Identifying Functional Families of Trajectories in Biological Pathways by Soft Clustering: Application to TGF-β Signaling

Jean Coquet, Nathalie Théret, Vincent Legagneux, Olivier Dameron

Identifying Functional Families of Trajectories in Biological Pathways by Soft Clustering: Application to TGF-β Signaling

Jean Coquet[1,2], Nathalie Theret[1,2], Vincent Legagneux[2], and Olivier Dameron[1]

[1] Université de Rennes 1 - IRISA/INRIA, UMR6074, 263 avenue du Général Leclerc, 35042 Rennes, Cedex, France

[2] INSERM U1085 IRSET, Université de Rennes 1, 2 avenue Pr Léon Bernard, 35043 Rennes, Cedex, France

Abstract. The study of complex biological processes requires to forgo simplified models for extensive ones. Yet, these models' size and complexity place them beyond understanding. Their analysis requires new methods for identifying general patterns. The Transforming Growth Factor TGF-β is a multifunctional cytokine that regulates mammalian cell development, differentiation, and homeostasis. Depending on the context, it can play the antagonistic roles of growth inhibitor or of tumor promoter. Its context-dependent pleiotropic nature is associated with complex signaling pathways. The most comprehensive model of TGF-β-dependent signaling is composed of 15,934 chains of reactions (trajectories) linking TGF-β to at least one of its 159 target genes. Identifying functional patterns in such a network requires new automated methods.

This article presents a framework for identifying groups of similar trajectories composed of the same molecules using an exhaustive and without prior assumptions approach. First, the trajectories were clustered using the Relevant Set Correlation model, a shared nearest-neighbors clustering method. Five groups of trajectories were identified. Second, for each cluster the over-represented molecules were determined by scoring the frequency of each molecule implicated in trajectories. Third, Gene set enrichment analysis on the clusters of trajectories revealed some specific TGF-β-dependent biological processes, with different clusters associated to the antagonists roles of TGF-β. This confirms that our approach yields biologically-relevant results. We developed a web interface that facilitates graph visualization and analysis.

Our clustering-based method is suitable for identifying families of functionally-similar trajectories in the TGF-β signaling network. It can be generalized to explore any large-scale biological pathways.

Keywords: TGF-β, Signaling pathways, Discrete dynamic model, Soft clustering, RSC model

1 Introduction

Living cells use molecular signaling networks to adapt their phenotype to the microenvironment modifications. In order to decipher the dynamic of signaling pathways, mathematical models have been developed using different strategies [10,4]. Differential equation-based models are limited to small networks due to the explosion in the number of variables in complex networks and the lack of known quantitative values for the parameters [1]. Qualitative modeling approaches based on events discretization have been successfully applied to large networks. In qualitative models, signaling networks are represented as a graph where each node (genes or proteins) is represented by a finite-state variable and edges describe interactions between biomolecules as rules [17]. Such models proved to be suitable for describing the qualitative nature of biological information whithin large and complex signaling pathways [19].

Signaling by the polypeptide Transforming Growth Factor TGF-β is one of the most intriguing signaling networks that govern complex multifunctional profiles. TGF-β was first described as a potent growth inhibitor for a wide variety of cells. It affects apoptosis and differentiation thereby controlling tissue homeostasis [7]. At the opposite, upregulation and activation of TGF-β has been linked to various diseases, including fibrosis and cancer through promotion of cell proliferation and invasion [24]. The pleiotropic effects of TGF-β are associated to the diversity of signaling pathways that depend on the biological context [13]. TGF-β binding to the receptor complex induces the phosphorylation of intracellular substrates, R-Smad proteins which hetero-dimerize with Smad4. The Smad complexes move into the nucleus where they regulate the transcription of TGF-β-target genes. Alternatively, non-Smad pathways are activated by ligand-occupied receptor to modulate downstream cellular responses [14]. These non-Smad pathways include mitogen-activated protein kinase (MAPK) such as p38 and Jun N-terminal kinase (JNK) pathways, Rho-like GTPase signaling pathways, and phosphatidylinositol-3-kinase/protein kinase B (PKB/AKT) pathways. Combinations of Smad and non-Smad pathways contribute to the high heterogeneity of cell responses to TGF-β. Additionally, many molecules from these pathways are involved in other signaling pathways activated by other microenvironment inputs, which leads to complex crosstalks [12].

