Accounting for noise when clustering biological data

Roman Sloutsky, Nicolas Jimenez, S. Joshua Swamidass and Kristen M. Naegle

Submitted: 22nd May 2012; Received (in revised form): 7th August 2012

Abstract
Clustering is a powerful and commonly used technique that organizes and elucidates the structure of biological data. Clustering data from gene expression, metabolomics and proteomics experiments has proven to be useful at deriving a variety of insights, such as the shared regulation or function of biochemical components within networks. However, experimental measurements of biological processes are subject to substantial noise—stemming from both technical and biological variability—and most clustering algorithms are sensitive to this noise. In this article, we explore several methods of accounting for noise when analyzing biological data sets through clustering. Using a toy data set and two different case studies—gene expression and protein phosphorylation—we demonstrate the sensitivity of clustering algorithms to noise. Several methods of accounting for this noise can be used to establish when clustering results can be trusted. These methods span a range of assumptions about the statistical properties of the noise and can therefore be applied to virtually any biological data source.

Keywords: clustering; noise; measurement variability; random effects; unsupervised learning; cluster ensemble

INTRODUCTION
High-throughput experimental technologies that capture large numbers of molecular measurements, such as gene expression, metabolomics and proteomics technologies, are increasingly common in routine biological research. To understand the data in high-throughput biology, researchers often use clustering algorithms to organize, visualize and infer relationships between objects (e.g. proteins, genes or samples) within a high-dimensional data set.

A variety of clustering algorithms have been employed to analyze biological data, such as Hierarchical clustering and K-means clustering. See Jain et al. [1] for a detailed review. Clustering algorithms partition a data set into clusters where measurements within a cluster are more similar to each other than they are to members of other clusters. The similarity measure is based on a distance metric, such as the Euclidean distance. This improves our ability to visualize complex data by reducing the number of objects into a smaller number of clusters. Doing so helps us understand the underlying process that generated the data. Clustering has been used in a variety of contexts for elucidating a variety of biochemical processes [2–5].

However, biological data are noisy, and clustering algorithms are sensitive to this noise. Noise in experiments arises from the techniques used to make
TOY EXAMPLE

The toy example in Figure 1A could represent a variety of experiments, such as the measurement of 100 messenger RNA transcripts in cancer cells versus normal tissue or of 100 metabolites in untreated versus cells treated with a drug. Using this toy model, we can start to understand how uncertainty in experimental measurements affects our confidence in the clustering solution.

The toy system is made up of five Gaussian processes, with 20 points generated from each process. We will refer to this as the ‘true’ data. The clusters in Figure 1A were generated by K-means clustering with $K = 5$. This solution matches the underlying processes. Black lines indicate cluster boundaries, such that for every point inside a cluster’s boundary area, its Euclidean distance to that cluster’s centroid is smaller than its Euclidean distance to any other centroid. Of course, true empirical measurements would have some noise associated with them, so noise has been superimposed on each of the 100 data points in the toy data set. Shaded areas in Figure 1B indicate 1 standard deviation (SD) of each noise distribution. It now becomes obvious that noisy measurements near dense cluster boundaries, such as those between clusters 1, 2 and 4, could lead to misinterpretations of the relationships between members of these clusters.

Unfortunately, most clustering algorithms do not explicitly account for the underlying uncertainty of measurements.

Our goal is to explore how noise within real data sets impacts the clustering results and interpretation of clustering. We will cover four methods of accounting for noise, which can be combined with any clustering algorithm of choice. These methods span a range of assumptions regarding the independence of measurements and requirements for the number of replicates. The focus of this work is on algorithm-independent methods that can easily be combined with virtually any commonly used clustering algorithm. First, we present a toy example to demonstrate explicitly how noise affects a controlled clustering problem. Next, we will introduce four example methods of accounting for noise. Finally, we will discuss two case studies using real biological data: a phosphoproteomic data set of insulin signaling and a dynamic microarray experiment of epidermal growth factor (EGF)-induced gene expression.

