substantial alterations with this data set. However, we performed

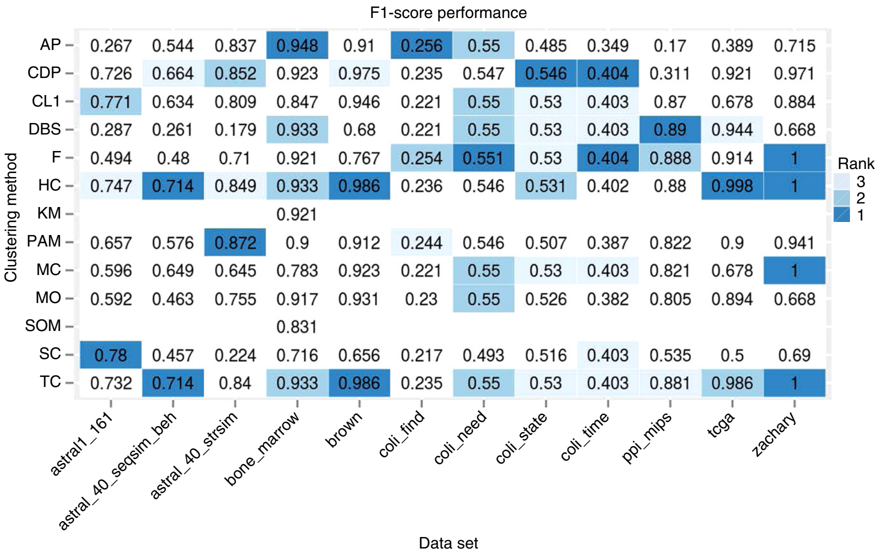

Figure 1 | Performance of all clustering tools on all nonartificial data sets on the basis of F1 scores. The top-performing tools are highlighted for each data set. Empty fields correspond to an inability of the corresponding tool to cluster the data set or an inability to compute a cluster validity index. An extended version of this figure including the artificial data sets and the silhouette values is provided in **Supplementary Figure 1**. Abbreviations for clustering methods are defined in **Table 1**.

Table 4 | Properties and classification of the artificial data sets

	Shape		Separation			
Data set	**Spherical**	**Convex**	**Separated**	**Highly overlapping**	**Nested**	**Difficulty**
veenman_r15	X	X	X			Easy
synthetic_cuboid		X	X			Easy
synthetic_cassini		X	X			Easy
gionis_aggregation		X				Medium
fu_flame						Medium
twonorm_50d	X					Medium
twonorm_100d	X					Medium
synthetic_spirals			X		X	Hard
chang_spiral			X		X	Hard
zahn_compound					X	Hard
chang_pathbased					X	Hard
fraenti_s3	X	X		X		Hard

This classification is based on manual or visual inspection and on inherent properties of the data sets. Easy, convex (or spherical) and well-separated clusters; medium, non-convex or nonseparated clusters; hard, nested or highly overlapping clusters.

the analysis to check the algorithmic stability of the methods (i.e., the likelihood that slight modifications of the input will lead to highly varying results). HC showed a decrease in F-measure by approximately 0.2 with density reduction and with noise addition on synthetic_cassini. The other methods seemed more stable, although affinity propagation (AP), CL1 and Markov clustering (MC) performed at a very low level for all three synthetic data sets.

DISCUSSION

We acknowledge that our conclusions are limited by the number of methods and data sets that we incorporated, but we believe that our real-world data covered many typical bioinformatics clustering tasks and that our synthetic data sets covered the most important factors and properties influencing the tools' output. Our analysis framework ClustEval is free, available online, flexible and extendible, and new methods and data sets can be seamlessly evaluated and compared with existing results in the future. Every step follows a predefined workflow, and we believe that our work is a major step toward a more objective means of comparing clustering methods. Essentially, our framework standardizes data formats and streamlines the most important cluster-analysis steps, from the creation of a similarity matrix to parameter optimization. We believe we have tackled the presented key problems by applying many clustering methods to a range of biomedical data sets, by identifying optimal parameters for all combinations of methods and data sets, by evaluating sets of validity indices and by providing an open platform for standardized evaluation and reproducibility.

