Bioinformatics, 33(17), 2017, 2691–2698
doi: 10.1093/bioinformatics/btx167
Advance Access Publication Date: 24 March 2017
Original Paper

OXFORD

Gene expression

Entropy-based consensus clustering for patient stratification

Hongfu Liu[1], Rui Zhao[2,3], Hongsheng Fang[2,4], Feixiong Cheng[5,6], Yun Fu[1,7,*] and Yang-Yu Liu[2,6,*]

[1]Department of Electrical and Computer Engineering, Northeastern University, Boston, MA 02115, USA, [2]Channing Division of Network Medicine, Brigham and Women's Hospital, Harvard Medical School, Boston, MA 02115, USA, [3]School of Engineering and Applied Sciences, Harvard University, Cambridge, MA 02138, USA, [4]Department of Statistics, Stanford University, Stanford, CA 94305, USA, [5]Center for Complex Network Research and Department of Physics, Northeastern University, Boston, MA 02115, USA, [6]Center for Cancer Systems Biology, Dana-Farber Cancer Institute, Boston, MA 02115, USA and [7]College of Computer and Information Science, Northeastern University, Boston, MA 02115, USA

*To whom correspondence should be addressed.
Associate Editor: Ziv Bar-Joseph

Received on December 2, 2016; revised on February 24, 2017; editorial decision on March 17, 2017; accepted on March 22, 2017

Abstract

Motivation: Patient stratification or disease subtyping is crucial for precision medicine and personalized treatment of complex diseases. The increasing availability of high-throughput molecular data provides a great opportunity for patient stratification. Many clustering methods have been employed to tackle this problem in a purely data-driven manner. Yet, existing methods leveraging high-throughput molecular data often suffers from various limitations, e.g. noise, data heterogeneity, high dimensionality or poor interpretability.

Results: Here we introduced an Entropy-based Consensus Clustering (ECC) method that overcomes those limitations all together. Our ECC method employs an entropy-based utility function to fuse many basic partitions to a consensus one that agrees with the basic ones as much as possible. Maximizing the utility function in ECC has a much more meaningful interpretation than any other consensus clustering methods. Moreover, we exactly map the complex utility maximization problem to the classic *K*-means clustering problem, which can then be efficiently solved with linear time and space complexity. Our ECC method can also naturally integrate multiple molecular data types measured from the same set of subjects, and easily handle missing values without any imputation. We applied ECC to 110 synthetic and 48 real datasets, including 35 cancer gene expression benchmark datasets and 13 cancer types with four molecular data types from The Cancer Genome Atlas. We found that ECC shows superior performance against existing clustering methods. Our results clearly demonstrate the power of ECC in clinically relevant patient stratification.

Availability and implementation: The Matlab package is available at http://scholar.harvard.edu/yyl/ecc.

Contact: yunfu@ece.neu.edu or yyl@channing.harvard.edu

Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

High-throughput technologies, such as next-generation sequencing, have enabled us to rapidly accumulate a wealth of various molecular data types, from genome to transcriptome, proteome and epigenome (Uhlen *et al.*, 2016; Zhu *et al.*, 2015). Those massive genomics studies offer us great opportunities to characterize human pathologies and disease subtypes, identify driver genes and pathways and nominate drug targets for precision medicine (Biankin *et al.*, 2015; Bolouri *et al.*, 2016; Lapointe *et al.*, 2004; Kamburov *et al.*, 2015). In particular, development of novel computational approaches for patient stratification leveraging high-throughput molecular data would significantly facilitate precision medicine and personalized treatment, which target discrete molecular subclasses of complex diseases with specific genetic or epigenetic profiles (Gentles *et al.*, 2015).

Clustering, an unsupervised exploratory analysis, has been widely used for patient stratification or disease subtyping (Chang *et al.*, 2005; Chen *et al.*, 2013). However, traditional clustering algorithms, such as *K*-means, hierarchical clustering, and spectral clustering, suffer from noise, data heterogeneity and high dimensionality that are associated with high-throughput molecular data (Andor *et al.*, 2016; Arnedos *et al.*, 2015). Ensemble clustering (a.k.a. consensus clustering) can merge some individually generated basic partitions, and ensure the final consensus partition maximally agrees with the basic ones (Strehl and Ghosh, 2002). This significantly helps us generate more robust partition, uncover interesting clusters, resist to noises, outliers and data variations, and fuse diverse solutions from multiple heterogeneous data sources (Liu *et al.*, 2015a). However, existing consensus clustering algorithms based on co-association matrix (Fred and Jain, 2005) are computationally expensive and require a large storage space, preventing them to handle high-throughput molecular data. Moreover, their interpretation of the consensus partition is often obscure. Last but not least, the existing consensus clustering methods become struggled to handle missing data, which significantly limits their use in practice.

In light of this, we propose the Entropy-based Consensus Clustering (ECC) for patient stratification. Our key idea is to introduce an entropy-based utility function to measure the similarity between each basic partition and the consensus one. We then rigorously map the complex consensus clustering problem into a simple *K*-means optimization problem, leveraging our previous theoretical framework that bridges consensus clustering and *K*-means clustering (Wu *et al.*, 2015). The entropy-based utility function is chosen due to its quick convergence and high performance (Liu *et al.*, 2015a). Our ECC method can also be easily extended to handle missing values and integrate various molecular data types. Extensive analysis of 110 synthetic datasets, 35 benchmark datasets and 13 cancer types with four molecular data types from The Cancer Genome Atlas (TCGA) demonstrate the significant advantages of ECC in terms of effectiveness, efficiency and robustness compared with several traditional clustering methods and state-of-the-art consensus clustering methods. Moreover, we systematically explored some key impact factors of ECC, such as the number of basic partitions and the generation strategy of basic partitions, which are crucial for practical use.