The most common approach to clustering experimental data with replicates is to cluster means of replicate measurements. See Method A in the next section. Important information may be lost when replicates are condensed in this manner before clustering analysis. For example, outlier points may seriously diminish or overemphasize a relationship between objects. Some data points may simply mis-cluster because their measurement average does not accurately reflect the underlying data.

To test typical clustering approaches on our toy data set, we generated two additional replicates for each point in the original data set by randomly drawing from the noise distributions indicated in Figure 1B. For visualization purposes, replicates of only one object from each cluster are plotted in Figure 1C. Although the objects from clusters 3 and 5 had no misclassified replicates, objects from clusters 1, 2 and 4 had at least one replicate that mis-clustered. Figure 1D shows the resulting empirical means and SDs (outlines) compared with the means and SDs of underlying noise distributions (shaded regions). This is a demonstration of how limited replicates may not accurately reflect the true underlying process. Although the average value representations of most of the objects in Figure 1D clustered accurately, the average of the object from Cluster 2 is mis-clustered. Additional experimental replicates improve our confidence regarding measurement accuracy and can also improve our confidence in clustering solutions. Please see Dougherty et al. [8] for an in-depth discussion of the relationship between experimental replication and clustering precision.

We repeated the in silico experiment of generating two additional replicates from the noise models in Figure 1B and clustering the empirical averages. On average, we observed three objects per experiment, out of 100 total, mis-clustered when represented by empirical averages. Mis-clustering occurred for objects in all clusters except Cluster 5. This demonstrates that impact of noise on clustering is influenced by the distribution of data in the multidimensional space.
Measurements for objects in Cluster 5 could effectively tolerate higher amounts of noise without impacting their association with that cluster, whereas measurements falling close to boundaries were often assigned incorrectly in the clustering solution. For real biological data, unlike this toy example, knowing whether correct partitioning has occurred is impossible. The probabilistic nature of the result from this type of noise analysis underscores the facts that hard cluster boundaries may not be meaningful and that measurements with noise, which span multiple cluster boundaries, could be considered to belong partially to multiple clusters.

CLUSTERING STRATEGIES
Some mixture-model-based clustering methodologies have been developed, which solve for clustering solutions while taking noise or replicates into account [9–11]. However, there are clustering methodologies that may work particularly well for a given type of data or that a researcher may be particularly well equipped to implement, for which a model-based incorporation of noise or replicate handling does not exist. Therefore, we are going to focus on methods that can be used in combination with any clustering algorithm and chosen set of clustering parameters.

We consider four methods of handling noise in clustering (Figure 2). In Method A, the data are collapsed by averaging each replicate experiment, and this averaged data are clustered. In Method B, the complete data with all the replicates expanded are clustered, and the concordance of replicate clustering is quantified, see related work in Yeung et al. [12]. Methods C and D are different ways of ‘ensemble’ clustering, which are combinations of clustering new instances of a data matrix, which themselves are likely representations of the data. In Method C, new data matrixes are formed by shuffling the data between replicates. In Method D, new data matrixes are produced by sampling from an
analytic distribution of the data. Examples of handling noise by ensemble clustering can be found in Kerr and Churchill [13] and Bittner et al. [14].

Although most methods of accounting for noise in clustering can be viewed as special cases of these four methods, this section does not constitute a fully comprehensive review of the field. Rather, the focus of these methods is on testing a single clustering algorithm’s sensitivity to noise. We do not address whether a particular clustering solution is in fact optimal for a given data set or desired information outcome. Moreover, we do not discuss previously developed methods that directly modify clustering algorithms, in non-trivial ways, to handle noise. Some of these methods can be found in the following references: Medvedovic et al. [9], Ng et al. [10] and Cooke et al. [11]. In this way, we hope to focus on the methods most broadly applicable across a wide range of biological analysis.