We found that even though some methods were among the top performers more often than others were, there was no universally best method that outperformed all others across all settings. We emphasize that our suggested guideline for biomedical data clustering without a gold standard is limited by the indices, data sets and methods that we evaluated here. The correlation we obtained between the silhouette value and the F1 score is reasonable (0.71). Nevertheless, we emphasize that the silhouette value cannot be a general *a priori* replacement for external measures, which take the ground truth into account. When we considered all data sets, the correlation between the silhouette value and the F1 score over all tools and parameter sets was 0.49 (more moderate but still reasonably high). The silhouette value is a particularly poor measure for entangled and highly overlapping data sets. Examples are the correlations on the spiral data sets "chang_spiral" (0.04) and "synthetic_spirals" (0.22). We believe that these data sets were not representative for a classical biomedical analysis. Still, one should be aware of the possible limitations of the silhouette value as a quality measure. Our intentions here were to provide an initial guideline and to proof the value of a standardized, well-structured and comprehensive evaluation.

Only a few methods achieved a perfect clustering result (F1 score of 1). In particular, on synthetic data sets of medium or hard difficulty with touching, overlapping, non-convex or even nested clusters, many methods failed. However, surprisingly, many methods also failed to perfectly cluster very easy data sets, such as "cuboid" (4 out of 13 with F1 scores of <0.7). Density-based methods, headed by CDP, were the best performers across all synthetic data sets (on average).

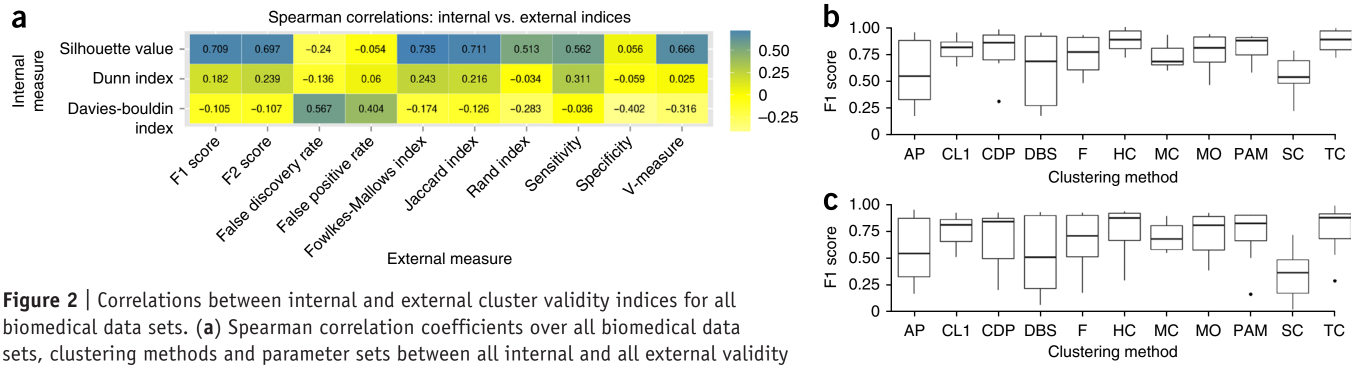

Figure 2 | Correlations between internal and external cluster validity indices for all biomedical data sets. (**a**) Spearman correlation coefficients over all biomedical data sets, clustering methods and parameter sets between all internal and all external validity indices. (**b**) Distribution of the best F1 scores achieved by each method over all biomedical data sets. (**c**) Distribution of F1 scores obtained with the parameter sets that achieved the best silhouette values. Note that the Davies-Bouldin index, false discovery rate and false positive rate are best if minimized and hence should be negatively correlated with the other measures, which are best if maximal. In **b** and **c**, boxes indicate the first and third quartiles, and upper and lower whiskers extend to the highest and lowest values, respectively, within 1.5× the interquartile range. Data beyond the whiskers were considered outliers and are plotted as points. KM and SOM were excluded from **b** and **c** because they work on coordinate-based data sets only. Abbreviations for methods are defined in **Table 1**.

A very brief investigation of the synthetic high-dimensional data sets “twonorm_50d” and “twonorm_100d” indicated that the performance of all methods dropped with increasing dimensionality. Note that this observation should not be generalized without further dedicated, in-depth analyses yielding scientifically sound conclusions.