2 Materials and Methods

2.1 Background

Here we introduce the basic ideas of consensus clustering in the context of omics data (e.g. gene expression) analysis. Consensus clustering is originally developed for fusing several existing partitions into a robust one (Strehl and Ghosh, 2002), which can be roughly divided into two categories. The first category summarizes these basic partitions into a co-association matrix, which measures how many times a pair of instances simultaneously occur in the same cluster. Then based on the co-association matrix, any graph partition method can be conducted for the final solution. For example, HCC employs the agglomerative hierarchical clustering on the co-association matrix (Fred and Jain, 2005); SEC conducts the spectral clustering and solves it by a weighted *K*-means (Liu *et al.*, 2015b); IEC fuses infinite basic partitions based on the expectation of co-association matrix (Liu *et al.*, 2016). Another category is to design a utility function to measure the similarity among partitions and achieve the consensus partition by maximizing the utility function between the basic partitions and the consensus one. The pioneering work by Topchy *et al.* (2003) employed the Quadratic entropy as the utility function and solved it by *K*-means clustering. Along this line, Wu *et al.* gave a theoretic framework for *K*-means-based consensus clustering (Wu *et al.*, 2015; Liu *et al.*, 2015a). Generally speaking, via the guidance of objective functions, the methods in the second category offer better interpretability and greater robustness to clustering results than methods in the first category. However, it is quite challenging to design a proper utility function and make a balance between the high execution efficiency and high clustering quality.

Due to its superior performance, consensus clustering has recently been applied to gene expression data analysis (Iam-on *et al.*, 2010; Galdi *et al.*, 2014). For example, the linkage-based cluster ensemble (LCE) method first summarizes several basic partitions into a co-association matrix; then modifies the zero entries in the co-association matrix with the distance derived from the original data; and finally conducts spectral clustering to obtain the consensus partition (Iam-on *et al.*, 2010). As a variant of the LCE method, the Approximate SimRank-based (ASRS) method employs very similar idea with slightly different modification on the zero entries in the co-association matrix (Galdi *et al.*, 2014). Different from the existing consensus clustering methods (LCE and ASRS), our ECC method employs an entropy-based utility function for the guidance of fusing all the basic partitions into a consensus one, which has a more meaningful interpretation than existing consensus clustering methods. This can be understood as follows: (i) in our ECC method, the utility function measures the similarity between each basic partition and the consensus one, which naturally clarifies the contribution of each basic partition to the consensus one. In contrast, in co-association matrix based consensus clustering methods, the contribution of each basic partition to the consensus one is obscure. (ii) In the co-association matrix based consensus clustering methods, the number of co-occurrence of a pair of subjects is used to quantify their similarity. In our ECC method, we transform the consensus clustering problem into a *K*-means clustering problem with a modified distance. By this means, we can easily calculate the similarity between subjects in terms of the modified distance, which provides more meaningful interpretation than the number of co-occurrence.

Let $\mathcal{X}$ denote a gene expression dataset with n subjects and m genes. A partition of $\mathcal{X}$ into K crisp clusters is represented as either a collection of K subsets of instances in $\mathcal{C} = \{C_k | k = 1, \ldots, K\}$, with $C_k \cap C_{k'} = \varnothing$, $\forall k \neq k'$, and $\cup_{k=1}^{K} C_k = \mathcal{X}$, or as a label vector $\pi = (L_\pi(x_1), \ldots, L_\pi(x_n))^\top$, where L_π maps x_l to some label in $\{1, 2, \ldots, K\}$, $1 \leq l \leq n$. Suppose we have r basic partitions denoted as $\Pi = \{\pi^{(1)}, \pi^{(2)}, \ldots, \pi^{(r)}\}$ generated by some traditional clustering method (e.g. *K*-means) and there are K_ν clusters in $\pi^{(\nu)}$, for $1 \leq \nu \leq r$. The goal of consensus clustering is to find a consensus partition π by solving the following optimization problem:

$$\max_{\pi} \sum_{v=1}^{r} U(\pi, \pi^{(v)}), \tag{1}$$

where U is a utility function measuring the similarity at the partition-level between each basic partition and the consensus one. In other words, we expect to find an optimal partition that agrees with the basic ones as much as possible. Different utility functions measure the similarity of two partitions in different aspects, rendering different objective functions for consensus clustering. In this work, we employ an entropy-based utility function for its fast convergence and high quality (Wu *et al.*, 2015).

2.2 Methodology overview of ECC

Here, we summarize Entropy-based Consensus Clustering (ECC), for patient stratification. Consider an $n \times m$ matrix of molecular data of n subjects (or experiments, conditions, samples; corresponding to n rows) and m features (such as mRNAs; corresponding to m columns). Each subject can be represented by a point in the m-dimensional feature space, with different shapes representing different clusters the subjects belong to (Fig. 1A). There are three steps in the ECC pipeline. Step 1: We generate r basic partitions using K-means clustering with parameter K (i.e. the number of clusters) randomly chosen from 2 to $\sqrt{n}$ (Fig. 1B) (Wu *et al.*, 2015). Hereafter, we call this kind of basic partitions generation strategy *Random Parameter Selection* (RPS). Note that in this step we can use any basic clustering method. Here, we just choose K-means for its simplicity and high efficiency. In this work we set the number of basic partitions $r = 100$, which is large enough for a robust partition (see Supplementary Materials Section IV for performance of ECC with different numbers of basic partitions). Step 2: We derive a binary matrix from each basic partition via 1-of-Kcoding, where K is the cluster number in this basic partition and only one element in each row is 1, others are 0. We concatenate all those binary matrices into a large binary matrix **B** (Fig. 1C). Step 3: We employ an entropy-based utility function U_H to guide the fusion of all the r basic partitions into a consensus one (Fig. 1D). This is achieved by conducting K-means clustering on the binary matrix with a modified distance function and a user-defined cluster number K.

Our ECC method has three key features. First, it solves the consensus clustering problem in a *utility* way, which has more meaningful interpretation than any other consensus clustering methods. Here the utility function is applied to quantify the similarity between each of the r basic partitions and the consensus one. Maximizing the utility function requires us to find a single consensus partition that agrees with the basic ones as much as possible. Second, we uncover a remarkable equivalence relationship between an entropy-based utility function and a K-means distance function so that the complex utility maximization problem can be efficiently solved by the classic K-means method with a modified distance function (see Supplementary Materials Section I.A). Consequently, both the time and space complexity of ECC are linear in n (see Supplementary Materials Section I.B). This dramatically improves the efficiency of ECC in real-world applications (Wu *et al.*, 2015). Finally, ECC can naturally integrate multiple molecular data types measured from the same set of subjects, and easily handle missing values without any imputation. This significantly increases the power of ECC in clinically relevant patient stratification.

