Dimitrakopoulou, K., Vrahatis, A.G., Wilk, E., Tsakalidis, A.K., Bezerianos, A.
OLYMPUS: An automated hybrid clustering method in time series gene expression. Case study: Host response after Influenza A (H1N1) infection
(2013) Computer Methods and Programs in Biomedicine, 111 (3), pp. 650-661.

OLYMPUS: An automated hybrid clustering method in time series gene expression. Case study: Host response after Influenza A (H1N1) infection

Konstantina Dimitrakopoulou[1†], Aristidis G. Vrahatis[2†], Esther Wilk[3], Athanasios K. Tsakalidis[2] and Anastasios Bezerianos[1,4*]

[1]School of Medicine, University of Patras, Patras 265 00, Greece

[2]Department of Computer Engineering and Informatics, University of Patras, Patras 265 00, Greece

[3]Department of Infection Genetics, Helmholtz Centre for Infection Research and University of Veterinary Medicine Hannover, 38124 Braunschweig, Germany

[4]SINAPSE Institute, Center of Life Sciences, National University of Singapore, Singapore 117456

†The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

*To whom correspondence should be addressed.

Email: bezer@upatras.gr

ABSTRACT

The increasing flow of short time series microarray experiments for the study of dynamic cellular processes poses the need for efficient clustering tools. These tools must deal with three primary issues: first, to consider the multi-functionality of genes; second, to evaluate the similarity of the relative change of amplitude in the time domain rather than the absolute values; third, to cope with the constraints of conventional clustering algorithms such as the assignment of the appropriate cluster number. To address these, we propose OLYMPUS, a novel unsupervised clustering algorithm that integrates Differential Evolution (DE) method into Fuzzy Short Time Series (FSTS) algorithm with the scope to utilize efficiently the information of population of the first and enhance the performance of the latter. Our hybrid approach provides sets of genes that enable the deciphering of distinct phases in dynamic cellular processes.

We proved the efficiency of OLYMPUS on synthetic as well as on experimental data. The discriminative power of OLYMPUS provided clusters, which refined the so far perspective of the dynamics of host response mechanisms to Influenza A (H1N1). Our kinetic model sets a timeline for several pathways and cell populations, implicated to

participate in host response; yet no timeline was assigned to them (*e.g.* cell cycle, homeostasis). Regarding the activity of B cells, our approach revealed that some antibody-related mechanisms remain activated until day 60 post infection.
The Matlab codes for implementing OLYMPUS, as well as example datasets, are freely accessible via the Web (http://biosignal.med.upatras.gr/wordpress/biosignal/).

1. Introduction

It is already established that physiological processes such as cell cycle or immune response are dynamical due to the invariant activities of the underlying genes. The increasing flow of time series microarray experiments poses the need for optimized and tailored clustering algorithms that set appropriately the optimal number of clusters, capture the multi-functionality of genes and consider the relative change of expression as well as the temporal information, regardless of absolute values.

With the advent of time series microarray experiments, many standard clustering algorithms were recruited for analysis such as hierarchical clustering, k-means [1] and self-organizing maps [2], which however treat measurements taken at different time points as independent, ignoring the sequential nature of time series data. Lately, a significant number of methods were designed specifically for the time series analysis [3-6]. For instance, Ramoni *et al.* [7] suggested the grouping of genes whose dynamics can be expressed roughly with the same auto-regressive equation. Also, works like [8-9] proposed the mixed-effects model using B-splines, which can estimate the continuous curves of gene profiles and cluster them simultaneously. Similarly, Song *et al.* [10] employed the B-splines to reconstruct the expression profiles, with the functional principal component analysis applied before clustering for achieving dimensionality reduction. Further, Nascimento *et al.* [11] proposed a Bayesian method that considers simultaneously an autoregressive panel data model fit and a hierarchical clustering of the parameter estimates from this model, combining so the temporal autocorrelation of the gene expression with the model-based clustering.

Further progress was observed when many standard clustering algorithms in gene expression analysis, such as k-means and fuzzy c-means, were coupled with Evolutionary Algorithms (EAs) [12-13] to optimize partitioning. EAs are meta-heuristics widely believed to be effective on NP-hard problems, such as clustering, being able to provide near-optimal solutions in reasonable time. Hybrid methods like the study of [14] have combined k-means with Genetic Algorithms GAs (a subcategory of EAs), resulting to the Incremental Genetic K-means Algorithm (IGKA), which converges to a global optimum faster than the stand alone GA and without sensitivity to the initialization of prototypes. Extensive is the research regarding the recruitment of EAs to solve the fuzzy problem (see the review of Horta *et al.* [15]). Gong *et al.* [16] attempted to improve the Differential Evolutionary (DE) algorithm by integrating the one-step fuzzy c-means algorithm. Other works like [17]

introduced an evolutionary fuzzy clustering algorithm with knowledge based evaluation based on Bayesian validation and prior knowledge. With respect to time series gene expression analysis, hybrid approaches like [18-19] attempted to consider the shape of profiles rather than the distance but ignored the unevenness in time sampling and failed to embed the temporal information in the similarity metric.

One fundamental deal in the design of clustering algorithms is the development or usage of internal (based on the information intrinsic to the data alone), external (based on previous knowledge about data) and relative (tools for comparing and choosing the best parameters among different algorithm settings) validation measures that assess the quality of the clustering solution obtained. However, external validation measures like F-measure, purity, entropy and Normalized Mutual information (NMI) are not applicable in transcriptome analysis, in which the true biological structure is far from known. Also, relative indices like Figure of Merit (FOM) and instability index have been proved to perform lower than other indices (e.g. internal) and be highly time consuming. In literature, several unsupervised clustering algorithms employed internal measures [20] like Hubert's correlation, Rand Statistics, Jaccard's coefficient, Calinski-Harabasz, Davies-Bouldin, Bayesian Information Criterion (BIC), Dunn and Silhouette indices to overcome limitation such as the need for specifying the optimal number of clusters. However, the choice of the relevant measure is usually taken after prioritizing the needs for precision, accuracy, time-related computational resources and the level of algorithm dependencies.

Deciphering the dynamics of complex diseases such as Influenza A via time series microarray experiments is common strategy. Influenza infection is a major health problem worldwide and causes many fatalities every year [21]. The level of severity as well as the outcome during Influenza A infection is defined by many host and viral factors. Moving further, the triggered host responses are highly dynamic [22] and long lasting [23], where the related cell populations and pathways are recruited in specific time intervals even after two months after the onset of infection. The ultimate goal in this kind of experimental design is to define a kinetic model that sets the timeline of response mechanisms throughout the acute innate immune response phase, to the clearance of the virus until the establishment of the long-lasting immunity and the restoration of homeostasis. In such a scenario, an optimal clustering result would assist significantly in the monitoring of the participating pathways and cell types throughout the distinct phases and, differentiating even between overlapping activities, *e.g.* such as between NK and T cells. However, challenging remains the choice of the suitable clustering algorithm, which in turn should meet the following criteria: (a) include the temporal information in the similarity metric, (b) be insensitive to the initialization parameters and (c) deal with the determination of the optimal cluster number.

Towards this orientation, we present evOLutionary fuzzY teMPoral clUStering (OLYMPUS) algorithm, a hybrid approach that integrates Differential Evolution (DE) algorithm into Fuzzy Short Time Series (FSTS) method, with the latter employing a similarity metric that considers the temporal features of dynamic expression data. Additionally, the Bayesian Information Criterion (BIC) was integrated into

OLYMPUS in an attempt to determine the best number of clusters for a given data set [24]. The FSTS method revised the standard fuzzy c-means method by incorporating the Short Time Series (STS) distance into the equations for computing the membership matrix and the prototypes (or cluster centers), thus developed a fuzzy time series clustering algorithm [25]. Despite the fact that this method converges fast, it is easily trapped in local optima, since it is sensitive to initialization parameters. In order to overcome this limitation, we embedded the DE algorithm into our partitioning scheme, which is a stochastic and global optimization search method, efficient at exploring the search space and locating the region of global minimum due to its population-based nature [26].