Method A: clustering replicate averages

The majority of studies cluster biological data one time on single vector representations of the data. In the case where multiple replicates exist, it is common to use the average of the replicates to represent the data. This method of average-value clustering will be referred to as Method A, Figure 2A. If there are enough replicates, this is a reasonable way of managing experimental noise, because the average of the replicates converges to the average of the true distribution. In high-throughput biology experiments, however, there are usually a limited number of replicates. With few replicates, as we will see, this is a poor method of managing experimental noise.

Method B: replicate co-clustering

The second method works by clustering all the data, with all the replicates, and measuring the robustness of a result by quantifying if replicates of each object are placed in the same cluster. The advantage of this method is that it is easily implemented and evaluated. A visual demonstration of replicate co-clustering is shown in Figure 1C for a subset of the 100 objects. The results are also summarized in Figure 2B, which indicate the percentage of times that each gene is correctly classified. In practice, the percentage of times each gene pair clusters together would be reported.

Method C: Permutation sampling

Repeat:
• Make $D^*$ (with size $j \times k$) sampled from replicates
• Evaluate the ensemble of cluster solutions

Results for Toy Data: Percentage of correctly clustered replicates
<table>
<thead>
<tr>
<th>Gene 1</th>
<th>Gene 2</th>
<th>Gene 3</th>
<th>Gene 4</th>
<th>Gene 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>56%</td>
<td>46%</td>
<td>100%</td>
<td>66%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Method D: Model-based sampling

Establish a statistical distribution for measurements in D
Repeat:
• Make $D^*$ (with size $j \times k$) by sampling from distribution
• Evaluate the ensemble of cluster solutions

Results for Toy Data: Percentage of correctly clustered replicates
<table>
<thead>
<tr>
<th>Gene 1</th>
<th>Gene 2</th>
<th>Gene 3</th>
<th>Gene 4</th>
<th>Gene 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>48%</td>
<td>44%</td>
<td>100%</td>
<td>58%</td>
<td>100%</td>
</tr>
</tbody>
</table>

Figure 2: Robust clustering methods. We examine four methods of clustering. The results of five genes from the toy data set (Figure 1C) are reported. In this case, we know the true correct cluster of each gene and report the percentage of times that each gene is correctly classified. In practice, the percentage of times each gene pair clusters together would be reported.
measurements. Other methods that directly use replicates have been used. For example, in Yeung et al. [12], they explore forcing replicates into the same subtree as a seed for further Hierarchical clustering.

**Ensemble clustering**

In ensemble clustering, the clustering algorithm is applied to resampled versions of the data to generate multiple clustering solutions. These clustering solutions are combined to create a consensus or ensemble solution [15–19]. The critical reason ensemble clustering is more powerful than replicate co-clustering is that it enables generation of many more samples than there are replicates. This enables better resolution of co-clustering confidence. Furthermore, as we shall see, this method can be applied even to single-replicate data sets.

The ensemble has been used to address a variety of issues that arise in clustering including the effect of initialization on non-deterministic algorithms (such as K-means) [15, 20, 21], sensitivities to algorithm, distance metric and data transformation selections [22] and incorporating the effect of noise [13, 14, 18, 23]. In essence, the ensemble has been used to address how to handle variations in clustering results that arise from the factors that could alter the solution, such as the distance metric used or data variability.

Different groups have used different methods of generating an ensemble solution in the context of noise sensitivities [13, 14]. For example, Bellec et al. [23] use ensemble clustering to find stable features of brain networks in resting-state functional magnetic resonance imaging with sampling from noise distributions of data and allowing the initialization of K-means to change randomly. Unfortunately, studies like this, which account for the effect of experimental noise on a clustering solution, are not published frequently enough.

In this discussion, we focus on a variation of the method used by Bellec et al. [23] because of its ease of implementation. In this method, clustering results are combined to create an ensemble solution by computing a co-occurrence matrix, which indicates the fraction of times each pair of objects cluster together across all clustering sets. Objects that co-cluster frequently are said to ‘robustly’ cluster and are the connections that can be most trusted.