2.3 Entropy-based utility function

The core of ECC is to fuse these basic partitions into a consensus one based on an entropy-based utility function, which assures the consensus clustering algorithm to be highly efficient and robust. As formulated in Equation (1), a utility function is defined on two partitions π and $\pi^{(v)}$ to measures their similarity at the partition-level. We can employ the contingency table (Table 1) to calculate the entropy-based utility function.

Consider two partitions π and $\pi^{(v)}$, which contain K and K_v clusters, respectively. Let $n_{ij}^{(v)}$ denote the number of data objects belonging to both cluster C_j' in π' and cluster C_i in π. Define $n_{i+}^{(v)} = \sum_{j=1}^{K_v} n_{ij}^{(v)}$, and $n_{+j}^{(v)} = \sum_{i=1}^{K} n_{ij}^{(v)}$, $1 \le i \le K, 1 \le j \le K_v$.

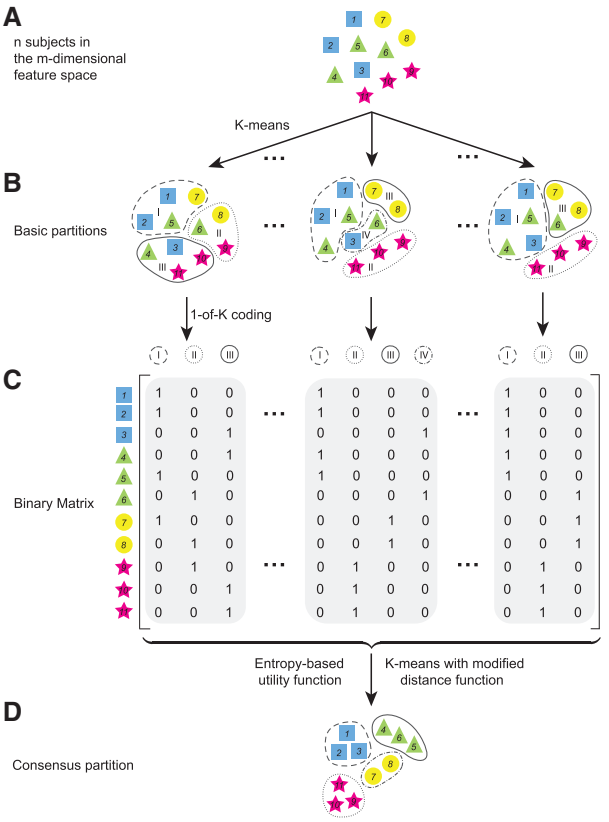

Fig. 1. Schematic diagram of the ECC pipeline. (**A**) n subjects are presented by n points (in this example, $n = 11$) and the features of these subjects can be mRNA expression, Protein expression, or any other molecular data. Different shapes represent the subjects in different disease subtypes (clusters). (**B**) K-means clustering is applied to the molecular data of the n subjects to obtain r basic partitions. For each basic partition, the cluster number K is randomly chosen from 2 to $\sqrt{n}$, and we highlight the K clusters using dashed line, dotted line, solid line, etc. (**C**) Each basic partition is transformed into 1-of-K coding, where K is the cluster number in each basic partition and only one element in each row is 1, others are 0. Concatenating all the basic partitions in 1-of-K coding form yields a large binary matrix B, which is a new representation of the original molecular data. (**D**) K-means clustering with a modified distance function derived from an entropy-based utility function is conducted on the binary matrix B for the final consensus clustering

Table 1. The contingency matrix

		$\pi^{(v)}$				
		$C_1^{(v)}$	$C_2^{(v)}$	$\cdots$	$C_{K_v}^{(v)}$	$\sum$
π	C_1	$n_{11}^{(v)}$	$n_{12}^{(v)}$	$\cdots$	$n_{1K_v}^{(v)}$	$n_{1+}^{(v)}$
	C_2	$n_{21}^{(v)}$	$n_{22}^{(v)}$	$\cdots$	$n_{2K_v}^{(v)}$	$n_{2+}^{(v)}$
	$\cdot$	$\cdot$	$\cdot$	$\cdots$	$\cdot$	$\cdot$
	C_K	$n_{K1}^{(v)}$	$n_{K2}^{(v)}$	$\cdots$	$n_{KK_v}^{(v)}$	$n_{K+}^{(v)}$
	$\sum$	$n_{+1}^{(v)}$	$n_{+2}^{(v)}$	$\cdots$	$n_{+K_v}^{(v)}$	n

Based on the contingency table, for π and π^v we define two discrete distributions, $P_i^{(v)} = (n_{i1}^{(v)}/n_{k+}, \ldots, n_{iKv}^{(v)}/n_{k+}), \forall\ i$, and $P^{(v)} = (n_{+1}^{(v)}/n, \ldots, n_{+j}^{(v)}/n, \ldots, n_{+Kv}^{(v)}/n)$. Then we define the entropy-based utility function,

$$U_H(\pi, \pi^{(v)}) = -\sum_{i=1}^{K} \frac{n_{i+}}{n} H(P_i^{(v)}) + H(P^{(v)}), \tag{2}$$

where H denotes the Shannon entropy.

Since H is a concave function, according to the Jensen's inequality, we have $U_H \geq 0$. A larger U_H indicates the higher utility from the two partitions in greater similarity. Note that U_H is asymmetric, with $U_H(\pi, \pi^{(v)}) \neq U_H(\pi^{(v)}, \pi)$, if $\pi \neq \pi^{(v)}$.

2.4 Entropy-based consensus clustering

Although it is crucial to design an appropriate utility function, how to efficiently optimize the objective function is another challenge. Thanks to the general K-means based Consensus clustering (Wu *et al.*, 2015), which has substantial advantage in terms of efficiency; we can transform the optimization problem in Equation (1) into a modified K-means clustering problem as follows.