There are few review papers dedicated to evaluating the performance of different clustering methods on gene expression data sets, usually derived using microarrays (prominent examples are refs. 17–21). One very popular roadmap paper by Andreopolus *et al.*[13] broadly covers biomedical clustering tools. Similarly, Xu *et al.*[12] discuss a list of different methods and validity measures with potential application for biomedical data. Both publications, however, lack systematic and quantitative evaluations, parameter optimization and cross–data set comparisons. In addition to our presentation of extensive analysis of hundreds of thousands of clustering results, we have provided the community with a platform for the standardized evaluation of clustering tools, ClustEval. We believe that this will feed a trend toward more objective and more systematic evaluations of clustering software.

METHODS

Methods and any associated references are available in the online version of the paper.

Note: Any Supplementary Information and Source Data files are available in the online version of the paper.

ACKNOWLEDGMENTS

C.W. is supported by the SDU2020 funding initiative at the University of Southern Denmark. R.R. was partially supported by the International Max Planck Research School on Computer Science and the Saarland University Graduate School for Computer Science. J.B. is grateful for financial support from the Cluster of Excellence for Multimodal Computing and Interaction (MMCI).

AUTHOR CONTRIBUTIONS

C.W. implemented ClustEval and performed the study. J.B. and R.R. jointly directed this work and designed the study. All authors contributed equally to the manuscript.

COMPETING FINANCIAL INTERESTS

The authors declare no competing financial interests.