Numerical approaches using differential models have been developed to describe the behavior of TGF-β canonical pathway involving Smad proteins [27]. Because of the numerous components and the lack of quantitative data, the non canonical pathways have never been included in these TGF-β models. To solve this problem, Andrieux et al. recently developed a qualitative discrete formalism compatible with large-scale discrete models [2]. The Cadbiom language is a state-transition formalism based on a simplified version of guarded transition [16]. It allows a fine-grained description of the system's dynamic behavior by introducing temporal parameters to manage competition and cooperation between parts of the models (http://cadbiom.genouest.org). Based on the Cadbiom formalism, Andrieux et al. integrated the 137 signaling pathways from the Pathway Interaction Database (PID) [20] and derived an exhaustive TGF-β signaling network

that includes canonical and non-Smad pathways [2]. Using this model they identified 15,934 signaling trajectories regulating 145 TGF-β target genes and found specific signatures for activating TGF-β-dependent genes.

Characterizing these 15,934 signaling trajectories remains a challenging task. They are mainly composed of signaling molecules whose modularity and combination are the base of cell response plasticity and adaptability [9,15,21]. We developed a methodological approach to identify families of trajectories with functional biological signature based on their signaling molecules content. The major difficulty were the inner complexity of the networks, and the fact that some molecules may be involved in multiple families, as suggested by TGF-β's context-dependent roles. To address these challenges, we used an unsupervised soft-clustering method to compare signaling trajectories according to their molecular composition. The clusters correspond to families of trajectories, and can share common molecules. Our analysis does not rely on a priori knowledge on the number of clusters nor on the membership of a molecule to a cluster. Based on this approach, we identified five groups of signaling trajectories. Importantly we further show that these five groups are associated with specific biological functions thereby demonstrating the relevance of soft clustering to decipher cell signaling networks.

2 Materials and Methods

Cellular signaling pathways are chains of biochemical reactions. Typically, they encompass the interaction of signaling molecules such as growth factors with receptors at the cell surface, the transmission of signal through signaling cascades involving many molecules such as kinases and finally the molecular networks involved in regulation of target gene transcription within the nucleus. In order to decipher the complexity of signaling TGF-β-dependent networks and for characterizing these trajectories, we focus on the proteins involved in the reactions (reactants, products and catalyzers). Note that a gene can encode for a protein implicated elsewhere in the pathway, so proteins and genes form non-disjoint sets.

The trajectories are first submitted to a pre-processing step to generate a non redundant set of signaling trajectories. The second step groups similar trajectories using soft clustering. The third step characterizes the specificity of groups of trajectories by determining the over-represented proteins and their biological function using semantic annotations.

2.1 Available data & Pre-processing

The original data-set contained the 15,934 signaling trajectories involved in the regulation of 145 TGF-β-dependent genes as previously described in [2]. A signaling trajectory is defined as a set of molecules required for activation of TGF-β-dependent genes (fig. 1A). Each original trajectory T_k was composed of TGF-β, signaling molecules and a single target gene (fig. 1B). There were 321 signaling

molecules (identified by their uniprot ID) involved in at least one of the 15,934 signaling trajectories. To compare the trajectories based on their molecule composition, we first discarded TGF-β which was belonging to all the trajectories. Next we observed that several trajectories were composed of the same signaling molecules but differed only by the target genes. We decided to discard the target genes from the trajectories, and to represent separately the associations between trajectories and target genes (fig. 1C). The motivation was (i) to avoid the artificial duplication of trajectories, and (ii) to have a model that represents explicitly the fact that a single chain of reactions can influence several genes. In the remainder of the article, the pre-processed trajectories are noted t_k and their set is noted S.