The efficacy of our approach on synthetic as well as experimental datasets is highlighted next. In synthetic analysis, we created random datasets, each one having a pre-fixed dimension size and number of optimal clusters to be detected. The objective was to show the efficiency of OLYMPUS in a two-way scheme; on one side, we present the ability of OLYMPUS in recovering the optimal cluster number against known unsupervised clustering methods and on the other side, we displayed its enhanced performance, in terms of accuracy, against other (un)supervised fuzzy or hybrid algorithms. In the latter analysis, we used mouse unevenly sampled time series gene expression data, which record the host response mechanisms over a period of 60 days after infection with influenza A (H1N1) virus. Our goal was to refine the kinetic model of host response mechanisms as presented in recent study [23] and show the discriminative power of OLYMPUS in capturing the coordinated activity of genes that were implicated to be involved in the response, yet not still clearly identified. We managed to set a fine-tuned timeline for the activation as well as suppression of several involved pathways and cell populations, which accords with prior biological knowledge and complements substantially with the addition of other biological processes that participate directly or indirectly in the host response (*e.g.* cell cycle).

2. Methodology

2.1. Differential Evolutionary Algorithm (DE)

Evolutionary algorithms (EAs) are stochastic search methods that mimic natural biological evolution. The crucial idea behind EAs is to have a number of potential solutions after applying the Darwinian evolutionary processes in order to obtain better approximations of the optimum solution [26]. Evolutionary algorithms (EA) are suitable for overcoming the problem of local optima as introduced by many conventional algorithms such as Fuzzy C-Means (FCM). A sub-category of EAs, the Differential Evolution (DE) approach, is a population-based stochastic parallel direct search method that optimizes a problem by iteratively trying to improve a candidate solution with regard to a given measure of quality, which in many cases is the minimization of a problem-oriented objective fitness function. Below, we provide the description of DE steps.

2.1.1. Initialization

The initial population represents potential solutions (individuals) over the optimization search space (a potential solution is the identification of cluster centers in time series gene expression data). Each individual (each set of cluster centers) is represented as $x_g^i = \left[x_{g,1}^i . x_{g,2}^i \ldots x_{g,D}^i \right]$, for $i = 1,2,\ldots NP$, where *NP* is the size of population, $g = 0,1,\ldots, g_{max}$ is the current generator and g_{max} the maximum number of generations. Initialization is realized through a random number distribution to generate the potential individuals in the optimization search space. The optimization search space is defined by lower and upper bound values, i.e. $L = \left[L_1, L_2 \ldots L_D \right]$ and $U = \left[U_1, U_2 \ldots U_D \right]$. Hence, we initialize the j-th dimension of the i-th individual according to:

$$x_{0,j}^i = L_j + rand_j\left(0,1\right) \cdot \left(U_j - L_j\right) \quad (1),$$

where $rand_j\left(0,1\right)$ is a uniformly distributed random number defined in the $\left[0,1\right]$ range.

2.1.2. Mutation:

After initialization, the mutation operator is introduced. For each individual a new one is derived through the combination of randomly selected and pre-specified individuals, called mutant individual $\tilde{x}_g^i$. The most well-known mutation strategies reported in literature are:

$$\tilde{x}_g^i = x_g^{r1} + F\left(x_g^{r2} - x_g^{r3}\right) \quad (2),$$

$$\tilde{x}_g^i = x_g^{best} + F\left(x_g^{r1} - x_g^{r2}\right) \quad (3),$$

where $x_g^{r1}, x_g^{r2}, x_g^{r3}$ are the different individuals of the population, $r1, r2, r3$ are distinct integers uniformly chosen from the set $\left\{1,2,\ldots, NP\right\}$, x_g^{best} indicates the fittest (best) individual of the current generation and F is a fixed parameter ranging in the interval [0,1], which is used to generate all mutation vectors in all generations. These particular mutation schemes are named DE/rand/1 and DE/best/1 respectively. The DE schemes are defined as DE/x/y/z, where DE stands for 'differential evolution', *x* is the type of vector to be perturbed, *y* is the number of difference vectors associated to the perturbation of *x*, and *z* stands for the type of crossover used (exp: exponential; bin: binomial). In literature, a substantial amount of research has been devoted to the development and analysis of efficient mutation operators and their dynamics [27-28].

2.1.3. Crossover:

The operation of crossover is employed so as to increase the diversity of the population. A trial individual $\tilde{x}_{g,j}^{i}$ is constructed as a competitor for each target individual x_g^i, based on its parents x_j^i and $\tilde{x}_j^i$ according to the following crossover rules:

$$\tilde{x}_{g,j}^{i} = \begin{cases} \tilde{x}_{g,j}^{i} \ \ if \ \ R^i \le C_R \ \ or \ j = I_j \\ x_{g,j}^{i} \ \ if \ \ R^i > C_R \ \ and \ \ j \ne I_j \end{cases} . \qquad (4)$$

where R^i is a uniformly distributed random number in [0,1], different for every j^{th} component of every individual. $I_j \in \{1, 2,\ldots, D\}$ is a randomly chosen integer, which ensures that at least one component of the mutant vector will be assigned to the target individual. $C_R \epsilon [0,1]$ is a constant drawn randomly for each *j*, used to generate all trial vectors in all generations. In literature there are two main crossover operators, namely the binomial and the exponential, each one having its own characteristics (a detailed description can be found in [28]). Here, we adopt the widely used binomial crossover operator which exhibits stable and robust performance.

2.1.4. Selection:

Last, in order to acquire the best trial individuals in the next generation, the selection operation is initiated. The target individual x_g^i is compared with the trial vector $\tilde{x}_g^i$ and the one with the lowest objective function value is introduced to the next generation

$$x_{g+1}^{i} = \begin{cases} \tilde{x}_g^i \ \ if \ \ \mathrm{f}(\tilde{x}_g^i) < \mathrm{f}(x_g^i) \\ x_g^i \ \ otherwise \end{cases} , \qquad (5)$$

where *g* denotes the number of generations and *f(x)* is the problem-defined objective function to be minimized. Finally, mutation, crossover and selection procedures continue until some stopping criterion is reached.

2.2. Evolutionary Fuzzy Temporal Clustering algorithm (OLYMPUS)

The OLYMPUS algorithm (Supplementary File 1) is an unsupervised hybrid approach that integrates DE algorithm into Fuzzy Short Time Series (FSTS) method to deal with the sensitivity of the latter to the initialization parameters (Fig.1). In other words, DE serves as an optimization strategy that avoids FSTS being trapped in local optima by guiding it in a more promising search space, in which FSTS will approach fast the optimal solution. The FSTS method is a variant of the classic Fuzzy C-Means algorithm in the sense that it incorporates a similarity metric, Short Time Series (STS), which in turn manages to find the differences in the shapes, as defined by the

relative change of expression and the corresponding temporal information, regardless of the difference in absolute values [25].

2.2.1. Short Time Series (STS) Similarity Metric

The STS distance corresponds to the square root of the sum of the squared differences of the slopes obtained by considering time-series as linear functions between measurements. The STS distance between two time-series *x* and *v* (e.g. two gene expression profiles across all recorded time points) is defined as:

$$d_{STS}^2 = \sum_{k=1}^{n_t} \left(\frac{v_{(k+1)} - v_k}{t_{(k+1)} - t_k} - \frac{x_{(k+1)} - x_k}{t_{(k+1)} - t_k} \right)^2 \quad (6),$$

where n_t is the number of time-points and $t_{(k+1),} t_k$ are two successive time-points. The aforementioned framework addresses the questions posed in cases where the goal is to unravel the dynamics of the biological system under investigation. On one hand, the fuzziness correlates to the realistic nature of genes, and on the other hand DE overcomes the local optima and provides near-optimal solution. Finally, the STS metric enables the grouping of genes with similar expression trends in terms of absolute values along with genes, whose fold change values are dissimilar, however with similar shapes. This attribute bridges a gap common in transcriptome analysis; some genes are recruited among other during a process, however they are transcribed at a low, nevertheless significant, rate (relative to control state) in comparison to the rest. This trait would be lost if Euclidean distance was employed. In the case of Pearson correlation coefficient, the similarity in shape between two profiles is included but not the varying length between recorded time points. Other temporal informative metrics, such as the one proposed in Smith *et al.* [4], group genes after incorporating the time shift between profiles; however, this kind of metrics is efficient and biologically realistic in experiments with high sampling resolution and low-order time scale. In other words, time shift is informative in experiments that record a cellular process at several time points at the level of hours, and not at few time points at the level of days (such as the experimental data used in this paper).

2.2.2. Algorithmic Stages

In general, the OLYMPUS algorithm operates in two levels. At first, the DE strategy is applied to obtain the most promising cluster centers by increasing the possibility to avoid the local minima. At second level, the FSTS algorithm is recruited until it converges to the optimal position.