Several groups are exploring methods of evaluating the co-occurrence matrix generated from the cluster ensemble. These methods range from linkage-based clustering of the co-occurrence matrix to define a final ensemble clustering solution [15, 18, 21, 24, 25] to graph theoretical-based methods, where the co-occurrence matrix is viewed as a weighted association matrix between objects [26].

An advantage of ensemble clustering is that the final ensemble result can take on shapes that are different from the constraints of the underlying clustering algorithm used [20]. For example, as we will see in the case studies, the network visualization of the co-occurrence matrix can have a different number of clusters than the K-means clustering algorithm from which it was sampled. Additionally, an important piece of information contained in the ensemble is not just the decrease in probability of one object clustering with a second object, but in balance, what other clusters that object could alternately be associated with.

Ensemble clustering naturally lends itself to a probabilistic, fuzzy clustering interpretation. When the same clusters are consistently identified across clustering sets, the probability of an object belonging to a cluster is simply estimated by the frequency of this occurring in the ensemble, thereby defining fuzzy cluster boundaries. For many real data sets, however, cluster identities cannot be mapped between clustering sets because the identified clusters are so different between clustering solutions. Nevertheless, co-clustering frequencies between pairs of objects may be treated as probabilities of belonging to the same cluster. Robust clusters may then be built up from the most robust pairwise relationships.

We will introduce two methods of producing an ensemble result that accounts for noise. In the first method, the replicates themselves will be reshuffled to produce a new data set. In the second method, an alternate data set will be created by sampling from a noise model. The case studies have been chosen to explore various nuances of these methods, including how to deal with single-replicate data.

**Method C: permutation sampling**

In permutation sampling, which could be considered a form of bootstrapping, the data vectors for each object’s replicates are shuffled to generate each sample (Figure 2C). This is done by randomly picking a value of each measurement to form a novel replicate. Exhaustively, enumerating all permutations of the data set’s replicates is computationally expensive for realistically sized data sets.
A permutation sampling method is attractive because it is easy to implement and makes no explicit assumptions regarding the underlying distribution of the data samples. It is not entirely assumption free. It implicitly assumes, for example, that sample to sample noise for a given object is independent, an assumption sometimes violated if data are not effectively normalized. Also this method assumes that there are sufficient replicates to adequately sample each object. The problem with this requirement is subtle. There are a large number of objects in typical biological data sets, so with a handful replicates it is almost certain that a few of these objects will have non-representative samples [27].

**Method D: model-based sampling**

In the final and most powerful method, samples are drawn from a mathematical model of the data-generating process that explicitly defines the experimental noise. This model is computed from the observed data and appropriate, domain-specific assumptions about the noise.

Sampling consists of generating random values from probability distributions characterizing the replicate data for each point in each vector to create a resampled data set, \( D^* \). Normal distributions are used when normality conditions apply to the data. These distributions can be defined by the means and SDs of replicates or may be inferred from the data by more sophisticated methods [27]. However, experimental data may deviate from normality, and other parametric models of noise can be used for sampling. The advantage of this method is that if a noise model can be defined for a certain process, even data sets with no replicates can be evaluated for the effect of noise on the analysis by clustering. The potential limitation of this application is when assumed distributions deviate drastically from the sampled distribution leading to a skewed or over- or under-representation of the noise.

**MODELING NOISE**

To use Method D, some effort must be made toward defining an appropriate model of the experimental uncertainty in the data. When there are two or more replicates, it is commonly assumed that the noise is normally distributed and that each object’s noise is independent. The sample mean and SD of each object are used to parameterize normal distributions from which sampling replicates are drawn [28]. For this to be sensible, the data must be transformed onto a scale where the data are approximately normally distributed. For example, expression data are usually normalized and transformed onto a log scale to accomplish this.

With domain- and technology-specific studies, it is possible to use better models. In the case of gene expression, several researchers have proposed better noise models [27, 29, 30]. For example, Posekany et al. [29] argues that microarray noise has a fatter tail than a normal distribution and suggest using a t-distribution instead.