Let $\mathbf{B} = (b_1, \ldots, b_n)^\top$ be the concatenated binary matrix derived from r basic partitions $\pi^{(1)}, \ldots, \pi^{(r)}$, with

$$b_l = (b_l^{(1)}, \ldots, b_l^{(v)}, \ldots, b_l^{(r)}), 1 \leq l \leq n, \tag{3}$$

$$b_l^{(v)} = (b_{l,1}^{(v)}, \ldots, b_{l,j}^{(v)}, \ldots, b_{l,K_v}^{(v)}), \tag{4}$$

$$b_{l,j}^{(v)} = \begin{cases} 1, & \text{if } L_{\pi^{(v)}}(l) = j \\ 0, & \text{otherwise} \end{cases}. \tag{5}$$

Apparently, $\mathbf{B}$ is a $n \times \sum_{i=1}^{r} K_i$ binary matrix, with $|b_l^{(v)}| = 1,\ \forall\ l, v$. A K-means clustering is directly conducted on $\mathbf{B}$ with the modified distance function as follows:

$$\begin{aligned} f(b_l, m_k) &= \sum_{v=1}^{r} D(b_l^{(v)} || m_k^{(v)}) \\ &= \sum_{v=1}^{r} \sum_{j} b_{l,j}^{(v)} \log b_{l,j}^{(v)} / m_{k,j}^{(v)}, \end{aligned} \tag{6}$$

where $m_k = (m_k^{(1)}, \ldots, m_k^{(r)})$ with $m_k^v = \sum_{b_l \in C_k} b_l^{(v)}/|C_k|$, and $D(b_l^{(v)}||m_k^{(v)})$ is the KL-divergence from $b_l^{(v)}$ to $m_k^{(v)}$.

From the utility function view, we measure the similarity at the partition-level, while we calculate the distance between subjects at the instance-level. Therefore, ECC combines two kinds of similarity together within the simple K-means framework to solve the consensus clustering problem. Moreover, ECC largely benefits from the entropy-based utility function, which has been shown to deliver high performance partition with fast convergence (Liu *et al.*, 2015a).

Consequently, the complex consensus clustering can be exactly mapped into a classic K-means clustering with a modified distance function, which has roughly linear time complexity and its convergence can also be guaranteed as well. The exactness of the mapping can be rigorously proved (see Supplementary Materials Section I.A for details). This mapping makes ECC very practical for large-scale molecular data analysis. Indeed only r elements are non-zero entries in each row of $\mathbf{B}$, thus the positions for these non-zero elements are needed, which leads the time complexity from $O(IKn\sum_{v=1}^{r} K_v)$ to $O(IKnr)$, where I is the number of iterations.

2.5 Handling missing values

Missing values are quite common in practice due to data collection or device failure, especially for the pan-omics data of a large population (in computer science, this kind of data is called multi-view). Typically there are two ways to handle those missing values. One is to just remove the instances (i.e., subjects) that have missing values in any single molecular data type (or any single view). Apparently, this is of great waste because those instances (subjects) might have values for many other views (molecular data types). The other way is to replace these missing values by default or average values. This would harm the original data structure and degrade the clustering performance. We can naturally resolve this issue within the framework of ECC. In particular, we consider that those missing values, which lead to missing labels in the basic partitions, do not provide any utility for the consensus fusion. If a basic partition has missing labels, we call it an *incomplete basic partition* (IBP). For IBP, we directly denote $b_{l,j}^{(v)}$ as an all-zero vector, which will not be involved in the distance calculation and centroid update. The following is the distance function for IBP:

$$f(b_l, m_k) = \sum_{v=1}^{r} 1(b_l^{(v)} \in \pi^{(v)}) D(b_l^{(v)} || m_k^{(v)}), \tag{7}$$

where 1 is the judgment function, which returns 1 with satisfied conviction and returns 0 otherwise. And $m_k = (m_k^{(1)}, \ldots, m_k^{(v)}, \ldots, m_k^{(r)})$ with

$$m_k^v = \frac{\sum_{b_l \in C_k \cap \pi^{(v)}} b_l^{(v)}}{|C_k \cap \pi^{(v)}|}. \tag{8}$$

3 Results

3.1 Datasets

In this work, we use 110 synthetic datasets to systematically evaluate the performance of ECC. The 110 synthetic datasets are generated by a well-established dynamical gene regulation model (Schaffter *et al.*, 2011):

$$\begin{aligned} F_i^{\text{mRNA}}(\mathbf{x}, \mathbf{y}) &= \frac{dx_i}{dt} = m_i \cdot f_i(\mathbf{y}) - \lambda_i^{\text{mRNA}} \cdot x_i, \\ F_i^{\text{Prot}}(\mathbf{x}, \mathbf{y}) &= \frac{dy_i}{dt} = r_i \cdot x_i - \lambda_i^{\text{Prot}} \cdot y_i, \end{aligned} \tag{9}$$

where m_i is the maximum transcription rate, r_i is the translation rate, λ_i^{mRNA} and λ_i^{Prot} are the mRNA and protein degradation rates, $\mathbf{x} \in \mathbf{R}^n$ and $\mathbf{y} \in \mathbf{R}^n$ are vectors of mRNA and protein concentration levels, respectively. $f_i(\cdot)$ computes the relative activation of gene. The topology of the gene regulatory network is encoded in the activation functions.

Among the 110 synthetic datasets, 55 of them are based on an Erdős–Rènyi random network of 500 genes, and the other 55 are based on a real human transcriptional regulation network of 2,723 genes (Chang *et al.*, 2005). Each dataset contains 200 subjects divided evenly into four subtypes (clusters). To simulate gene expression data for different subtypes (clusters), we assume that each subtype is characterized by a specific set of knocked-out genes. A more detailed description of the synthetic data generation can be found in Supplementary Materials Section I.