Specifically, at first level, DE initializes randomly the individuals, with each individual representing all the cluster center profiles in the gene expression dataset, which are reshaped into a vector (cluster vector). This initialization occurs in the search space, whose upper (U) and lower (L) bounds are defined by the maximum and

minimum value in gene expression dataset respectively. For n time points and n_c clusters, each individual X^i, with $i = 1, 2, \ldots NP$, where *NP* is the size of population, is expressed as:

$$X^i = X_{11} X_{12} \ldots X_{1n} \ldots\ldots X_{j1} X_{j2} \ldots X_{jn} \ldots\ldots X_{n_c 1} X_{n_c 2} \ldots X_{n_c n} \quad (7).$$

Then, the processes of mutation, crossover, and selection are executed for a user-defined number of iterations (*MaxIterDE*). Here, we utilize binomial crossover and the classical mutation strategy, DE/rand/1, in these equations (1, 2, 3, 4 and 5). Through the evolutionary process, individuals get better positions in search space, and after *MaxIterDE* iterations we obtain the best individual which minimizes optimally our objective fitness function. At this point, we describe the fitness function of OLYMPUS, which is applied both in DE and later in FSTS algorithm. This function derives after substituting the STS distance in the fitness function of the standard Fuzzy C-Means (FCM) algorithm. In detail:

$$J(x, v, u) = \sum_{i=1}^{n_c} \sum_{j=1}^{n_g} u_{ij}^w \sum_{k=0}^{n_t - 1} \left(\frac{v_{i(k+1)} - v_{ik}}{t_{(k+1)} - t_k} - \frac{x_{j(k+1)} - x_{jk}}{t_{(k+1)} - t_k} \right)^2 \quad (8),$$

where n_c is the number of clusters, n_g the number of genes, t is each time-point and n_t is the number of time-points as provided in the microarray experiment, u_{ij} is the degree of membership of vector x_j in the cluster j , x_j is the j^{th} gene expression profile and v_i is the i^{th} cluster center profile. w is often called the fuzzifier parameter and determines the fuzziness of the clustering. The clustering becomes fuzzier as w gets larger.

On second level, the best individual is introduced to FSTS method as a set of initial cluster centers. This initialization helps FSTS to avoid being trapped in local minima. Next, we run the FSTS method for each initial cluster center and for a user-defined number of iterations (*MaxIterFSTS*). The backbone idea of FSTS algorithm is to acquire a prototype (center) v_k that minimizes the fitness function (Eq. 8). Therefore, it is necessary to obtain prototypes that take into account the similar slopes according to temporal information. The prototypes are estimated after calculating the partial derivative of the fitness function and solve for v_k after setting it equal to zero, as described in [25]:

$$v(i,n) = \sum_{r=2}^{n-3}\left[\frac{m_{ir}\prod_{q=1}^{r-1}c_q + \left(\prod_{q=r+1}^{n-1}a_q + \prod_{q=r+1}^{n-1}c_q + \sum_{p=r+3}^{n}\prod_{j=p-1}^{n-1}c_j + \prod_{j=r+1}^{p-2}a_j\right)}{\prod_{q=2}^{n-1}c_q}\right] + \left[\frac{m_{i(n-1)}\prod_{q=1}^{n-2}c_q + m_{i(n-2)}\prod_{q=1}^{n-3}c_q\,(a_{(n-1)}+c_{(n-1)})}{\prod_{q=2}^{n-1}c_q}\right] \quad (9),$$

where $1 \le i \le n_c$ and $3 \le n \le n_t$,

$$m_{ik} = \frac{\sum_{j=1}^{n_g} u_{ij}^{w}(d_k x_{j(k-1)} + e_k x_{jk} + f_k x_{j(k+1)})}{\sum_{j=1}^{n_g} u_{ij}^{w}},$$

$a_k = -(t_{(k+1)} - t_k)^2$, $b_k = -(a_k + c_k)$, $c_k = -(t_k - t_{(k-1)})^2$, $d_k = (t_{(k+1)} - t_k)^2$, $e_k = -(d_k + f_k)$, and $f_k = (t_k - t_{(k-1)})^2$.

Each element of the partition matrix is setting the membership degree of each gene to each cluster (Eq. 10) and depends on STS distance to segregate the clusters in the framework stated above.

$$u_{ij} = \frac{1}{\sum_{q=1}^{n_c}\left(d_{STS}\left(x_i, v_j\right) / d_{STS}\left(x_i, v_q\right)\right)^{\frac{1}{w-1}}} \quad (10),$$

2.2.3. Bayesian Information Criterion

The Bayesian Information Criterion (BIC) has been successfully applied to the problem of determining the number of components in model-based clustering [29]. We integrated BIC into OLYMPUS in an attempt to estimate the optimal cluster number. The equation for BIC is shown below:

$$BIC_{STS} = n_g \ln\left(\frac{RSS_{STS}}{n_g}\right) + n_c \ln\left(n_g\right) \quad (11),$$

where n_g is the number of data points (genes) and n_c is the number of clusters being considered. The first term of the formula is the log-likelihood and the second term is penalty. Given any two estimated models, the model with the lower value of BIC is the one to be preferred. The BIC is an increasing function of RSS and an increasing function of n_c. RSS is the residual sum of squared errors, usually based on Euclidean distance, but in our case adapted to STS distance. In detail:

$$RSS_{STS} = \sum_{i=1}^{n} d_{STS}^{2}\left(x_i, c_j\right) \quad (12),$$

where x_i is the i^{th} vector and c_j is the centroid with the largest membership value.

The two levels (DE and FSTS method) run iteratively until a minimum BIC value is achieved. We propose the user to run OLYMPUS after setting the cluster number in the range $\{N_{cmin}, N_{cmax}\}$, where N_{cmin} is set to 2 and N_{cmax} equal to $\sqrt{n_g}$ according to the rule of thumb used by many investigators in literature [30-31]. The behavior of BIC function monotonically decreases until it hits the minimum at the correct cluster number and monotonically increases as the n_c increases (see Supplementary File 2 – Fig.1). When OLYMPUS hits a minimal BIC value after which the value increases gradually, the respective BIC score is considered as a candidate global minimal value (optimal cluster number). Then, OLYMPUS compares this minimal value with the derived BIC values after sampling in a specified rate the rest search space; in case a smaller value is found during sampling procedure, the local region around is further explored and a new candidate global minimal value is set for comparison until sampling ends. However, the synthetic analysis (as described in the next section) showed that the behavior of BIC function (in over 1000 runs) remains stable and the sampling procedure does not introduce novel candidate solutions (Chi-square test, $P < 0.0001$).

3. Synthetic Analysis

At this section, we evaluate the algorithmic implementation proposed in the previous one. To this end, we employ a two-way scheme of simulation analysis. This procedure gives the opportunity to pre-design (and hence know beforehand) the structure of the data that the clustering procedure aims to recover. At first level, we compared the ability of OLYMPUS in determining the correct cluster number against other known unsupervised clustering algorithms such as Density-Based Spatial Clustering of Applications with Noise (DBSCAN) [32], Data Spectroscopic (DaSpec) [33] and Gaussian Mixture Model-BIC, as proposed in [34], as well as with Evolutionary k-prototype (EKP) [35], an unsupervised hybrid method that integrates evolutionary optimization with clustering categories. At second level, we tested the accuracy of OLYMPUS against the original FSTS and other hybrid approaches such as the

evolutionary fuzzy algorithm of Anand *et al.* [18] - called hereafter Anand - and the Improved Differential Evolutionary Fuzzy Clustering (IDEFC) [19].

3.1. Simulated datasets

Initially, we constructed 16 synthetic datasets consisting of 1000 time series profiles, each one having a pre-fixed dimension size, i.e. sampling resolution (5, 10, 15 and 20 dimensions), and number of optimal clusters (10, 15, 20 and 50) to be detected. In detail, for each of the 16 clustering scenarios, we aimed at constructing groups of time series with almost equal shapes, which in turn fulfill the corresponding optimal cluster number. For each group of time series, we initialize a profile with a random value *X(0)* with normal distribution in a user defined range [0, threshold]. The threshold varied between the groups, after repetitive trials, between [0, 12] with the scope to produce distinct clusters. Then, for each time-point $i \in \{1, \text{dimension size}\}$, we iteratively compute the values of *X(i)* based on random values for slope and offset (obtained by normal distribution too) according to their user-defined values (mean, standard deviation):

$$X(i) = X(i-1) * slope + offset \quad (13).$$

After exhaustive trials the slope was set 0.18 ± 0.18 (mean ± std) in order to achieve almost identical shapes, while the offset was set 0.5 ± 3.5 so as to resemble the abrupt up- and down-regulation patterns of dynamical processes.