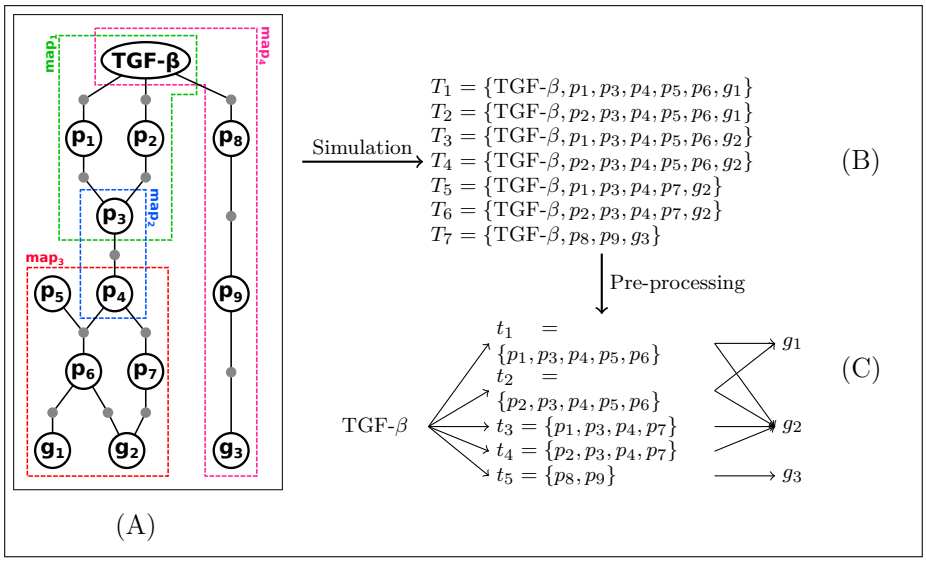

Fig. 1: *Example of the generation of trajectories from PID maps and their pre-processing. (A) The signaling network made of 4 maps and is composed of proteins, TGF-β and genes. (B) Trajectories are defined by a set of proteins containing TGF-β, signaling proteins (p_i) and target genes (g_i). (C) Pre-processed trajectories are restricted to signaling proteins. After pre-processing, the trajectories T_1 and T_3 are represented by the trajectory t_1; T_2 and T_4 are represented by t_2; T_5 is represented by t_3; T_6 is represented by t_4; T_7 is respresented by t_5.*

2.2 Clustering Method

We used the Relevant Set Correlation (RSC) model to identify clusters of trajectories [6]. This model uses as input a function $Q(t)$ that returns for every trajectory $t \in S$ a list of all the other trajectories in S sorted by their decreasing correlation with t.

$Q(t)$ function for ranking the trajectories by decreasing correlation to t A trajectory $t_i \in S$ is represented by a binary vector v_i whose dimension is equal to the number of all proteins. The coordinate value of "1" indicates that the trajectory contains the protein, and the coordinate value of "0" indicates that the trajectory does not (see table 1).

	p_1	p_2	p_3	p_4	p_5	p_6	p_7	p_8	p_9
t_1	1	0	1	1	1	1	0	0	0
t_2	0	1	1	1	1	1	0	0	0
t_3	1	0	1	1	0	0	1	0	0
t_4	0	1	1	1	0	0	1	0	0
t_5	0	0	0	0	0	0	0	1	1

Table 1: *Example of binary matrix representing the protein composition of trajectories. If a protein p_j is present in a trajectory t_i then the cell (i, j) is "1" else "0".*

Based on the binary vectors, we apply the Pearson correlation formula and construct a similarity matrix (see table 2) :

$$r(t_i, t_j) = \frac{\sum_{k=1}^{n}(t_{i,k} - \overline{t_i})(t_{j,k} - \overline{t_j})}{\sqrt{\sum_{k=1}^{n}(t_{i,k} - \overline{t_i})^2 \sum_{k=1}^{n}(t_{j,k} - \overline{t_j})^2}} \tag{1}$$

where $(t_{i,1}, t_{i,2}, ..., t_{i,n})$ and $(t_{j,1}, t_{j,2}, ..., t_{j,n})$ are the vectors of trajectories t_i and t_j with $\overline{t_i}$ and $\overline{t_j}$ their respective average.

	t_1	t_2	t_3	t_4	t_5
t_1	1.000	0.550	0.350	-0.100	-0.598
t_2	0.550	1.000	-0.100	0.350	-0.598
t_3	0.350	-0.100	1.000	0.550	-0.478
t_4	-0.100	0.350	0.550	1.000	-0.478
t_5	-0.598	-0.598	-0.478	-0.478	1.000

Table 2: *Example of correlation matrix of trajectories t_i obtained from the trajectories' composition of table 1. If two trajectories t_i, t_j have exactly the same proteins the value of the cell (i, j) is* 1.0*. If the trajectories do not share any proteins the value is* 0.0*.*