Besides the 110 synthetic datasets, 35 widely used cancer gene expression benchmark datasets (de Souto *et al.*, 2008) are employed to test the cluster validity of ECC. Also, 13 cancer types with four molecular data types from TCGA with survival information are

used for practical evaluation of ECC (Supplementary Tables S3 and S4).

3.2 Competitive methods

To demonstrate the advantages of ECC in terms of effectiveness and efficiency, we compared the performance of ECC with five traditional clustering methods: Agglomerative Hierarchical Clustering with Average-Linkage (AL), Single-Linkage (SL) and Complete-Linkage (CL), *K*-means Clustering (KM) and Spectral Clustering (SC); and two state-of-the-art consensus clustering methods: the Link-based Cluster Ensemble (LCE) (Iam-on *et al.*, 2010) and Approximate SimRank-based (ASRS) methods (Galdi *et al.*, 2014).

3.3 Evaluation metrics

Since the true labels for synthetic and benchmark datasets are available, we apply external measurements to objectively evaluate the performance of different clustering algorithms. Although there are many external measurements, some of them are biased. According to Wu *et al.* (2009), two normalized external metrics, *NMI* and R_n are chosen for proper evaluation of clustering performance. Both can easily be calculated from the contingency table.

Normalized Mutual Information (*NMI*), measures the mutual information between resulted cluster labels and ground truth labels, followed by a normalization operation to assure *NMI* ranges from 0 to 1. Mathematically, it is defined as:

$$NMI = \frac{\sum_{i,j} n_{ij} \log \frac{n \cdot n_{ij}}{n_{i+} \cdot n_{+j}}}{\sqrt{\left(\sum_i n_{i+} \log \frac{n_{i+}}{n}\right)\left(\sum_j n_{+j} \log \frac{n_{+j}}{n}\right)}}. \quad (10)$$

Normalized Rand Index, denoted as R_n measures the similarity between two partitions in a statistical way, which is defined as:

$$R_n = \frac{\sum_{i,j}\binom{n_{ij}}{2} - \sum_i \binom{n_{i+}}{2} \cdot \sum_j \binom{n_{+j}}{2} / \binom{n}{2}}{\sum_i \binom{n_{i+}}{2}/2 + \sum_j \binom{n_{+j}}{2}/2 - \sum_i \binom{n_{i+}}{2} \cdot \sum_j \binom{n_{+j}}{2} / \binom{n}{2}}. \quad (11)$$

Note that both *NMI* and R_n are positive measurements, i.e. a better partition has a larger *NMI* or R_n value. Although R_n is normalized, it can still be negative, which means that the partition is even worse than random label assignment.

To compare the overall performance of those clustering algorithms over the 35 benchmark cancer expression datasets, we propose an average performance score as follows:

$$Avg(A_i) = \frac{1}{d}\sum_{j=1}^{d} \frac{V(D_j, A_i)}{\max_i V(D_j, A_i)}, \quad (12)$$

where $V(D_j, A_i)$ denotes the performance (i.e., R_n or *NMI*) of Algorithm A_i on dataset D_j and d is the total number of benchmark datasets.

3.4 Evaluation on synthetic data

We first applied all those clustering methods to synthetic gene expression datasets with built-in cluster structure (see Supplementary Materials Section II) (Schaffter *et al.*, 2011). We found that ECC generally outperforms other methods in terms of its robustness against noise (see Supplementary Materials Section IV.A).

3.5 Evaluation on benchmark cancer gene expression data

We then evaluated ECC and other clustering methods on 35 widely used benchmark cancer gene expression datasets (de Souto *et al.*, 2008) (Supplementary Materials Section III.A). The detailed description of the 35 datasets was provided in Supplementary Table S3. Figure 2A shows the clustering performance of different algorithms measured by *NMI*. (Similar results were obtained for R_n, see Supplementary Materials Section IV.) We found that for most datasets, the three consensus clustering methods (LCE, ASRS and ECC) are superior to the five traditional clustering methods. Moreover, our ECC method achieves promising results on several datasets by a large margin, such as *Armstrong-2002-v2*, *Yeoh-2001-v1*, *Chowdary-2006* and *Golub-1999-v1*. Although LCE and ASRS yield reasonable performance on several datasets, they suffer from low robustness. For example, ASRS achieves 100 accuracy on *Nutt-2003-v3*, but it yields even worse results than that of random assignment on *Chen-2002*. We emphasize that, for unsupervised tasks, robustness is much more important than performance in practice when dealing with highly heterogeneous molecular data types (such as mRNA expression). Different from LCE and ASRS, ECC fuses the basic partitions in a utility way, which ensures highly meaningful interpretations with high stability for the final consensus partition. To compare the overall performance of those clustering methods over the 35 benchmark datasets, we employed the $Avg(A_i)$ in Equation (12) for evaluation, finding that ECC revealed significant advantages over all other methods in terms of average performance score.

We noticed that there are four specific datasets (*Gordon-2002, Khan-2001, Ramaswamy-2001* and *Shipp-2002*) for which all clustering methods yield very poor performance, most likely due to the presence of irrelevant or noisy features. We pointed that this difficulty cannot be easily resolved by any existing clustering methods. Yet, it can be alleviated by a complementary basic partition generation strategy of RPS, i.e. the *Random Feature Selection* (RFS) strategy, within the framework of ECC. To achieve that, we generated different sub-datasets by randomly selecting certain percentage of features (e.g. mRNAs) and then applied traditional clustering (e.g. *K*-means) to those sub-datasets to obtain basic partitions. Indeed, we found that for these four datasets, the performance of RFS exceeds RPS with all sampling ratios. This indicates that RFS helps us avoid noisy and irrelevant mRNA expressions (see Supplementary Materials Section IV for details).

Moreover, we also evaluated the robustness of ECC with missing treatments. By removing the labels in the basic partitions with different ratios, we found that ECC delivers high-quality partitions even with high missing ratios (see Supplementary Materials Section IV for details). It indicates that ECC is of good robustness and a suitable candidate for incomplete multi-view data analysis (Section 3.6).