The sliding-window prior, proposed by Baldi and Long [27], merits special attention because it is likely to be applicable across several types of biological data. They note that there is a strong, non-linear relationship between the mean and SD of genes in expression data (Figure 3A). Furthermore, when there are few replicates, the most problematic error in modeling is underestimating sample variance, which implies there is more confidence in the true mean than is warranted. To mitigate the risk of dramatically underestimating the variance, they propose averaging the variance of genes with similar mean expression levels (Figure 3B). There is often a strong, but non-linear, dependence between the mean and variance in real biological data, so this regularization is often sensible.

Most methods of modeling noise require that there is at least one replicate of the experimental data. Replicate data sets are not always collected. It is still possible to model noise in this context if the technology is well understood. An example of this is demonstrated in the second case study.

**CASE STUDIES**

To illustrate how these methods work in practice, we will present two case studies, one using phosphoproteomic profiling data and one using gene expression data.

**Case 1: phosphoproteomic data with replicates**

This case study focuses on a quantitative liquid chromatography MS/MS phosphoproteomic experiment, which captures phosphotyrosine signaling dynamics in 3T3-L1 adipocytes stimulated with insulin [31]. The data set was downloaded from PTMScout [32] for analysis, and it represents 120 phosphopeptides measured at 0, 5, 10 and 30 min after stimulation with insulin. Although biological
triplicates were measured, because of technological
limitations, 15% of the phosphopeptides have no
replicate information and 29% of the phosphopep-
tides are only measured in duplicate. Because this
data set contains replicates, we applied all four meth-
ods to this data set to compare the results of each, for
a particular data set and clustering implementation.
In this section, we will present results relative to
Method A results, i.e. clustering using the average
of the replicates.

Method A: clustering replicate averages
Hierarchical clustering, with a Euclidean distance
metric and average linkage, was chosen as the
method for clustering. The dendrogram was cut,
such that 12 clusters were formed (i.e. K = 12).
This set of clustering parameters was chosen based
on relatively good performance for producing
clusters enriched for biological terms, such as Gene
Ontology labels (determined by using the
PTMScout interface [32]). The heat map of the
co-occurrence matrix for average-value clustering is
shown in Figure 4A, a matrix of ones and zeros. The
size of each cluster is shown in a bubble diagram in
Figure 5. The clustering set solution for this imple-
mentation is composed of two large clusters, four
smaller clusters and outlier clusters composed of, at
most, two members. The two largest clusters can be
seen in the upper-right and lower-left portions of the
co-occurrence matrix, Figure 4A.

Method B: replicate co-clustering
Because 85% of the data set has at least one replicate,
we applied Method B and clustered all replicates
together to see how co-clustering was affected,
Figure 4B. Using the same order of phosphopeptides
as shown in Figure 4A, this heat map clearly shows
the two largest clusters have replicates which cluster
between both groups, which is indicated by the
appearance of clustering between the upper-left and
lower-right clusters, Figure 4B. Additionally,
the next two largest clusters also have replicates
co-clustering between the two groups. The two
smallest clusters, which are separated by the clusters
made of single members, do not appear to change as
dramatically in structure when replicates are con-
sidered. For those phosphopeptides with at least
one replicate, 27.5% of them have replicates that
cluster differently. Even this simplistic attempt at
considering noise within an experiment has informed
our understanding about the relationship among
clusters beyond that of average-value clustering. In
particular, it indicates that replicates between the
largest groups co-cluster when considered as individ-
ual vectors within clustering.

Ensemble results: methods C and D
Ensemble clustering by permutation and model-
based sampling were accumulated across 5000 iter-
ations. Figure 4C shows the co-occurrence matrix
that results from sampling a normal distribution
model of noise, defined by a measurement’s experi-
mental mean and SD. For measurements with no
replicates, the experimental mean was assumed to
be the single replicate data, and the SD was derived
from a global estimate of the ratio of sample SD to
sample mean, known as the coefficient of variation
(CV). Global CV was estimated by averaging CVs of
measurements with replicates. This simple approach
assumes a linear relationship between variance and
mean across the data set, but a more complicated