For each trajectory $t_k \in S$, the Pearson correlation gives a partial ordering $< t_i >_{i=1}^{|S|}$ of trajectories where $i < j$ implies that $r(t_k, t_i) \geq r(t_k, t_j)$ (see

table 3). If two trajectories have the same correlation score, they are sorted alphabetically. We define the $Q(t)$ function as follows:

$$Q(t_k) =< t_i >_{i=1}^{|S|} \qquad \forall (i,j) \in [1,|S|]^2, i < j \Rightarrow r(t_k,t_i) \geq r(t_k,t_j) \tag{2}$$

	Q				
	1	2	3	4	5
t_1	t_1	t_2	t_3	t_4	t_5
t_2	t_2	t_1	t_4	t_3	t_5
t_3	t_3	t_4	t_1	t_2	t_5
t_4	t_4	t_3	t_2	t_1	t_5
t_5	t_5	t_3	t_4	t_1	t_2

Table 3: *Example of partial ordering of all trajectories for every trajectory t_i. All trajectories are sorted for each trajectory t_k in function of their Pearson correlation score.*

Heuristic algorithm for clustering the trajectories The GreedyRSC method is an heuristic algorithm to apply the RSC model [6]. It performs a soft clustering, where the clusters may overlap and do not necessarily cover the entire data set. In addition to the $Q(t)$ function, it requires four parameters:

- x_1 : Minimum size of cluster
- x_2 : Maximum size of cluster
- x_3 : Maximum interset significance score between two clusters.
- x_4 : Minimum significance score.

Houle [6] defines the significance score by the function $Z_1(A)$ and the interset significance score by the function $Z_1(A,B)$ where A and B are two clusters.

The minimum size x_1 of pattern means that all clusters would be composed of at least x_1 trajectories. To respect this constraint, we have to choose the minimum significance score $x_4 = \sqrt{x_1(|S|-1)}$ where $|S|$ is the number of trajectories. We can prove the computation of the minimum significance score as follows:

Let A be a cluster (set of trajectory),

$$\begin{aligned} |A| &\geq x_1 \geq 0 \\ SR_1(A)\sqrt{|A|(|S|-1)} &\geq SR_1(A)\sqrt{x_1(|S|-1)} \\ Z_1(A) &\geq SR_1(A)\sqrt{x_1(|S|-1)} \end{aligned}$$

where $SR_1(A)$ is the intra-set correlation measure. A value of 1 indicates total identity among the trajectories of A, whereas a value approaching 0 indicates total difference. Because $0 \leq SR_1(A) \leq 1$, we need a minimum significance score x_4 equal to $\sqrt{x_1(|S|-1)}$ to ensure that all clusters have a minimum of x_1 trajectories.

For studying the RSC clustering robustness, we performed 64 ($= 4 \times 4 \times 4$) analyses with four different values covering a wide range for the variables x_1, x_2 and x_3:

- $x_1 = [2, 5, 10, 50]$
- $x_2 = [1500, 2000, 3000, 6000]$
- $x_3 = [0.1, 0.5, 1.0, 2.0]$

Because RSC is a non-deterministic clustering method, we performed five replicates of each of the 64 clustering analyses.

Next hierarchic clustering based on Jaccard index permitted to compare the different clusters obtained by the 320 clustering. The clusters were classified in several groups and we extracted the intersection for each group. We named "core i" the intersection to the "group i", for example i.e. the set of trajectories that belong to all the clusters of "group i".

2.3 Identification of the over-represented proteins in each core

Trajectories clustering was performed using correlation score based on the presence and the absence of proteins. The core of each group can be characterized by a set of over-represented proteins, i.e. the proteins that appear more often in the trajectories of the core than we would expect if we had selected the same number of trajectories randomly (fig. 2).

We can compute the protein level of representation for each cluster with a zScore of protein frequency:

$$Z_A(p) = \frac{N_A(p) - F_S(p)|A|}{\sqrt{F_S(p)|A|(1 - F_S(p))}} \tag{3}$$

where p is a protein and A is a cluster of trajectories, $N_A(p)$ is the number of trajectories in A involving p, $F_S(p)$ is the frequency of p in all trajectories S and $|A|$ is the size of cluster.

The zScore allows to normalize the frequency of proteins in the cluster of trajectories compared to all trajectories. For each core, we computed the zScore of all the proteins. We then identified a list of over-represented proteins with a high zScore.