In addition, ECC has tremendous merits in terms of computational cost. Figure 2B shows the execution time (in logarithmic scale) of the three consensus clustering methods (LCE, ASRS and ECC). The time complexity of ECC is $\mathcal{O}(InKr)$, where I is the number of iterations, n is the number of subjects, K is the number of clusters and r is the number of basic partitions. The space complexity of ECC is $\mathcal{O}(nr)$. For LCE and ASRS, the space complexities are both $\mathcal{O}(n^2)$; and the time complexities are $\mathcal{O}(n^2 \log n)$ and $\mathcal{O}(n^3)$, respectively. Naturally, ECC is more suitable for high-throughput molecular data analysis. For example, on *Yeoh-2002-v1*, ECC is 115 times and 1,600 times faster than LCE and ASRS, respectively.

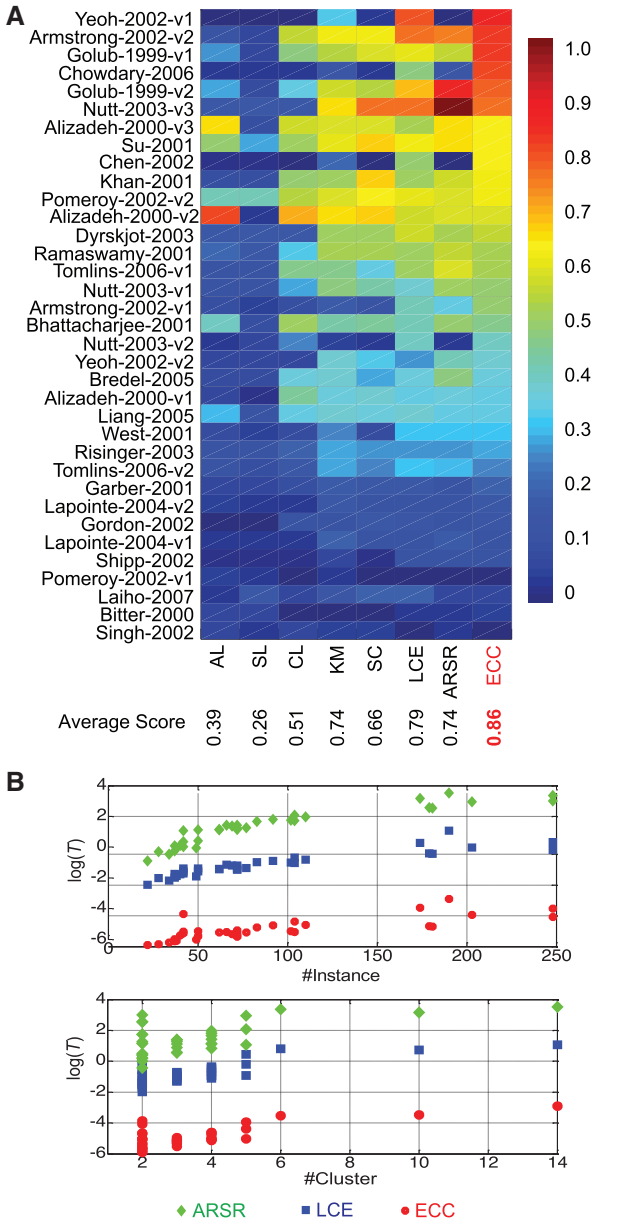

Fig. 2. The performance of ECC on 35 benchmark cancer gene expression datasets. (**A**) The performance of different clustering methods five traditional clustering methods: Agglomerative Hierarchical Clustering with Average-Linkage (AL), Single-Linkage (SL) and Complete-Linkage (CL), *K*-means Clustering (KM) and Spectral Clustering (SC) and two state-of-the-art consensus clustering methods: the Link-based Cluster Ensemble (LCE) and Approximate SimRank-based (ASRS) methods. The performance is measured by the Normalized Mutual Information (NMI). (**B**) The execution time (in logarithmic scale) of different consensus clustering methods as a function of the number of instances or the number of clusters

3.6 Translational application of ECC

The availability of massive and various molecular data types generated from large-scale and well-characterized cohorts across multiple cancer types provides an unprecedented opportunity for patient stratification. Here we demonstrated the translational applications of ECC based on 13 major cancer types from TCGA project with sufficient sample size and clinical profiles for four molecular data types: mRNA expression (RNA-seq V2), microRNA (miRNA) expression, protein expression, and somatic copy number alterations (SCNAs), as shown in Supplementary Table S4. For fair comparison, we collected well-established clinical subtypes from previous studies and hence empirically determined the number of clusters for the 13 TCGA cancer types. Then we applied survival analysis to evaluate the performance of different clustering methods in terms of $-\log_{10}P$ with P the log-rank test P-value. Here we employ survival analysis to evaluate clustering methods for real-world data that have no label information, because we expect that subjects in different clusters will have different survival distributions. The log-rank test is to compare the survival distributions of two or more groups, which determines if the observed number of events in each group is significantly different from the expected number (Supplementary Materials Section V and Tables S8–S11).

For each molecular data type (Protein, miRNA, mRNA and SCNA), we calculated the clustering performance of ECC against other clustering methods across the 13 TCGA cancer types. We found that ECC outperformed other methods (in terms of the number of significant survival analysis results across the 13 TCGA cancer types, as highlighted in dotted red rectangles in Fig. 3A–D) for any single molecular data type.