Figure 3: The relationship between mean and variance. (A) In gene expression data, there is a strong relationship
between the mean expression and the sample variance. The higher the expression level, the higher the variance.
(B) In log space, the relationship flips. The higher the expression level, the lower the variance. The sliding-window
prior averages the variance of genes with similar expression levels. Variance estimates are shown as a line.
As mentioned earlier, there are many methods for finding a single clustering solution from an ensemble. Here, we used the Ward algorithm to hierarchically cluster the co-occurrence matrix and chose a cutoff to assemble eight clusters [15], which appeared to be the naturally occurring breakdown of the co-occurrence matrix for both Methods C and D (Figure 4). Figure 5 illustrates the results of the sampling methods, both compared with the cluster structure of Method A results. The basic structures match well to the information contained in the heat maps of the co-occurrence matrixes (Figure 4),

model could be fit to that relationship if necessary. SDs for measurements without replicates were calculated from their means and the global CV estimate. The co-occurrence matrixes were only subtly different between the two methods of sampling. The approximate behavior of the ensembles is the same as co-clustering replicate data; those cluster boundaries that break down when replicates are considered are also blurred when ensemble clustering is used. In this method, probabilities of pairwise relationships can take on a finer range of values versus simply clustering replicate data directly.
i.e. the two largest clusters are remixed to varying extents in the ensemble result, and there are two relatively stable smaller clusters (clusters 3 and 5 in the middle panel). It becomes clear from this illustration that one of the strengths of ensemble clustering is the ability to naturally capture outliers (cluster number 4). These phosphopeptides are outliers based on two pieces of evidence, they do not cluster robustly with any other phosphopeptide and they undergo drastically different dynamics compared with the rest of the data set (Figure 5). Interestingly, despite being outliers, when joined by ensemble
analysis, they form a potentially biologically meaningful subset of phosphorylation sites belonging to proteins involved in regulation of vesicle fusion.

Mapk1 T183/Y185 (Mapk1-p2), Mapk1 Y185 (Mapk1-p1) and Irs1 Y935 are particularly interesting examples to explore in more depth because their cluster membership changes with consideration of noise. Mapk1-p2 and Mapk1-p1 are members of a relatively stable cluster enriched for members of the MAPK cascade; however, permutation sampling indicates that the singly phosphorylated form of Mapk1 instead resides in the cluster containing Irs1 Y935. The disparate results among the methods can be more clearly understood by observing the data used in clustering for each method (Figure 5). The middle panels of Figure 5 demonstrate that Mapk1-p2 and Irs1 Y935 are relatively low variance measurements, with distinctly different down-regulation profiles. In contrast, Mapk1-p1 replicates have much higher variance, and that variance acts in such a way as to make its association with either Irs1 or Mapk1-p2 indeterminable. Both Mapk1 and Mapk3 singly and doubly phosphorylated forms behave identically, indicating that this could in fact represent biologically meaningful information. In cases such as these, it is perhaps best to view these as fuzzy clustering relationships. In this way, one would describe the singly phosphorylated forms of Mapk1 and Mapk3 as belonging partially to the cluster containing Irs1 Y935 and partially as belonging to the doubly phosphorylated forms of the Map kinases.

Because this data set represents a real-world example with an unknown ‘ideal’ clustering solution, it is impossible to say which method of handling noise is best. However, from this study, it is possible to see that average-value clustering is the least informative when it comes to understanding the robustness of a given solution with respect to experimental noise. The remaining three methods, clustering replicates and the ensemble methods, follow surprisingly similar trends when it comes to highlighting relationships that are not robust to noise. The advantage of the ensemble methods appears to be their ability to define a finer range of co-clustering values, which could be helpful in defining either an improved, definitive clustering solution, or a fuzzy clustering solution.

Case 2: gene expression data without replicates

In this case, we examine a single measurement microarray gene expression experiment. HeLa cells were stimulated with EGF for 0, 20, 40, 60, 120, 240 and 480 min, followed by gene expression profiling by hybridization to the Affymetrix HG-U133A array [33]. The data set is publicly available from www.ncbi.nlm.nih.gov/geo/, record GSE6783.