3.2. Evaluation of cluster number prediction

As described above, DBSCAN and DaSpec are algorithms that can automatically appoint the number of clusters in a data set. DBSCAN algorithm is a well-known representative of the density-based class and DaSpec of the spectral clustering class. GMM is density-based model where k multivariate normal density components are combined by assuming that each component represents a cluster. In our analysis, we applied the GMM-BIC after executing GMM for a varying number of clusters and then selecting the best amongst them using BIC criterion. EKP combines evolutionary optimization strategy with K-prototype algorithm, a partition-based algorithm for mixed type datasets. All algorithms were run over 100 experiments with the corresponding optimal parameters which could be different than the default. Before testing OLYMPUS, we carried out repetitive trials in examining the parameter values and strategies that optimize its performance. In detail, regarding DE we employed the classic mutation scheme, named *DE/rand/1*. We set the population size of individuals $NP = 8$, DE constants $F = 0.5$, $C_R = 0.75$, $w = 1.4$, $h = 0.3$, $MaxIterDE = 15$ and $MaxIterFSTS = 30$. With respect to FSTS, we set fuzziness $w = 1.4$ and $h = 0.3$. The results of this experimental setting are exhibited in Table 1, in which we report the number of retrieved clusters per cluster number case and per dimension size.

We employed the Wilcoxon signed rank test (P-value < 0.01) after observing that the median differences between pairs of observations (between OLYMPUS and each

algorithm under investigation) follow non-normal distribution (one sample Kolmogorov-Smirnov test with P-value < 0.01). Each entry in the table displays the mean, median and the standard deviation of median over the 100 experiments. The sign next to each entry, denotes that OLYMPUS performed better and the difference was either significant (+) or not significant (=). As shown, OLYMPUS outperformed EKP, GMM-BIC and DBSCAN almost in all cases. With respect to DaSpec, OLYMPUS outperformed significantly in the dimension size less than 15 and in large number of clusters (equal to 50). The results indicate that DaSpec is not appropriate for short time series data, a typical case in microarray experiments, where the number of available time points is usually less than 10 [34].

3.3. Evaluation of Correct Classification Rate (CCR)

At second level, we evaluated the accuracy of OLYMPUS, based on the same set of 16 synthetic datasets, against the aforementioned unsupervised methods, FSTS and other hybrid evolutionary fuzzy clustering methods like Anand and IDEFC, which however cannot determine the correct cluster number *a priori*. It is worth mentioning that Anand assigns genes to model profiles based on a fuzzy membership function and selects model profiles using an evolutionary algorithm that finds trade-off between minimizing quantization errors and minimizing the number of profiles. IDEFC algorithm combines an improved version of differential evolutionary strategy with standard Fuzzy C-Means algorithm. All algorithms were run over 100 experiments after setting the cluster number as defined by the simulated datasets, whereas OLYMPUS estimated it automatically. The Pearson correlation coefficient was used for IDEFC and STS distance for FSTS and OLYMPUS respectively. In the case of Anand, Euclidean distance was used to calculate the quantization error. The clustering user-defined parameters for FSTS algorithm were the exponent of membership, w = 1.4 and the threshold of membership to form the clusters, h = 0.3. Regarding Anand and IDEFC, the parameters were assigned as proposed by the authors. All aforementioned algorithms, besides OLYMPUS, were re-run with different parameters, though no improvement was detected. In Table 2, the comparison of the four algorithms is reported with regard to Correct Classification Rate (CCR), which was computed as the ratio between the number of genes clustered into the true clusters and the total number of genes.

Similarly, we employed the Wilcoxon signed rank test (P-value < 0.01) after observing that the median differences between pairs of observations (between OLYMPUS and each algorithm under investigation) follow non-normal distribution (one sample Kolmogorov-Smirnov test with P-value < 0.01). The sign next to each entry, denotes that OLYMPUS performed better and the difference was either significant (+) or not significant (=). It is evident that OLYMPUS outperformed GMM-BIC, DBSCAN, EKP, IDEFC and Anand in almost all datasets. With regard to FSTS, the results were similar to OLYMPUS only when the dimension size was large (20 samples) and the number of clusters was small (10 and 15 clusters). This finding states clearly that the integrated framework, proposed by OLYMPUS, improves

significantly the clustering potential of the original FSTS and circumvents the limitations confronted in typical microarray experiments. With respect to DaSpec, OLYMPUS performed substantially better in the dimension size less than 15 and in large number of clusters (equal to 50).

4. Influenza A Kinetic Model

4.1. Gene expression data

We tested the ability of OLYMPUS in a time series gene expression dataset with the scope to present the kinetic model of host response mechanisms after infection with Influenza A (H1N1) and set a timeline regarding the activation of several processes as well as the recruitment of various signaling pathways and cell populations. As described in Pommerenke *et al.* [23] – called hereafter reference study, C57BL/6J mice were infected with a mouse-adapted influenza A virus (PR8). Three replicates, from three individually infected mice, were taken for each time point after infection (1, 2, 3, 5, 8, 10, 14, 18, 22, 26, 30, 40, 60 days) and nine replicates from three mock-infected mice (day 0). The complete dataset is accessible through ArrayExpress database under the accession number E-MTAB-764. In the reference study, density clustering was applied [36], which differentiates a region with higher density than its neighborhood based on KNN (k-nearest neighbor) algorithm. Nevertheless, this method is not suitable for unevenly sampled time series gene expression data, since it does not consider the time domain features. Also, the low clustering resolution (eight clusters were identified) failed to capture the expression profile 'signals' of genes which on one hand are few compared to the complete gene set but yet are distinct of significantly involved cellular processes.

In this work, we focused on the differentially expressed set of genes as defined in the reference study, in which 3,595 genes exhibited at least a two-fold change in expression levels compared to the control day and the fold-changes were significant with an False Discovery Rate (FDR) corrected p-value of 0.1 using the rank product method. On next level, we applied OLYMPUS at the dataset with the same parameter values used in simulation application. With regard to mutation scheme, we examined all four alternative strategies and observed that DE\rand\1 performed better in terms of minimizing the fitness function (Supplementary File 2 - Fig.2). The number of clusters was set to 25 as appointed by the minimum BIC.

The dendrogram (Supplementary File 2 - Fig. 3) displays the $\log_2$ fold changes of the cluster centroids profiles with respect to day 0. The centroids were grouped (only across rows) based on STS distance. In general, fifteen clusters displayed an increase in their expression levels compared to control, nine clusters displayed decrease and one cluster exhibited a unique temporal profile with exchangeable up- and down-regulation across all time points. After analyzing all clusters with the use of DAVID tool [37], we observed that the fifteen clusters are involved in acute innate and adaptive immune response processes, the nine clusters are characterized by

‘metabolic process’, ‘system development’, ‘lung alveolus development’ and ‘cell differentiation’ Gene Ontology (GO) terms and the last one remaining cluster is represented by ‘muscle system process’ and ‘tissue development’ GO terms (Supplementary Files 3 and 4).

4.2. Innate immune response

We investigated the kinetics of several regulatory pathways involved in the innate and adaptive host response mechanisms after viral infection (Fig. 2). One of the most important host responses to viral infections is the expression of interferons (IFNs). IFNs are generally divided into three types: type I (IFN-α and IFN-β), type II (IFN-γ) and type III (IFN-λ). Some representative example genes are *Ifnb1* (clusters 7, 13 and 17), *Il28b* (clusters 13 and 17) and *Ifng* (clusters 1, 7 and 17). These genes displayed increase in their fold change after day 1 p.i. with peak at day 3 or day 5 p.i. (an exception was found in the case of *Ifng* with peak at day 8 p.i., as defined by cluster 1). These observations are consistent with the idea that interferons are mainly produced by infected epithelial cells and activated Dendritic Cells (DCs) in the early phase, and that *Ifng* is mainly expressed in activated and infiltrating Natural Killer (NK) and T cells in the early and intermediate phase of an infection.