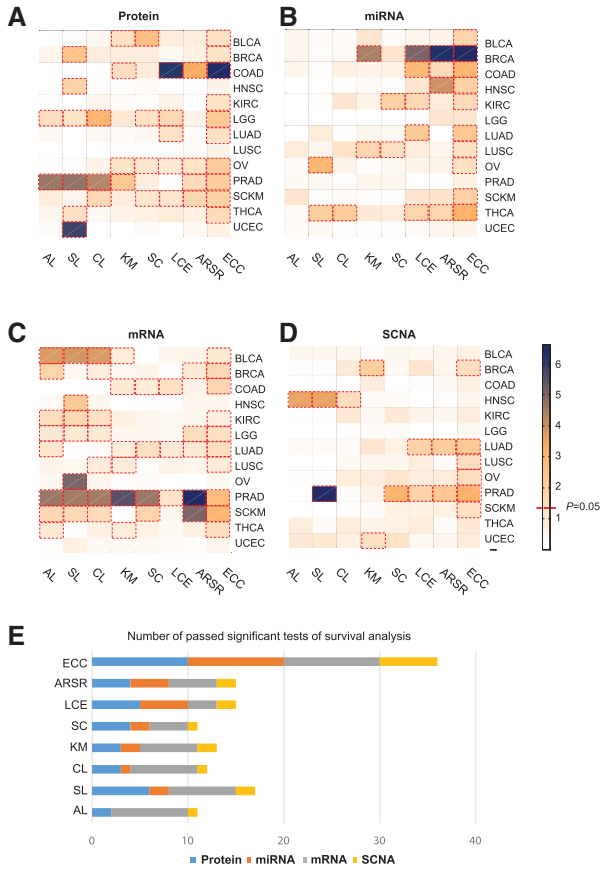

Fig. 3. Performance of seven different clustering methods on four molecular data types across 13 major cancer types from TCGA. Heatmaps show the survival analysis for 13 major cancer types using seven different clustering methods based on four molecular data types: (**A**) protein expression (protein), (**B**) miRNA expression (miRNA), (**C**) mRNA expression (mRNA) and (**D**) somatic copy number alterations (SCNA), respectively. We use the $-\log(P)$ to draw the heatmap and elements with dotted red rectangles have $P<0.05$. (**E**) This plot displays for each clustering method the times that it passes the significant tests of survival analysis, i.e. the number of dotted red rectangles in (A–D), over the 13 cancer types and the four different molecular data types (Color version of this figure is available at *Bioinformatics* online.)

Table 2. Performance of ECC on four molecular data types and its integration across 13 major cancer types from TCGA

ECC	Protein	miRNA	mRNA	SCNA	Integration
BLCA	**0.0212**	**0.0124**	**0.0187**	0.1910	**0.0027**
BRCA	**0.0313**	**6.4E−08**	**0.0011**	**0.0375**	**0.0131**
COAD	**3.3E−09**	**5.9E−04**	**7.8E−04**	0.2340	**8.7E−06**
HNSC	0.1820	**0.0090**	0.1160	0.3800	**0.0323**
KIRC	**0.0313**	**0.0223**	**0.0314**	0.1730	**9.8E−04**
LGG	**0.0016**	0.0751	**0.0039**	0.4130	**0.0119**
LUAD	**0.0245**	**0.0028**	**0.0028**	**0.0067**	**2.9E−05**
LUSC	0.1980	**0.0442**	**0.0258**	**0.0425**	**0.0393**
OV	**0.0021**	**0.0375**	0.2210	**0.0359**	**8.0E−04**
PRAD	**0.0020**	0.1840	**8.6E−06**	**7.3E−04**	**5.7E−04**
SKCM	**0.0035**	**0.0076**	**3.9E−06**	**0.0491**	**0.0131**
THCA	**0.0138**	**3.8E−04**	**0.0024**	0.1080	**0.0035**
UCEC	0.1310	0.2680	0.1240	0.1260	**0.0043**

Note: The performance is quantified by the log-rank test *P*-value of the survival analysis over the identified clusters (cancer subtypes). We highlight $P < 0.05$ in bold. With the integration of the four molecular data types, i.e. the pan-omics, ECC yields clusters that pass the significant test for all the 13 cancer types.

Table 2 shows the performance of ECC on four molecular data types and its integration across 13 major cancer types from TCGA. By integrating the four different molecular data types, ECC generated significant clusters (cancer subtypes) for all the 13 TCGA cancer types ($P < 0.05$, log-rank test, see Supplementary Table S12). Note that traditional clustering methods and existing consensus clustering methods cannot easily integrate multiple molecular data types, due to the presence of missing values for certain molecular data type of certain subjects. Yet, ECC can naturally resolve this issue by utility fusion, where missing values in basic partition provide no utility for the final fusion (Supplementary Fig. S9). Moreover, by integrating multiple molecular data types, ECC is effectively more robust to noise present in the data (partially because it has more data types to generate basic partitions). For example, in the case of uterine corpus endometrial carcinoma (UCEC), using any of the four molecular data types, ECC cannot yield significant clusters (Fig. 4A). Yet, by integrating multiple molecular data types (pan-omics), ECC yielded four significant clusters (Fig. 4B) with distinct patient survival curves ($P = 0.0043$, Fig. 4C); while using any single molecular data type the clusters generated by ECC do not pass the significance test of survival analysis ($P > 0.05$, Supplementary Tables S8–S11). In addition, subtypes identified by ECC via integrating four molecular data types were closely associated with the clinical subtypes on a histological basis in UCEC (Fig. 4D). For instance, subtype 2 with most aggressive uterine tumor shows poor survival than subtype 1 with the less aggressive uterine tumors. Similar trends are also observed in ovarian serous cystadeno-carcinoma (OV, $P = 7.79 \times 10^{-4}$) and prostate adenocarcinoma (PRAD, $P = 5.27 \times 10^{-4}$).

4 Discussion

In sum, we showed that ECC owns significant advantages in terms of cluster validity, execution time and space complexity and robustness compared with other clustering methods in patient stratification. We demonstrated that ECC with RFS strategy can alleviate the detriment effect of irrelevant and noisy features. Moreover, ECC displays superior performance on the pan-omics data by integrating multiple molecular data types than that of single molecular data type. We anticipated that integrating more types of both molecular and clinical data, such as somatic mutations, DNA methylation, functional genomic data generated from CRISPR/Cas9 (Cong *et al.*, 2013), proteogenomics (Zhang *et al.*, 2014), radiomics(Aerts *et al.*, 2014) and electronic medical records (Denny *et al.*, 2013), will further improve patient stratification. Altogether, our ECC method paves the way to a much more refined representation and understanding of various molecular data types, facilitating the development of precision medicine.