Based on the scores of over-representation of proteins in trajectories, we next searched for the biological significance of the protein signatures that characterized the three cores. The Gene Set Enrichment Analysis (GSEA) is a method

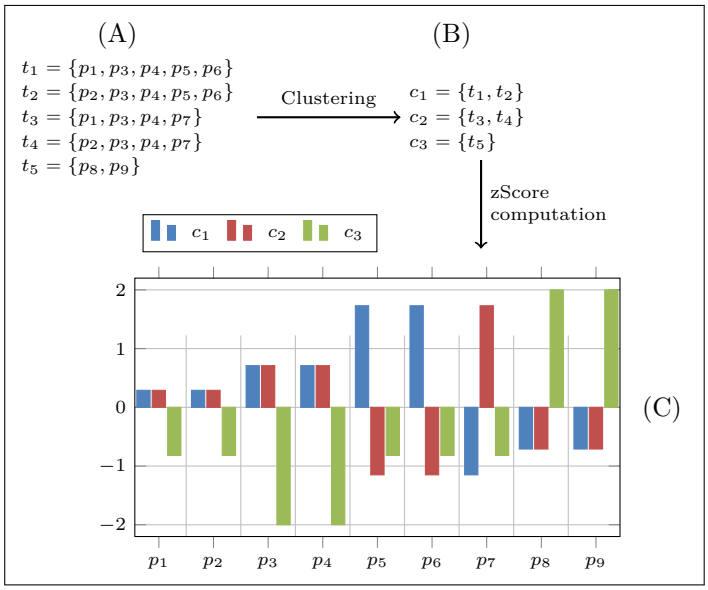

Fig. 2: *Example of calculation for determining over-represented proteins between three cores of trajectories. (A) t_1, t_2, t_3, t_4 and t_5 are five trajectories containing proteins p_1, p_2, p_3, p_4, p_5, p_6, p_7, p_8 and p_9. (B) the clustering method identifies three cores c_1, c_2 and c_3. (C) distribution of representation level of proteins in c_1, c_2 and c_3 cores. For example, p_1 and p_2 are slightly over-represented in the cores c_1 and c_2 but not over-represented in c_3, contrary to p_9. The core c_3 can be characterized by p_8 and p_9.*

which permits to identify significantly enriched classes of genes or proteins in a large set of genes or proteins, that are associated with specific biological functions. The analyses were performed using the GSEA tool developed by the Broad Institute [22]. The lists of proteins and their respective score frequency were used as input and biological processes from Gene Ontology database were selected as *gene sets database*. The outputs were the "biological processes" terms significantly enriched in the submitted lists of proteins from each core when compared with the other cores.

3 Results

3.1 TGF-β signaling trajectories are highly connected

In order to identify functional families of signaling trajectories based on the comparison of their signaling molecules (proteins) content, we performed a preprocessing step as described in material and method. Discarding TGF-β and the target genes from the 15,934 trajectories led to 6017 trajectories composed of 321 different proteins.

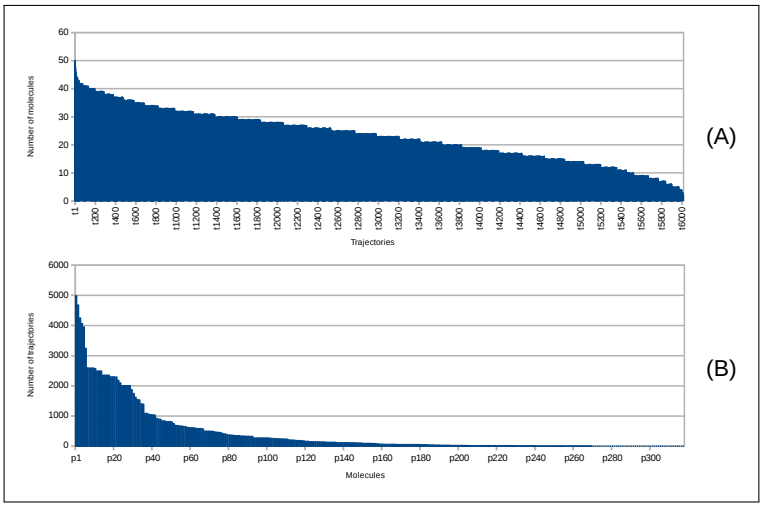

Fig. 3: *Distribution of (A) the number of molecules for each trajectory and (B) the number of trajectories involving each molecule. These results showed that most proteins are shared by many trajectories suggesting high degree of connectivity of TGF-β-dependent signaling pathways.*

As illustrated in figure 3, the number of proteins per trajectory varied from 1 to 50, with more than 90 percent of trajectories containing at least 10 proteins. Analyses of the distribution of each protein in all trajectories showed a great heterogeneity. More than 70 proteins were present in at least 500 trajectories, and 6 proteins were present in more than 3000 trajectories (FOS, JUN, ATF2, MAP2K4, ELK1, JAK2). Conversely 75 proteins appeared in fewer than 10 trajectories. Together these results showed that many proteins are shared by many trajectories suggesting high degree of connectivity of TGF-β-dependent signaling pathways.