1. Brohée, S. & van Helden, J. Evaluation of clustering algorithms for protein-protein interaction networks. *BMC Bioinformatics* **7**, 488 (2006).
2. Wittkop, T., Baumbach, J., Lobo, F.P. & Rahmann, S. Large scale clustering of protein sequences with FORCE—a layout based heuristic for weighted cluster editing. *BMC Bioinformatics* **8**, 396 (2007).
3. Salton, G. Developments in automatic text retrieval. *Science* **253**, 974–980 (1991).
4. Navigli, R. Word sense disambiguation: a survey. *ACM Comput. Surv.* **41**, 10:11–10:69 (2009).
5. Verhaak, R.G.W. *et al.* Integrated genomic analysis identifies clinically relevant subtypes of glioblastoma characterized by abnormalities in PDGFRA, IDH1, EGFR, and NF1. *Cancer Cell* **17**, 98–110 (2010).
6. Wirapati, P. *et al.* Meta-analysis of gene expression profiles in breast cancer: toward a unified understanding of breast cancer subtyping and prognosis signatures. *Breast Cancer Res.* **10**, R65 (2008).
7. Wittkop, T. *et al.* Comprehensive cluster analysis with Transitivity Clustering. *Nat. Protoc.* **6**, 285–295 (2011).
8. Röttger, R. *et al.* Density parameter estimation for finding clusters of homologous proteins–tracing actinobacterial pathogenicity lifestyles. *Bioinformatics* **29**, 215–222 (2013).
9. King, A.D., Przulj, N. & Jurisica, I. Protein complex prediction via cost-based clustering. *Bioinformatics* **20**, 3013–3020 (2004).
10. Nepusz, T., Yu, H. & Paccanaro, A. Detecting overlapping protein complexes in protein-protein interaction networks. *Nat. Methods* **9**, 471–472 (2012).
11. Milligan, G. & Cheng, R. Measuring the influence of individual data points in a cluster analysis. *Journal of Classification* **13**, 315–335 (1996).
12. Xu, R. & Wunsch, D.C. Clustering algorithms in biomedical research: a review. *IEEE Rev. Biomed. Eng.* **3**, 120–154 (2010).
13. Andreopoulos, B., An, A., Wang, X. & Schroeder, M. A roadmap of clustering algorithms: finding a match for a biomedical application. *Brief. Bioinform.* **10**, 297–314 (2009).
14. Dubes, R.C. How many clusters are best? - An experiment. *Pattern Recognit.* **20**, 645–663 (1987).
15. Jain, A.K., Murty, M.N. & Flynn, P.J. Data clustering: a review. *ACM Comput. Surv.* **31**, 264–323 (1999).
16. Röttger, R., Kreutzer, C., Duong Vu, T., Wittkop, T. & Baumbach, J. Online transitivity clustering of biological data with missing values. *Proc. German Conference on Bioinformatics* (eds. Böcker, S., Hufsky, F., Scheubert, K., Schleicher, J. & Schuster, S.) 57–68 (Schloss Dagstuhl–Leibniz-Zentrum für Informatik, 2012).
17. Belacel, N., Wang, Q. & Cuperlovic-Culf, M. Clustering methods for microarray gene expression data. *OMICS* **10**, 507–531 (2006).
18. Boutros, P.C. & Okey, A.B. Unsupervised pattern recognition: an introduction to the whys and wherefores of clustering microarray data. *Brief. Bioinform.* **6**, 331–343 (2005).
19. D’Haeseleer, P. How does gene expression clustering work? *Nat. Biotechnol.* **23**, 1499–1501 (2005).
20. Kerr, G., Ruskin, H.J., Crane, M. & Doolan, P. Techniques for clustering gene expression data. *Comput. Biol. Med.* **38**, 283–293 (2008).
21. Thalamuthu, A., Mukhopadhyay, I., Zheng, X. & Tseng, G.C. Evaluation and comparison of gene clustering methods in microarray analysis. *Bioinformatics* **22**, 2405–2412 (2006).
22. Frey, B.J. & Dueck, D. Clustering by passing messages between data points. *Science* **315**, 972–976 (2007).
23. Rodriguez, A. & Laio, A. Clustering by fast search and find of density peaks. *Science* **344**, 1492–1496 (2014).
24. Ester, M., Kriegel, H.-P., Sander, J. & Xu, X. A density-based algorithm for discovering clusters in large spatial databases with noise. *Kdd* **96**, 226–231 (1996).
25. Maechler, M., Rousseeuw, P., Struyf, A., Hubert, M. & Hornik, K. cluster: cluster analysis basics and extensions. R package version 2.0.1 (2015).
26. R Core Team. *A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2012).
27. Van Dongen, S. *A Cluster Algorithm for Graphs* Technical Report INS-R0010 (National Research Institute for Mathematics and Computer Science in the Netherlands, 2000).
28. Bader, G.D. & Hogue, C.W.V. An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics* **4**, 2 (2003).
29. Wehrens, R. & Buydens, L.M.C. Self- and super-organizing maps in R: the kohonen package. *J. Stat. Softw.* **21**, 1–19 (2007).
30. Karatzoglou, A., Smola, A., Hornik, K. & Zeileis, A. kernlab–an S4 package for kernel methods in R. *J. Stat. Softw.* **11**, 1–20 (2004).
31. Wittkop, T. *et al.* Partitioning biological data with transitivity clustering. *Nat. Methods* **7**, 419–420 (2010).
32. Monti, S., Tamayo, P., Mesirov, J. & Golub, T. Consensus clustering—a resampling-based method for class discovery and visualization of gene expression microarray data. *Machine Learning* **52**, 91–118 (2003).
33. Speicher, N. *Towards the Identification of Cancer Subtypes by Integrative Clustering of Molecular Data* M.S. thesis, Universität des Saarlandes (2012).
34. Pagel, P. *et al.* The MIPS mammalian protein-protein interaction database. *Bioinformatics* **21**, 832–834 (2005).
35. Brenner, S.E., Koehl, P. & Levitt, M. The ASTRAL compendium for protein structure and sequence analysis. *Nucleic Acids Res.* **28**, 254–256 (2000).
36. Brown, S.D., Gerlt, J.A., Seffernick, J.L. & Babbitt, P.C. A gold standard set of mechanistically diverse enzyme superfamilies. *Genome Biol.* **7**, R8 (2006).
37. Ortiz, A.R., Strauss, C.E. & Olmea, O. MAMMOTH (matching molecular models obtained from theory): an automated method for model comparison. *Protein Sci.* **11**, 2606–2621 (2002).
38. Zachary, W.W. An information flow model for conflict and fission in small groups. *J. Anthropol. Res.* **33**, 452–473 (1977).