Upon infection by influenza, host cells detect viral RNA through pathogen sensors, such as RIG-I, and induce interferons (IFNs) and an antiviral program that is common to many RNA viruses. Members of the RIG-I pathogen sensory pathway are located in cluster 3, which is activated from day 2 to day 8 p.i. with clear peak at days 2, 3 and 5 p.i. (Supplementary File 2 – Fig. 4). Moving forward, the NOD-like receptors (NLRs) are a specialized group of intracellular receptors that represent a key component of the host innate immune system [38]. Also, in addition to their primary role in host defense against invading pathogens, they regulate nuclear factor-kappa B (NF-kappaB) signaling, interleukin-1-beta (IL-1beta) production and cell death. This pathway is depicted through clusters 13 and 17 (Supplementary File 2, Fig. 4 and Fig. 5), which display up-regulation with clear peak in days 2, 3 and 5 p.i. (cluster 13) and days 5 and 8 p.i. (cluster 17). Similar results were observed with regard to Toll-like receptor pathway (clusters 3, 7 and 17), which appears to play a central role in mediating both the antiviral and inflammatory responses of the innate immunity in combating viral infections [39].

In the early phase of the host response, many signaling pathways are activated that lead to the transcription of early response genes, mainly interferons, chemokines and cytokines. Chemokines and cytokines are induced in the infected tissue and stimulate resident macrophages and dendritic cells (DCs), leading to a continual chemokine/cytokine production, which in turn attracts infiltrating cells of the innate immune system such as macrophages, granulocytes, NK cells and DCs. The respective cytokine-cytokine and chemokine signaling pathways are represented by clusters 1, 7, 13 and 17 that display an increase in their fold change in the interval 2-8 day p.i., with the peak mainly in days 3 and 5 p.i. (Supplementary File 2, Fig. 4 and

Fig. 5). Similarly, the NK cell cytotoxicity is reflected by clusters 1, 7, 14 and 22, which exhibited increase in their fold change mainly in the range day 2 - day 10 p.i, two days later than the time point stated in the reference study (Fig. 2). Similarly, the Fcγ R-mediated phagocytosis pathway, which represents the destruction of pathogens by specialized cellular types such as macrophages, neutrophils and monocytes, was activated in the same interval as NK cell cytotoxicity.

4.3. Adaptive immunity

Further, we examined the timeline of adaptive immunity. In particular, we monitored the activation and infiltration of T cells in the lung by examining known T cell signature cells [40]. Example genes are *Cd3d*, *Lck*, *Itk*, *Cd3g*, *Lat*, *Rasgrp1*, *Scap1*, *Bcl11b*, *Cd6* and *Cd5*. Cytotoxic T cells are recruited in the adaptive immunity phase to destroy virally infected cells. Also the late study of [41] revealed, based on human nasal samples, a large increase in influenza-specific T cell responses by day 7, the time point where the virus was completely cleared and serum antibodies were still undetectable. Our results are similar with this observation, with the cluster 22 highly enriched in T cell signature cells. This cluster showed increase in the fold change starting from day 5 (two days later from the time point proposed in the reference study), with clear peak at day 8, followed by a decline up to day 18 p.i. (Supplementary File 2, Fig. 4 and Fig. 5). It should be noted that NK cell cytotoxicity shows considerable overlap with the T cells and therefore, the two pathways are not specific for one or the other cell population. However, our results reveal that NK cells have an earlier activation than the T cells, an observation that accords with the findings of the reference study.

Subsequently, we searched for the respective B cell signature cells, as assigned in [40]. Indeed, the B cell genes present in our gene set (example genes: *Cd19*, *Cd79b*, *Faim3*, *Cd22*, *Fcrla*, *Cd37*, *Spib*, *Tnfrsf13b*, *Prkcb*, *Bcl11a*, *Tnfrsf13c* and *Igh-6*) are activated in the late phase of the host response and are located in clusters 11, 22, but mainly in cluster 25. Clusters 11 and 22 showed an increase in their fold change in the interval day 5 - day 14 p.i. with the peak at day 8 p.i., an observation that accords with the results of the reference study, where the decline of B cells is consistent with the elimination of viral antigens. However, cluster 25 displayed up-regulation after day 14 p.i. and remained so until day 60 p.i. This cluster exhibits the efficacy of our method, since it refines the observations of the reference study, in the sense that we identified a subset of active antibody-depending mechanisms two months after the viral entrance (Fig. 2). In addition, similar results were found with regard to MHC class II pathway genes that are mainly located in cluster 16, which in turn is activated after day 5 p.i. with peak at day 8 p.i., followed by a decline until day 10 p.i.

4.4. Homeostasis

Next step in our analysis was to examine the fate of homeostatic related genes. Homeostatic pathways prevent inflammation in the first instance, limit its magnitude

once initiated and mediate resolution. This process is represented by cluster 18, which showed up-regulation after day 18 p.i. and remained so until day 60 p.i. This finding explicitly states that the homeostatic pathways are highly activated immediately after the elimination of the viral agent and the establishment of the adaptive immunity, in order to regain the original environment. Recently, homeostasis within the lung is described as an active process controlled by the association of the lung-resident cells with site-specific mechanisms; loss of epithelia integrity leads inevitably to loss of homeostasis and pulmonary inflammation [42]. However, our related genes showed no difference in fold change until day 18 p.i., i.e., indicating that homeostatic pathways switch on in the late phase of the host response (Supplementary File 2, Fig. 4 and Fig. 5). In parallel, the steady increase of fold change difference until day 60 p.i. accords with the observation that the lung does not return to its initial state of homeostasis but rather sets a new threshold of responsiveness to antigen, which is retained for a long time period and probably implicated in susceptibility to secondary bacterial co-infection [42].

4.5. Metabolism

Along with the analysis of the homeostatic pathways, we checked the behavior of metabolic pathway related genes. The respective genes are mainly found in six clusters (2, 8, 10, 15 and 20), all of which displayed decrease in their fold change mainly in the interval day 5- day 14 p.i. with the nadir at day 8 p.i. (Supplementary File 2, Fig. 4 and Fig. 5). The recent work of [43], in which metabolic profiling approaches were applied, focused on a narrow time interval (until 30 hours p.i.) and showed that metabolic imbalance is caused not by the viral replication but rather from the onset of the cell death caused by apoptosis. Our findings raise a hypothesis that warrants further study, by setting the significant breakdown of the metabolism on a more ‘macroscopic’ level than the level of hours, an aspect recently revealed through long term time series transcriptome data.

4.6. Cell Cycle

Further, we searched for clusters enriched in cell cycle related genes and clusters 5 and 11 fulfilled this purpose. These clusters displayed increase in their fold change in the time interval day 3 - day 14 p.i., with peaks at day 8 and 10 p.i. (Supplementary File 2, Fig. 4 and Fig. 5). Late studies [44-45] showed that the host response to severe influenza infection is followed by increased activation the cell cycle pathways with significant up-regulation of the key cell cycle encoding proteins. An alternative explanation, based on the fact that the cell cycle time interval coincides with that of the T cells, is that the increased cell cycle activity represents the recruitment of premature and the subsequent expansion of immune cells as a central event, i.e. the clonal expansion of virus-specific T cells (Fig. 2).

4.7. Repair Processes

Finally, we explored the dynamics of the repair processes and clusters 4, 6, 9, 19, 20 and 21 showed significant over-representation in ‘system development’, ‘tissue development’, ‘cell differentiation’ and ‘lung alveolus development’ GO terms. These clusters exhibited decrease in their fold change mainly in the time interval day 3- day 14 p.i. (nadir at day 8 p.i.) and then restored to the mock values (Supplementary File 2, Fig. 4 and Fig. 5). Unlike the results of the reference study, in which these processes get activated after day 30 p.i., we reveal the significant suppression of these processes at the early time points and their restoration to mock values after day 14 p.i. Also, it should be mentioned that cluster 4 displayed a unique temporal profile with exchangeable increase and decrease in fold change throughout the whole time period. It is enriched in genes related to ‘muscle system process’ GO term and probably reflects processes related to tracheal epithelium; the respective expression profiles could be an artifact caused, during experimental procedure, by the extraction of some trachea tissue along with the lung extraction.

In conclusion, the discriminative power of our clustering technique managed to reveal the dynamics of several pathways and physiological processes, which were implicated to be affected after the infection, yet no timeline was assigned to them. Our findings corroborate, complement and fine-tune the so far perspective of the temporal kinetic model of Influenza A host response.

Acknowledgements

This research has been co-financed by the European Union (European Social Fund – ESF) and Greek national funds through the Operational Program "Education and Lifelong Learning" of the National Strategic Reference Framework (NSRF) - Research Funding Program: Thalis. Investing in knowledge society through the European Social Fund.

Conflict of Interest

None declared.