3.2 Relevant Set Correlation method identifies five families of trajectory clusters

Using a greedy strategy and a large variety of parameters, we performed 320 clusterings over the 6017 trajectories. Each clustering generated 3, 4 or 5 clusters leading to 1139 different clusters of trajectories. In order to compare their similarity, we calculated the Jaccard index based on the number of shared trajectories between two given clusters. Using a hierarchical classification of this similarity between clusters, we identified five groups of clusters (fig. 4).

To characterize the five groups of clusters, we analyzed the number of clusters associated with each group, the number of trajectories associated with these clusters (average cluster size) and the redundancy between clusters (union and intersection). As described in table 4, the groups 1 and 2 were characterized by clusters generated from 320 and 319 clusterings respectively, suggesting a robust

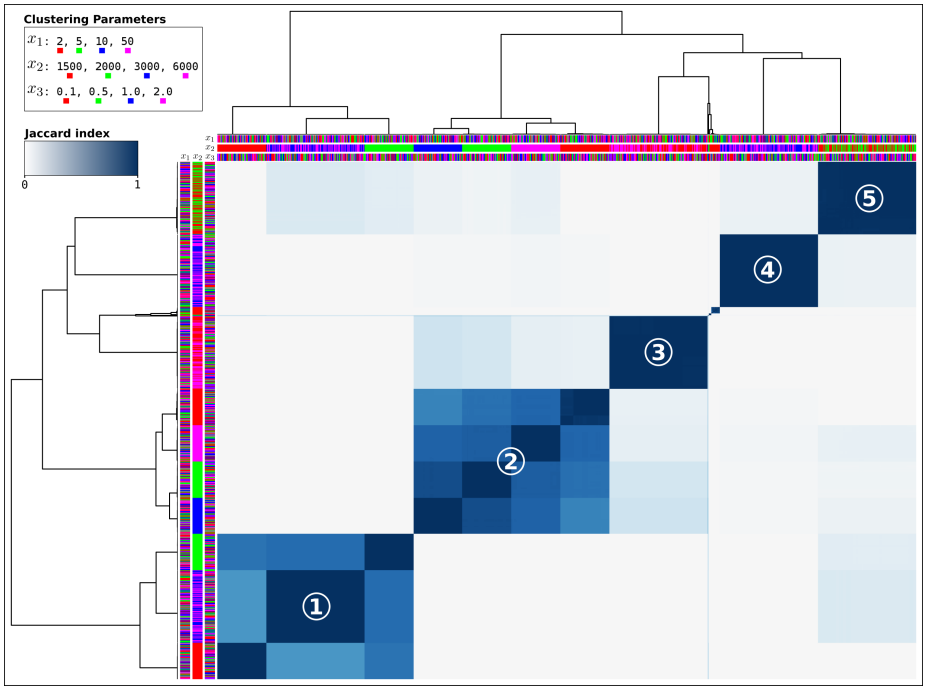

Fig. 4: *Hierarchical classification of the clusters generated by the 320 clusterings using varying parameters (x_1, x_2 and x_3) according to their similarities (Jaccard index). The parameter values are indicated by four different colors. Each cluster results from a clustering characterized by a combination of the three parameters. The five groups of clusters identified are numbered from 1 to 5 and the intensity of blue color indicates the Jaccard index between two clusters.*

classification of trajectories. The three other groups 3, 4 and 5 contained clusters generated from 160 clusterings suggesting higher sensitivity to parameters. The average cluster size expressed as the average number of trajectories contained in clusters varied from 202 in group 4 to 2170 in group 1. The core of a group is the intersection of the clusters of a group. It is the set of the trajectories that belong to all the clusters of the group, so it allows to focus on the most stable trajectories of the group. The cores of groups 1 and 2 contained 1485 (57%) and 1458(67%) trajectories respectively, while the core size of groups 3, 4 and 5 were either identical or very similar to the union of clusters. To further characterize these cores, we determined the number of proteins implicated in the trajectories and the number of target genes activated by these signaling trajectories. While the total number of proteins involved in trajectories from each group was almost similar, the number of target genes was highly variable. The trajectories from the most important core 1 (1485) were characterized by 114 proteins but only 3 target genes suggesting complex combinations of signaling for these genes. At the opposite the trajectories from core 4 that contained only 202 trajectories were characterized by 156 proteins that activate 19 genes.