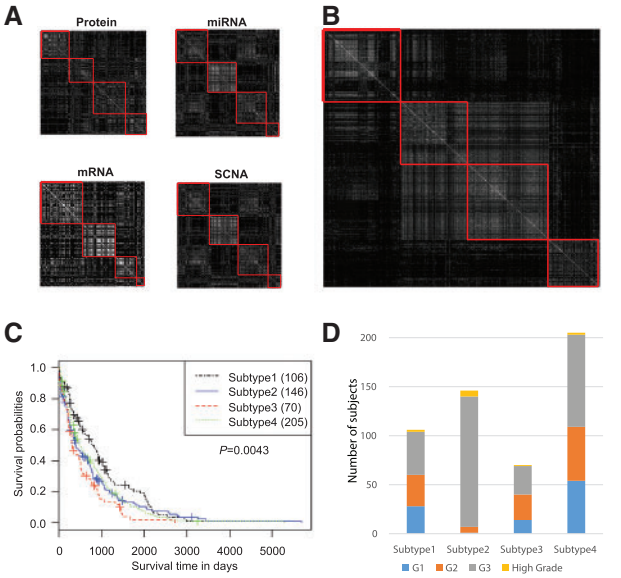

Fig. 4. Performance of ECC for uterine corpus endometrial carcinoma (UCEC) subjects from TCGA. The similarity matrices calculated from the four clusters generated by ECC using single molecular data type (**A**) and pan-omics data (**B**) of UCEC. The survival curves (**C**) and the composition of different clinical subtypes (**D**) for the four clusters generated by ECC using pan-omics data of UCEC

Acknowledgement

We thank Rulla Tamimi and Edwin Silverman for valuable discussions.

Funding

John Templeton Foundation (Award number 51977); National Academy of Sciences- Grainger Foundation Frontiers of Engineering Award (2000006959).

Conflict of Interest: none declared.

References

Aerts,H. *et al.* (2014) Decoding tumour phenotype by noninvasive imaging using a quantitative radiomics approach. *Nat. Commun.*, **4**, 4006.

Andor,N. *et al.* (2016) Pan-cancer analysis of the extent and consequences of intratumor heterogeneity. *Nat. Med.*, **22**, 105–113.

Arnedos,M. *et al.* (2015) Precision medicine for metastatic breast cancer—limitations and solutions. *Nat. Rev. Clin. Oncol.*, **12**, 693–704.

Biankin,A. *et al.* (2015) Patient-centric trials for therapeutic development in precision oncology. *Nature*, **526**, 361–370.

Bolouri,H. *et al.* (2016). Big data visualization identifies the multidimensional molecular landscape of human gliomas. *Proc. Natl. Acad. Sci. USA*, **113**:5394–5399.

Chang,H. *et al.* (2005). Robustness, scalability, and integration of a wound-response gene expression signature in predicting breast cancer survival. *Proc. Natl. Acad. Sci. USA*. **102**, 3738–3743.

Chen,G. *et al*. (2013). Biclustering with heterogeneous variance. *Proc. Natl. Acad. Sci. USA*, **110**, 12253–12258.

Cong,L. *et al*. (2013) Multiplex genome engineering using crispr/cas systems. *Science*, **339**, 819–823.

de Souto,M. *et al*. (2008) Clustering cancer gene expression data: a comparative study. *Bioinformatics*, **9**, 497.

Denny,J. *et al*. (2013) Systematic comparison of phenome-wide association study of electronic medical record data and genome-wide association study data. *Nat. Biotechnol*., **31**, 1102–1111.

Fred,A. and Jain,A. (2005) Combining multiple clusterings using evidence accumulation. *IEEE Trans. Pattern Anal. Mach. Intell*., **27**, 835–850.

Galdi,P. *et al*. (2014). Consensus clustering in gene expression. In *International Meeting on Computational Intelligence Methods for Bioinformatics and Biostatistics*, pp. 57–67. Springer, Cambridge, UK.

Gentles,A. *et al*. (2015) The prognostic landscape of genes and infiltrating immune cells across human cancers. *Nat. Med*., **21**, 938–945.

Iam-On,N. *et al*. (2010) Lce: a link-based cluster ensemble method for improved gene expression data analysis. *Bioinformatics*, **26**, 1513–1519.

Kamburov,A. *et al*. (2015). Comprehensive assessment of cancer missense mutation clustering in protein structures. *Proc. Natl. Acad. Sci. USA*, **12**, E5486–E5495.

Lapointe,J. *et al*. (2004). Gene expression profiling identifies clinically relevant subtypes of prostate cancer. *Proc. Natl. Acad. Sci. USA*, **101**, 811–816.

Liu,H. *et al*. (2015a). Dias: a disassemble-assemble framework for highly sparse text clustering. In: *Proceedings of SIAM International Conference on Data Mining*.

Liu,H. *et al*. (2015b). Spectral ensemble clustering. In: *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Ming*.

Liu,H. *et al*. (2016). Infinite ensemble for image clustering. In: *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*.

Schaffter,T. *et al*. (2011) Genenetweaver: in silico benchmark generation and performance profiling of network inference methods. *Bioinformatics*, **27**, 2263–2270.

Strehl,A., and Ghosh,J. (2002) Cluster ensembles—a knowledge reuse framework for combining partitions. *J. Mach. Learn. Res*., **3**, 583–617.

Topchy,A. *et al*. (2003). Combining multiple weak clusterings. In: *Proceedings of International Conference on Data Mining*.

Uhlen,M. *et al*. (2016) Transcriptomics resources of human tissues and organs. *Mol. Syst. Biol*., **12**, 862.

Wu,J. *et al*. (2009). Adapting the right measures for k-means clustering. In: *Proceedings of ACM SIGKDD International Conference on Knowledge Discovery and Data mining*.

Wu,J. *et al*. (2015) K-means-based consensus clustering: a unified view. *IEEE Trans. Knowl. Data Eng*., **27**, 155–169.

Zhang,B. *et al*. (2014) Proteogenomic characterization of human colon and rectal cancer. *Nature*, **513**, 382–387.

Zhu,Q. *et al*. (2015) Targeted exploration and analysis of large cross-platform human transcriptomic compendia. *Nat. Methods*, **12**, 211–214.