References

1. Dimitrakopoulou K, Tsimpouris C, Papadopoulos G, Pommerenke C, Wilk E, Sgarbas KN, Schughart K, Bezerianos A. **Dynamic gene network reconstruction from gene expression data in mice after influenza A (H1N1) infection.** *J. Clin. Bioinforma.* 2011; 1:27.
2. Mavroudi S, Papadimitriou S, Bezerianos A. **Gene expression data analysis with a dynamically extended self-organized map that exploits class information.** *Bioinformatics*. 2002; 18:1446-1453.

3. Magni P, Ferrazzi F, Sacchi L, Bellazzi R. **TimeClust: a clustering tool for gene expression time series.** *Bioinformatics*. 2008; 24:430-2.
4. Smith AA, Vollrath A, Bradfield CA, Craven M. **Clustered alignments of gene-expression time series data.** *Bioinformatics*. 2009; 25:i119-27.
5. Sivriver J, Habib N, Friedman N. **An integrative clustering and modeling algorithm for dynamical gene expression data.** *Bioinformatics*. 2011; 27:i392-i400.
6. Fujita A, Severino P, Kojima K, Sato JR, Patriota AG, Miyano S. **Functional clustering of time series gene expression data by Granger causality.** *BMC Syst. Biol.* 2012; 6:137.
7. Ramoni MF, Sebastiani P, Kohane IS. **Cluster analysis of gene expression dynamics.** *PNAS*. 2002; 99:9121-9126.
8. Bar-Joseph Z, Gerber GK, Gifford DK, Jaakkola TS, Simon I. **Continuous representations of time series gene expression data.** *J. Comput. Biol*. 2003; 3:341-356.
9. Luan Y, Li H. **Model-based methods for identifying periodically expressed genes based on time course microarray gene expression data.** *Bioinformatics*. 2004; 20:332-339.
10. Song JJ, Lee HJ, Morris JS, Kang S. **Clustering of time-course gene expression data using functional data analysis.** *Comput. Biol. Chem*. 2007; 31:265-274.
11. Nascimento M, Sáfadi T, Fonseca e Silva F, Nascimento AC. **Bayesian model-based clustering of temporal gene expression using autoregressive panel data approach.** *Bioinformatics.* 2012; 28:2004-7.
12. Hruschka ER, de Castro LN, Campello RJGB. **Evolutionary Algorithms for Clustering Gene-Expression Data.** *ICDM* 2004, pp. 403-406.
13. Hruschka ER, Campello RJGB, Freitas AA, de Carvalho ACPLF. **A Survey of Evolutionary Algorithms for Clustering.** *IEEE Transactions on Systems, Man, and Cybernetics*. 2009; 39:133-155.
14. Lu Y, Lu S, Fotouhi F, Deng Y, Brown SJ. **Incremental genetic kmeans algorithm and its application in gene expression data analysis.** *BMC Bioinformatics*. 2004; 5:172.
15. Horta D, Naldi M, Campello RJGB, Hruschka ER, de Carvalho ACPLF. **Evolutionary Fuzzy Clustering: An Overview and Efficiency Issues.** *Studies in Computational Intelligence.* 2009; 204:167-195.
16. Gong W, Cai Z, Ling CX, Du J. **Hybrid Differential Evolution based on Fuzzy C-means Clustering.** *GECCO*. 2009; 7:523-530.
17. Park HS, Yoo SH, Cho SB. **Evolutionary Fuzzy Clustering Algorithm with Knowledge-Based Evaluation and Applications for Gene Expression Profiling.** *J. Comput. Theor. Nanosci.* 2005; 2:1-10.
18. Anand A, Suganthan PN, Deb K. **A novel fuzzy and multiobjective evolutionary algorithm based gene assignment for clustering short time series expression data.** In *IEEE Congress on Evolutionary Computation.* 2007; pp. 297-304.

19. Saha I, Plewczynski D, Maulik U, Bandyopadhyay S. **Improved differential evolution for microarray analysis.** *Int. J. Data Min. Bioinform.* 2012; 6:86-103.
20. Dalton L, Ballarin V, Brun M. **Clustering algorithms: on learning, validation, performance, and applications to genomics.** *Curr. Genomics.* 2009; 10:430-45.
21. Fauci AS, Challberg MD. **Host-based antipoxvirus therapeutic strategies: turning the tables.** *J. Clin. Invest.* 2005; 115:231-3.
22. Pawelek KA, Huynh GT, Quinlivan M, Cullinane A, Rong L, Perelson AS. **Modeling within-host dynamics of influenza virus infection including immune responses.** *PLoS Comput. Biol.* 2012; 8: e1002588.
23. Pommerenke C, Wilk E, Srivastava B, Schulze A, Novoselova N, Geffers R, Schughart K. **Global transcriptome analysis in influenza-infected mouse lungs reveals the kinetics of innate and adaptive host immune responses.** *PLoS One.* 2012; 7:e41169.
24. Schwarz G. **Estimating the dimension of a model.** *Ann. Stat.* 1978; 6:461–464.
25. Möller-Levet CS, Klawonnb F, Choc KH, Yina H, Wolkenhauere O. **Clustering of unevenly sampled gene expression time-series data.** *Fuzzy Sets In Bioinformatics.* 2005; 152:49-66.
26. Storn R, Price K. **Differential evolution - a simple and efficient adaptive scheme for global optimization over continuous spaces.** *Journal of Global Optimization*. 1997; 11:341-359.
27. Swagatam D, Suganthan PN. **Differential evolution: A survey of the state-of-the-art.** *IEEE Transactions on Evolutionary Computation.* 2010; 15:4-31.
28. Epitropakis, MG, Tasoulis DK, Pavlidis NG, Plagianakos VP, Vrahatis MN. **Enhancing differential evolution utilizing proximity-based mutation operators.** *IEEE Transactions on Evolutionary Computation*. 2011; 15:99-119.
29. Zhao Q, Mantau X, Franti P. **Knee Point Detection in BIC for Detecting the Number of Clusters**. 20th *IEEE International Conference on Tools with Artificial Intelligence (ICTAI '08)*. 2008; 2:431-438.
30. Sheng W, Swift S, Zhang L, Liu X. **A Weighted Sum Validity Function for clustering with a Hybrid Niching Genetic Algorithm.** *IEEE Transactions on Systems, Man, and Cybernetics*. 2005; 35:1156-67.
31. Pal NR, Bezdek JC. **On cluster validity for the fuzzy c-means model.** *IEEE Transactions on Fuzzy Systems*. 1995; 3:370-379.
32. Sander J, Ester M, Kriegel HP, Xu X. **Density-based clustering in spatial databases: the algorithm GDBSCAN and its applications.** *Data Mining and Knowledge Discovery.* 1998; 2:169–194.
33. Shi T, Belkin M, Yu B. **Data spectroscopy: Eigenspaces of convolution operators and clustering**. *Annals of Statistics.* 2009; 37:3960-3984.
34. Tasoulis SK, Tasoulis DK, Plagianakos VP. **Enhancing principal direction divisive clustering.** *Pattern Recognition* 2010; 43:3391-3411.

35. Zheng Z, Gong M, Ma J, Jiao L, Wu Q**. Unsupervised evolutionary clustering algorithm for mixed type data.** *IEEE Congress on Evolutionary Computation (CEC)*, 18-23 July 2010; 1-8.
36. Chung S, Jun J, McLeod D. **Mining gene expression datasets using density-based clustering.** *Proceedings of the 13th ACM international conference on Information and knowledge management, New York, USA.* 2004; pp. 150-151.
37. Huang da W, Sherman BT, Lempicki RA. **Systematic and integrative analysis of large gene lists using DAVID Bioinformatics Resources.** *Nature Protoc*. 2009; 4:44-57.
38. Chen G, Shaw MH, Kim YG, Nuñez G. NOD**-like receptors: role in innate immunity and inflammatory disease.** *Annu. Rev. Pathol.* 2009; 4:365-98.
39. Wong JP, Christopher ME, Viswanathan S, Karpoff N, Dai X, Das D, Sun LQ, Wang M, Salazar AM. **Activation of toll-like receptor signaling pathway for protection against influenza virus infection.** *Vaccine*. 2009; 27:3481-3483.
40. Alberts R, Lu L, Williams RW, Schughart K. **Genome-wide analysis of the mouse lung transcriptome reveals novel molecular gene interaction networks and cell-specific expression signatures.** *Respir. Res.* 2011; 12:61.
41. Wilkinson TM, Li CK, Chui CS, Huang AK, Perkins M, Liebner JC, Lambkin-Williams R, Gilbert A, Oxford J, Nicholas B, Staples KJ, Dong T, Douek DC, McMichael AJ, Xu XN. **Preexisting influenza-specific CD4+ T cells correlate with disease protection against influenza challenge in humans.** *Nat. Med.* 2012; 18:274-80.
42. Snelgrove RJ, Godlee A, Hussell T. **Airway immune homeostasis and implications for influenza-induced inflammation.** *Trends Immunol.* 2011; 32:328-34.
43. Ritter JB, Wahl AS, Freund S, Genzel Y, Reichl U. **Metabolic effects of influenza virus infection in cultured animal cells: Intra- and extracellular metabolite profiling.** *BMC Syst. Biol.* 2010; 4:61.
44. He Y, Xu K, Keiner B, Zhou J, Czudai V, Li T, Chen Z, Liu J, Klenk HD, Shu YL, Sun B. **Influenza A virus replication induces cell cycle arrest in G0/G1 phase.** *J. Virol.* 2010; 84:12832-40.
45. Parnell G, McLean A, Booth D, Huang S, Nalos M, Tang B. **Aberrant cell cycle and apoptotic changes characterise severe influenza A infection--a meta-analysis of genomic signatures in circulating leukocytes.** *PLoS One.* 2011; 6:e17186.