39. Chang, H. & Yeung, D.-Y. Robust path-based spectral clustering. *Pattern Recognit.* **41**, 191–203 (2008).
40. Fränti, P. & Virmajoki, O. Iterative shrinking method for clustering problems. *Pattern Recognit.* **39**, 761–775 (2006).
41. Fu, L. & Medico, E. FLAME, a novel fuzzy clustering method for the analysis of DNA microarray data. *BMC Bioinformatics* **8**, 3 (2007).
42. Gionis, A., Mannila, H. & Tsaparas, P. Clustering aggregation. *ACM Trans. Knowl. Discov. Data* **1**, 4–es (2007).
43. Veenman, C.J., Reinders, M.J.T. & Backer, E. A maximum variance cluster algorithm. *IEEE Trans. Pattern Anal. Mach. Intell.* **24**, 1273–1280 (2002).
44. Zahn, C.T. Graph-theoretical methods for detecting and describing gestalt clusters. *IEEE Trans. Comput.* **C-20**, 68–86 (1971).
45. Leisch, F. & Dimitriadou, E. *mlbench: Machine Learning Benchmark Problems* R package version 2.1-1. (CRAN R Project, 2010).
46. Miller, G.A. WordNet: a lexical database for English. *Commun. ACM* **38**, 39–41 (1995).
47. Davies, D.L. & Bouldin, D.W. A cluster separation measure. *IEEE Trans. Pattern Anal. Mach. Intell.* **1**, 224–227 (1979).
48. Dunn, J.C. Well-separated clusters and optimal fuzzy partitions. *Cybern. Syst.* **4**, 95–104 (1974).
49. Rousseeuw, P.J. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *J. Comput. Appl. Math.* **20**, 53–65 (1987).
50. Powers, D.M.W. Evaluation: from precision, recall and F-factor to ROC, informedness, markedness and correlation. *Journal of Machine Learning Technologies* **2**, 1–24 (2007).
51. Benjamini, Y. & Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Series B Stat. Methodol.* **57**, 289–300 (1995).
52. Hastie, T., Tibshirani, R. & Friedman, J. *The Elements of Statistical Learning* (Springer, 2009).
53. Fowlkes, E.B. & Mallows, C.L. A method for comparing two hierarchical clusterings. *J. Am. Stat. Assoc.* **78**, 553–569 (1983).
54. Jaccard, P. *Etude comparative de la distribution florale dans une portion des Alpes et du Jura* (Corbaz, 1901).
55. Rand, W.M. Objective criteria for the evaluation of clustering methods. *J. Am. Stat. Assoc.* **66**, 846–850 (1971).
56. Rosenberg, A. & Hirschberg, J. V-Measure: a conditional entropy-based external cluster evaluation measure. In *Proc. 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL)* (ed. Eisner, J.) 410–420 (Association for Computational Linguistics, 2007).

ONLINE METHODS

Clustering tools. We selected 13 popular clustering tools commonly used in biomedical contexts (**Table 1**). One of the most popular roadmap papers for clustering biomedical data is a work by Andreopoulos and colleagues published in 2009 (ref. 13). Although it does not offer quantitative evaluations of clustering methods, it classifies them into five categories and provides an overview of the field. From each category (*k*-means, hierarchical, density-based, model-based and graph-based; detailed descriptions are included in the **Supplementary Note**), we picked the most prominent examples (on the basis of citations and publication impact). We also added some more recently developed tools (AP, CDP, CL1 and TC). Some methods have been specifically designed for certain settings. MO and CL1, for instance, are used to identify protein complexes in protein-protein interaction (PPI) networks[10,28]. They essentially identify local, tightly connected neighborhoods (cliques) in a graph and can thus be applied to any kind of similarity graph. To apply MO to non–PPI network data, we created a similarity graph and removed all edges below a defined similarity cutoff. We treated this cutoff as an MO parameter and optimized it with the main MO parameters. AP, KM and PAM detect clusters by identifying the best representatives (prototypes) for groups of similar objects[22,57]. The density-based methods CDP and DBS, in contrast, define clusters as regions of high density[23,58] and do not favor spherical clusters. Hierarchical approaches essentially build a so-called dendrogram in which the proximity of objects reflects the so-called co-cluster likelihood[26]. We used R implementations when they existed for methods (i.e., with DBS, F, HC, KM, PAM, SOM and SC) and standalone implementations otherwise. CDP was available only in Matlab; we re-implemented it in R and provided the source code on the ClustEval website. Introductions for all clustering methods are provided in the **Supplementary Note**.