The subject of this study was transcriptional response, and the authors focused their analysis on putative regulators of transcription induced by EGF. To model the intent of the original study, we performed ensemble clustering on a subset of 655 probe sets meeting the following criteria: (i) both ‘DNA’ and ‘transcription’ appear in Gene Ontology annotations of the quantified transcripts and (ii) at least a 2-fold increase in expression was observed at any time point between 20 and 240 min, relative to the basal condition.

In contrast with Case 1, no replicate data were collected, a common scenario in high-throughput biological experiments. The absence of replicate data restricts our methodological options.

Methods A, B and C

In the absence of replicate measurements, Methods A, B and C cannot be used to account for noise. Unlike Case 1, where averaging replicate measurements constitutes an accounting of noise, albeit a naive one, clustering single measurements completely ignores it. In this case, Method D, employing sampling from probability distributions, must be used to account for noise.

Sampling without replicate data

Single measurements can be used to estimate positions (first moments, expectations or means) of probability distributions from which the measurements were drawn. However, they provide no information about the shapes (second moments or variances) of those distributions. To estimate the variances we used, as ‘background’, other data were collected on the same microarray platform to create a mapping between means and variances, based on the assumption that there is a strong and non-linear relationship. Microarray experiments assay gene expression globally, but the majority of genes will not change significantly in any single experiment. Therefore, we assume that we can use data sets collected from the same cell line and on the same microarray platform to generate a model for noise in the absence of replicates in a particular experiment.

We selected two expression data sets, which were also collected in HeLa cells and on the Affymetrix
HG-U133A platform, accessible from www.ncbi.nlm.nih.gov/geo/ as records GSE1417 [34] and GSE7009 [35]. The first experiment was measured in triplicate and the second in quadruplicate. Because measurements are not comparable between data sets without adequate scaling (Figure 6, left panel), background data sets were rescaled by the factor (75th quantile expression in foreground/75th quantile expression in background) to map them to the expression range of the ‘foreground’ (GSE6783) experiment, while preserving all pairwise fold differences between probe sets within each background data set. Rescaling makes all three data sets comparable across the entire range of expression (Figure 6, middle panel).

For the purpose of this demonstration, we modeled the noise of log-transformed expression data with normal distributions [27, 36]. From rescaled background data, we generated a mapping from mean log 2(expression) to SD (Figure 6, right panel) using Cyber-T software [27], which employs the sliding window prior to calculate regularized SD estimates for each set of replicates under normality assumptions. This mapping was interpolated to select SD values for each foreground experimental measurement.

Normal distributions may not adequately model noise for some microarray expression data [37], and heavier-tailed t-distributions have been proposed as a suitable alternative [29, 38]. Background data may be similarly used to parametrize a t-distribution noise model, for example using the algorithm developed by Posekany et al. [29], which may then be interpolated to obtain distribution parameters for foreground data.

**Method D**

Replicate measurements were generated for the 655 selected probe sets by sampling from the normal distributions parametrized as described earlier. Those samplings were then clustered by K-means with cosine distance metric and K = 20. Here, we present results from 500 iterations of sampling, followed by clustering, because co-clustering frequencies >0.5 did not change appreciably with larger number of replications. K = 20 was selected because it roughly fits the square root of number of objects, a general rule of thumb for selection of K, and it seemed to produce relatively well-formed clusters.

Only 11% of probes met the threshold requirement of co-clustering with any other probe 50% of the time or more. A robustness cutoff of >0.65 was chosen for determining robust clusters, because it maximized the number and size of distinct clusters, shown in Figure 7. More stringent choices for the threshold significantly dissipated the formation of robust clusters and lowering the stringency resulted in clusters that were too large to interpret. At this cutoff, Clusters 1 and 2 are still clearly distinct (Figure 7, bottom panel), although AKAP17A may be thought of as partially belonging to both. As discussed earlier, robustness analysis by noise sampling ensembles naturally produces fuzzy clusters with probabilistic boundaries.