Table 1. Mean, median and standard deviation of the observed cluster number.

	10 Clusters				15 Clusters				20 Clusters				50 Clusters			
	Mean	**Median**	**Std**		**Mean**	**Median**	**Std**		**Mean**	**Median**	**Std**		**Mean**	**Median**	**Std**	
Dimensions 5																
OLYMPUS	10.08	10	0.56		16.48	16	0.41		20.00	20	0.00		45.95	46	0.22	
DASPEC	8.70	9	0.97	+	10.90	11	2.03	+	4.05	4	1.29	+	1.12	1	0.32	+
DBSCAN	7.58	8	1.32	+	9.10	9	2.07	+	9.92	10	2.51	+	9.00	9	3.01	+
GMM-BIC	8.50	8	0.52	+	12.15	12	1.03	+	15.65	17	2.72	+	7.5	8	2.00	+
EKP	7.60	7	2.04	+	10.80	10	1.06	+	17.90	9	2.20	+	30.80	30	2.40	+
Dimensions 10																
OLYMPUS	10.02	10	0.14		15.20	15	0.20		20.26	20	0.52		49.75	50	2.40	
DASPEC	9.10	9	0.00	+	14.37	14	0.17	+	15.95	16	1.94	+	6.05	1	2.11	+
DBSCAN	8.97	9	0.98	+	11.85	12	1.67	+	14.05	14	2.26	+	21.67	21	4.66	+
GMM-BIC	11.8	12	1.20	+	18.05	18	0.45	+	24.5	24	3.20	+	61.20	60	9.70	+
EKP	8.60	9	1.00	+	13.00	13	0.00	+	18.00	18	1.20	+	36.10	36	3.40	+
Dimensions 15																
OLYMPUS	10.00	10	0.00		15.00	15	0.00		20.10	20	0.30		51.95	51	0.94	
DASPEC	10.00	10	0.00	=	15.00	15	0.00	=	19.55	20	0.62	=	24.75	6	4.14	+
DBSCAN	9.30	9	0.73	+	13.53	14	1.23	+	16.91	17	1.83	+	31.50	31	4.92	+
GMM-BIC	10.90	11	0.55	+	17.20	17	1.78	+	23.10	22	1.50	+	33.00	33	0.00	+
EKP	8.80	9	0.20	+	13.50	13	0.05	+	18.50	18	1.00	+	39.30	40	4.50	+
Dimensions 20																
OLYMPUS	10.00	10	0.00		15.00	15	0.00		20.00	20	0.00		53.30	53	2.45	
DASPEC	10.30	10	0.32	=	15.00	15	0.00	=	19.97	20	0.17	=	24.49	24	4.05	+
DBSCAN	9.46	9	0.29	+	14.26	14	0.90	+	18.63	18	1.12	+	39.39	40	3.74	+
GMM-BIC	4.50	5	0.80	+	6.52	6	0.61	+	8.90	9	0.80	+	12.50	13	0.50	+
EKP	7.80	8	0.70	+	12.20	12	0.07	+	16.50	16	1.50	+	31.30	35	9.50	+

Mean, median and standard deviation of the observed cluster number, between OLYMPUS, DaSpec, DBSCAN, GMM-BIC and EKP, over 100 experiments. The sign next to each entry, denotes that OLYMPUS performed better and the difference was either significant (+) (Wilcoxon rank sum test, P-value < 0.01) or not significant (=).

Table 2. Mean, median and standard deviation of the observed CCR value.

	10 Clusters				15 Clusters				20 Clusters				50 Clusters			
	Mean	**Median**	**Std**		**Mean**	**Median**	**Std**		**Mean**	**Median**	**Std**		**Mean**	**Median**	**Std**	
Dimensions 5																
OLYMPUS	0.91	0.91	0.05		0.88	0.88	0.04		0.85	0.85	0.05		0.78	0.78	0.03	
FSTS	0.85	0.86	0.06	+	0.80	0.81	0.05	+	0.79	0.78	0.04	+	0.70	0.70	0.02	+
DASPEC	0.80	0.80	0.00	+	0.66	0.65	0.02	+	0.25	0.25	0.00	+	0.02	0.02	0.00	+
ANAND	0.79	0.80	0.08	+	0.78	0.77	0.06	+	0.75	0.73	0.04	+	0.70	0.71	0.05	+
DBSCAN	0.80	0.80	0.00	+	0.53	0.55	0.03	+	0.45	0.45	0.00	+	0.22	0.22	0.01	+
IDEFC	0.78	0.78	0.06	+	0.79	0.78	0.09	+	0.77	0.78	0.05	+	0.73	0.74	0.08	+
EKP	0.84	0.84	0.03	+	0.81	0.82	0.04	+	0.74	0.74	0.00	+	0.61	0.60	0.06	+
GMM-BIC	0.86	0.87	0.02	+	0.79	0.80	0.02	+	0.69	0.70	0.03	+	0.25	0.27	0.05	+
Dimensions 10																
OLYMPUS	0.95	0.95	0.03		0.94	0.94	0.02		0.92	0.92	0.03		0.87	0.88	0.03	
FSTS	0.92	0.93	0.02	+	0.92	0.92	0.02	+	0.90	0.90	0.02	+	0.81	0.80	0.03	+
DASPEC	0.90	0.90	0.00	+	0.79	0.80	0.03	+	0.78	0.79	0.02	+	0.02	0.02	0.00	+
ANAND	0.84	0.85	0.06	+	0.85	0.86	0.03	+	0.83	0.82	0.04	+	0.80	0.89	0.02	+
DBSCAN	0.80	0.81	0.01	+	0.65	0.60	0.05	+	0.75	0.72	0.06	+	0.42	0.40	0.03	+
IDEFC	0.91	0.90	0.05	+	0.91	0.91	0.04	+	0.87	0.87	0.04	+	0.84	0.84	0.05	+
EKP	0.82	0.83	0.04	+	0.83	0.83	0.01	+	0.81	0.81	0.01	+	0.78	0.80	0.02	+
GMM-BIC	0.90	0.91	0.02	+	0.92	0.92	0.03	+	0.89	0.90	0.04	+	0.85	0.84	0.90	+
Dimensions 15																
OLYMPUS	0.95	0.96	0.03		0.95	0.95	0.02		0.94	0.94	0.03		0.89	0.90	0.01	
FSTS	0.93	0.94	0.02	+	0.92	0.93	0.02	+	0.92	0.92	0.04	+	0.87	0.88	0.04	+
DASPEC	0.90	0.90	0.02	+	0.94	0.94	0.04	=	0.93	0.94	0.02	=	0.34	0.36	0.08	+
ANAND	0.90	0.91	0.02	+	0.90	0.90	0.03	+	0.88	0.89	0.03	+	0.86	0.87	0.04	+
DBSCAN	0.80	0.82	0.02	+	0.83	0.85	0.04	+	0.82	0.80	0.06	+	0.43	0.40	0.03	+
IDEFC	0.92	0.92	0.01	+	0.90	0.92	0.03	+	0.90	0.91	0.03	+	0.85	0.85	0.04	+
EKP	0.89	0.89	0.00	+	0.92	0.91	0.01	+	0.89	0.90	0.03	+	0.85	0.84	0.02	+
GMM-BIC	0.80	0.80	0.01	+	0.75	0.77	0.06	+	0.75	0.75	0.01	+	0.73	0.75	0.09	+
Dimensions 20																
OLYMPUS	0.96	0.97	0.03		0.95	0.95	0.03		0.93	0.93	0.03		0.93	0.94	0.03	
FSTS	0.96	0.96	0.02	=	0.94	0.95	0.03	=	0.88	0.88	0.04	+	0.91	0.92	0.03	+
DASPEC	0.92	0.92	0.03	+	0.94	0.95	0.04	=	0.92	0.93	0.01	=	0.52	0.50	0.06	+
ANAND	0.93	0.94	0.02	+	0.92	0.92	0.04	+	0.89	0.89	0.02	+	0.89	0.89	0.04	+
DBSCAN	0.83	0.80	0.04	+	0.80	0.81	0.02	+	0.90	0.89	0.01	+	0.82	0.85	0.06	+
IDEFC	0.91	0.92	0.08	+	0.90	0.91	0.05	+	0.87	0.87	0.03	+	0.89	0.90	0.05	+