Data sets. We selected 12 real-world data sets covering 7 biomedical data sets (including gene expression, protein sequences, protein structures and protein complexes), 4 language-processing data sets used for word-sense disambiguation and 1 social network (**Table 2**). In addition, we generated 12 synthetic data sets covering different dimensionality, compactness, separation and shape properties. We categorized them according to their degree of difficulty on the basis of cluster shape and separation. 'Easy' data sets contained convex and separated clusters. 'Medium'-difficulty data sets had non-convex or nonseparated clusters. 'Hard' data sets had nested or highly overlapping clusters (**Table 4**). For the synthetic data, we generated the gold standards together with the data sets. For all real-world data sets, gold standards were extracted from the original publications (**Table 2**).

Data standards. To streamline the evaluation, we designed data format standards (technical documentation is available at the ClustEval website). This ensured future extensibility, as new tools would be required only to read and write these standard input and output formats. All measures, plots, comparisons, etc. will work immediately, with new software required only to support minimal data format standards.

Data preprocessing. The R implementation of SC (kernlab package) requires non-0 rows in the input matrix, so we applied a preprocessor that removes objects containing only 0 values (i.e., objects with no similarity to any other object). Note that these objects are accounted for as singletons during evaluation. Only the brown data set contained such rows. KM and SOM require feature vectors (i.e., coordinates) for the input objects, whereas the others need a similarity matrix as input. For C1 and MC, we normalized the similarities (edge weights) to [0, 1]. We converted coordinates into pairwise similarities by subtracting their Euclidian distance from the maximal observed distance in the data set (except for bone_marrow, where we used the Spearman correlation). Alternatively, ClustEval also supports Pearson correlation and Manhattan distance.

Validity indices. We integrated 13 functions for assessing the performance of a clustering tool on a given data set. This comprised three internal and 10 external measures (**Table 3**). External indices refer to gold-standard data, whereas internal ones are based solely on the input data. Most internal measures reward compactness and separation of clusters; compactness refers to small distances within clusters, and separation to large distances between clusters. Internal indices differ mainly in how they weight and relate these two aspects, and how sensitive they are to outliers. Whereas with the silhouette value and Davies-Bouldin index all objects of two clusters are taken into account to assess their separation and compactness, the Dunn index is based on the 'extreme' objects and thus is very sensitive to outliers. The silhouette value is more sensitive to outliers than the Davies-Bouldin index because it compares only the distances to the closest other clusters[59,60]. The external cluster validity indices, except for the V-measure, are based on fractions of true positives (objects of the same class clustered together), true negatives (objects of different classes clustered separately), false positives (objects of different classes clustered together) and false negatives (objects of the same class clustered separately). The F1 score (also called the F1 measure) is the harmonic mean of precision and recall, and it is the quasi-standard in clustering evaluation with a given gold standard. It has proved useful in many biomedical contexts[60]. The true performance of CL1 might not be fully reflected, as it produces overlapping clusters, which are neglected by most external measures. The **Supplementary Note** contains formal definitions.

Evaluation. We applied each clustering method to each data set using 1,000 different parameter sets. We therefore obtained a large number of clusterings, which we evaluated individually using the 13 cluster validity indices. This resulted in 4,056,000 measures for 312,000 clustering results. We defined the performance of a tool for a given data set using a selected validity index as the best validity achieved over all clusterings on that data set. In the text, we present and discuss the performance rankings of the methods with respect to only the F1 score and the internal silhouette value index; all other combinations are available at the ClustEval website. We follow the argumentation of Handl *et al.*[60] here, by which internal measures are classified as judging compactness, connectedness or separation. Some measures, such as the silhouette value, combine all three types. In the main text, because of space restrictions, we chose to discuss the silhouette value, as it is quite insensitive to noise, which makes it an ideal choice for biomedical data. It also correlates well with the most important external measures (**Fig. 2a**). External measures evaluate a result with regard to the purity of individual clusters and the

doi:10.1038/nmeth.3583

completeness of clusters. The F1 score is a comprehensive measure that takes both of these into account, and it also combines two external measures (precision and recall) and is preferable to simpler techniques[60].