Noise, however, is not the sole determinant of clustering robustness, as previously observed with Cluster 5 of the toy example. In ensemble, clustering objects belonging to highly separable clusters co-cluster more robustly than objects with comparable...
noise but belonging to more densely packed clusters. Figure 8 shows trajectories of three genes, which cluster with varying degrees of robustness. Despite having comparable amounts of noise, JUNB and AKAP17A do not cluster with the same degree of robustness. JUNB likely clusters substantially more robustly than AKAP17A due to separability properties of their respective trajectories. Although SD of replicate measurements adequately estimates noise, distinguishing clustering properties of JUNB and AKAP17A requires robustness analysis of clustering.

**CONCLUSIONS**

Unfortunately, experimental measurements are associated with noise, which reduces our confidence in those values. Handling this uncertainty in the process of analyzing large biological data sets by clustering may greatly aid in highlighting those cluster associations, which in turn have low or high confidence in light of this noise. Although the incorporation of noise in clustering requires new layers of analysis, compared with not handling noise, these results will ideally help researchers avoid misinterpreting clustering results and allow them to focus on highly probable hypotheses for further study.

We focused this work on those methods that can be used in an algorithm-independent fashion. By analyzing an in silico toy data set and two real biological data sets with very different structures and noise values, we were able to explore how one can incorporate noise in clustering analysis. One of the recurrent lessons across all these data sets is that the amount of noise alone in a particular measurement does not determine its sensitivity to mis-clustering. The highest sensitivity to noise lies in those regions of high spatial density. Conversely, well-partitioned vectors are much less sensitive, indicating that noise

---

**Figure 7:** Expression data clustering results. Top two rows: Average representation of dynamics in each robust cluster at $K = 20$, co-clustering frequency >0.65. Emphasized black lines in Clusters 1 and 2 represent dynamics of AKAP17A, which co-clusters robustly with at least one member of both clusters. Bottom panel: Graph representation of Clusters 1 and 2. Nodes represent genes. Edges represent co-clustering relationships above the 0.65 robustness cutoff. KLF6_p1 and KLF6_p2 represent separate probe sets hybridizing to KLF6 and similarly for JUN. Node outlined in black represents AKAP17A, belonging partially to both clusters.
analysis at the very least can highlight uncertainties in clustering partitions. Additionally, this observation indicates that pre-filtering a data set to remove observations with a large degree of variation could remove data from consideration that might be separated well by clustering, despite noise.

The most appropriate method of those presented here is entirely dependent on the data set being evaluated. The advantage to Method D is that as long as a model for the data and its noise can be assumed, it can be used for single-, low- or missing-replicate data as a way to test the sensitivity of a solution to unobserved noise. However, erroneous noise models could result in over- or under-representation of sensitivity of the clustering solution for missing- or no-replicate data. Replicate co-clustering and replicate sampling are ways to avoid making assumptions about a particular model of noise but come at the cost of co-clustering resolution. Because the attainment of these multidimensional data sets typically comes with a large financial or resource burden, it is rare that average-value clustering will be sufficient to handle noise during analysis, given low numbers of replicates. Despite the differences in assumptions and requirements for each method covered here, in toy and phosphoproteomic data, they had fairly rough agreement and added a great deal of value to the understanding of the stability of a clustering solution, so perhaps the use of any method, despite potential flaws, is still an improvement over not accounting for noise at all in clustering results.

**Key Points**

- Variations in data due to experimental noise will impact the clustering solution, and therefore, effort should be spent ensuring conclusions drawn from clustered data are robust to experimental noise.
- Clustering replicates and clustering with ensemble techniques are useful at providing an indication of the robustness, or probability, of a given clustering relationship in the face of experimental noise.
- Measurements with a large degree of variance could cluster robustly depending on the separability of the data set.
- Ensemble techniques, in combination with appropriate models of experimental noise, can be used to evaluate the impact of noise on clustering even when no replicates exist.

**References**