EKP	0.90	0.91	0.02	+	0.91	0.92	0.02	+	0.89	0.88	0.03	+	0.86	0.87	0.04	+
GMM-BIC	0.35	0.36	0.04	+	0.30	0.31	0.03	+	0.35	0.36	0.02	+	0.27	0.28	0.03	+

Mean, median and standard deviation of the observed CCR value, between OLYMPUS, DaSpec, DBSCAN, GMM-BIC, EKP, FSTS, Anand and IDEFC, over 100 experiments. The sign next to each entry, denotes that OLYMPUS performed better and the difference was either significant (+) (Wilcoxon rank sum test, P-value < 0.01) or not significant (=).

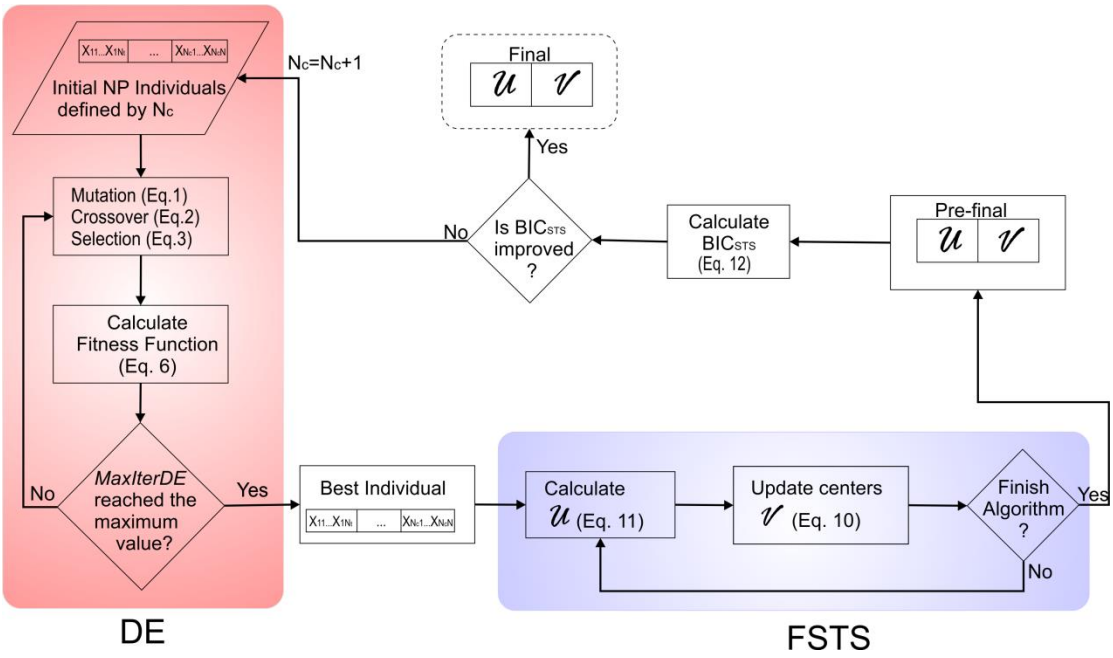

Fig. 1. Flowchart of OLYMPUS algorithm. It is an unsupervised hybrid approach that integrates algorithms from two major machine learning categories, namely evolutionary optimization and clustering. In detail, we integrated Differential Evolutionary (DE) algorithm into Fuzzy Short Time Series (FSTS) clustering method with the scope to utilize efficiently the information of population of the first and enhance the performance of the latter. On first level, a number of individuals are randomly generated, where each individual represents a different set of cluster centers. Then, the processes of mutation, crossover, and selection are executed for a user-defined number of iterations (*MaxIterDE*). Through the evolutionary process, individuals get better positions in search space, and after *MaxIterDE* iterations we obtain the best individual, i.e. the best set of cluster centers. On second level, the best individual is introduced to FSTS and this initialization helps FSTS to avoid being trapped in local minima. Next, FSTS method runs for a user-defined number of iterations (*MaxIterFSTS*), in which the cluster centers and the membership matrix are updated. Finally, the DE and FSTS steps are repeated for different numbers of clusters, ranging in $\{2, \sqrt{n_g}\}$ and the final number of clusters is set in the case where the BIC index reaches its minimal value.

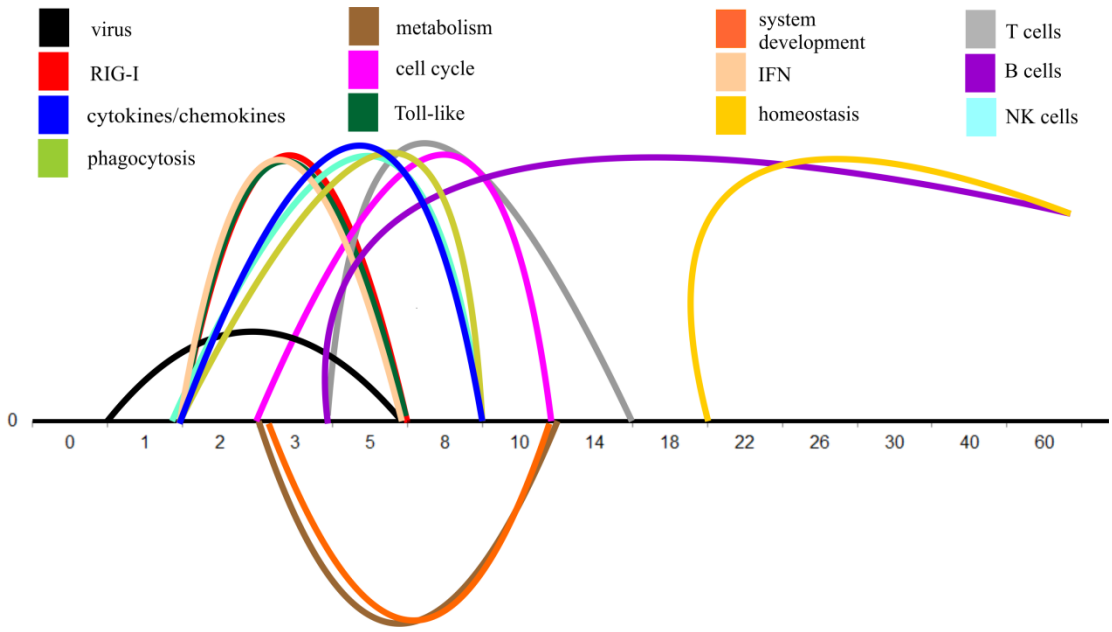

Fig. 2. Schematic representation of the kinetics of several distinct pathways, biological processes and cellular types in the host lung after the Influenza A infection. Curves above the axis of time correspond to activation, while the opposite stands for suppression. The beginning and end of each colored curve denotes the kinetics from the phase of activation followed by the respective decline (the opposite stands for suppression). It should be noted that the curve of virus is borrowed by (Pommerenke *et al.* [23]) and indicates the time period for the viral clearance. The curve of B cells was active until day 60 p.i., indicating so that certain antibody-related mechanisms decline at a later, currently un-recorded, time point.

SUPPLEMENTARY FILES

File 1. Matlab codes for implementing OLYMPUS.

File 2. Supplementary Figures 1, 2, 3, 4 and 5.

File 3. Table: Members per cluster named with the official gene names. Every column corresponds to a cluster.

File 4. Table: GO and KEGG pathway analysis of the identified 25 clusters based on DAVID tool. The mouse whole-genome was set as the population background in enrichment analysis. Count: observed number of genes; P-value: Modified Fisher exact P-value, EASE score; %: percentage of genes relative to the size of the whole cluster (annotated and un-annotated genes); nd: not determined in terms of enriched KEGG pathway terms.