Correlation of internal versus external validity indices. To quantify how well the internal validity indices correlated with the external measures, we computed pairwise Spearman rank correlation coefficients over all methods, data sets and parameter configurations (**Fig. 2a**). We then sought an internal index that correlated best with the external measure of choice (above). The silhouette value correlated best with the F1 score such that we investigated their relationship further, but for each clustering method separately and for biomedical data sets (bone_marrow, tcga, ppi_mips, astral1_161, astral_40_seqsim_beh and brown) only. This excluded KM and SOM, as they require coordinate-based data sets. We further investigated the distribution of the best F1 scores achieved for each method over the six data sets (**Fig. 2b**) and the F1 scores for the parameter set that maximized the silhouette value for each of the data sets (**Fig. 2c**). If multiple parameter sets yielded the same, maximal silhouette values, we computed F1 scores for all of them and used their average. A clustering method was particularly useful for our biomedical sets if (1) it had good performance (high median, low variance) in general (**Fig. 2b**) and (2) we were able to use the internal index to estimate a parameter set that yielded high-performance results (**Fig. 2c**).

Robustness. Biomedical data sets are often incomplete and noisy (because of technical limitations). We simulated the influence of these two factors by (1) randomly removing objects (density reduction) and (2) randomly adding artificial objects (noise addition). For density reduction, we randomly picked a certain percentage of objects (uniformly distributed) and removed them from the data set. For noise addition, we added a set of new objects with randomly generated values distributed uniformly between the minimal and maximal observed value in each dimension (or between minimal and maximal observed values in the similarity matrix). We used parameters optimized for each method with the non-distorted data on the distorted data sets, as we wanted to investigate the degree to which the clustering methods, given optimal parameter sets, are susceptible to small changes in the input. Added and removed objects were ignored when the validity indices were computed. We selected two biomedical (astral1_161 and bone_marrow) and three synthetic (chang_pathbased, synthetic_cassini and gionis_aggregation) data sets of varying levels of difficulty and then analyzed the robustness of all methods to the two types of noise. The level of noise was defined as the percentage of added (or removed) objects. For the two biomedical data sets (astral1_161 and bone_marrow), the noise levels were 5% (low) and 10% (high). For the three synthetic data sets, we assessed the performances with 20% (low) and 40% (high) noise.

The ClustEval platform. As discussed above, clustering is a complex task involving many individual steps that are not independent of each other and which greatly influence the clustering results (key problems 1–3). We streamlined and standardized all these processes and developed the platform ClustEval (key problem 4), which allowed us to exhaustively and comprehensively analyze all incorporated tools and store all results in a database. In addition, it comes with an interactive web interface for browsing all results and for generating plots summarizing comparisons of tools with different data sets and varying parameters using several quality measures. ClustEval is open source and can be downloaded and installed locally. Its back-end is programmed in Java with an interface to R. The web front-end is written in Ruby on Rails. We designed it in an extension-friendly manner such that future clustering tool developers may test their software in numerous settings when supporting the suggested minimal data format standards.

Code availability. All results discussed in the paper and the source code for ClustEval are freely available at our website (http://clusteval.mpi-inf.mpg.de).

57. Hartigan, J.A. & Wong, M.A. A K-means clustering algorithm. *J. R. Stat. Soc. Ser. C Appl. Stat.* **28**, 100–108 (1979).
58. Sander, J., Ester, M., Kriegel, H.-P. & Xu, X. Density-based clustering in spatial databases: the algorithm GDBSCAN and its applications. *Data Min. Knowl. Discov.* **2**, 169–194 (1998).
59. Lawson, R.G. & Jurs, P.C. New index for clustering tendency and its application to chemical problems. *J. Chem. Inf. Comput. Sci.* **30**, 36–41 (1990).
60. Handl, J., Knowles, J. & Kell, D.B. Computational cluster validation in post-genomic data analysis. *Bioinformatics* **21**, 3201–3212 (2005).

doi:10.1038/nmeth.3583