	Group 1	Group 2	Group 3	Group 4	Group 5
Number of clusters	320	319	160	160	160
Average cluster size (Number of trajectories)	2170.0	1905.58	899.62	202.0	877.12
Union of clusters (Number of trajectories)	2590	2289	904	202	888
Core size = Intersection of clusters (Number of trajectories)	1485	1458	894	202	870
Number of proteins for each core	114	188	110	156	151
Number of target genes for each core	3	68	58	19	16

Table 4: *Statistics of clusters.*

3.3 Cores are characterized by specific over-represented protein signatures associated with biological processes

In order to characterize the protein signature of each core, we investigated the level of representation of proteins within all the trajectories from each core. For that purpose, we calculated the zScore of protein frequency in each clusters. The list of protein zScores for each core was provided as supplementary tables [1]. As shown in figure 5, the zScore distribution of the 321 proteins from trajectories of each core was highly heterogeneous. Interestingly, the zScore distribution from core 1 was inversely correlated with that of core 2 suggesting different biological functions associated with trajectories. Together these observations suggested that each core of trajectories was characterized by specific protein signatures. During the course of the analysis of the zScore values, we showed that the probability to randomly find a protein in a group of trajectories with a zScore higher than 4.0 is less than 0.006%. As a consequence, we decided to select the proteins with a zScore superior to 4.0 to refine the protein signatures of the five cores of trajectories.

Based on the scores of proteins over-representation in trajectories, we next searched for the biological significance of the protein signatures that characterized the five cores. Gene Set Enrichment Analysis (GSEA) is a method for identifying significantly the elements of a set that appear more often in the set that one would expect if the set had been randomly assembled. It is typically used for determining which specific biological functions are specific of a set of genes or proteins. The analyses were performed using the GSEA tool developed by the Broad Institute [22]. The lists of proteins and their respective score frequency were used as input for GSEA analysis and the outputs are the lists of enriched biological processes (see supplementary tables [1]). As shown in figure 6, each core was characterized by specific set of biological functions since 57%, 90%, 80%, 81% and 88% of GO-terms were specific of core 1, core 2, core 3, core 4, and core 5, respectively. In order to identify the representative terms, we used Revigo [23] that reduces the list of GO terms on the basis of semantic similarity measures. Consequently trajectories from core 1 and core 2 were mainly

[1] `http://www.irisa.fr/dyliss/public/tgfbVisualization/supplementaryData`

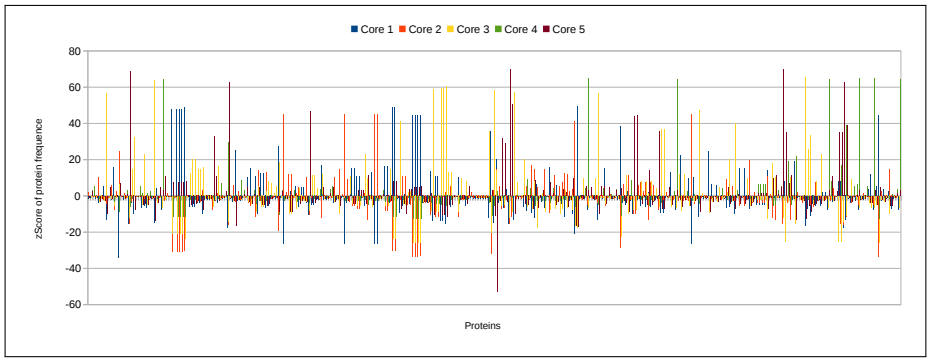

Fig. 5: *Distribution of zScore values of the frequences of 321 proteins in the trajectories from cores of the five cluster groups. The zScore distribution of the 321 proteins from trajectories of each core is highly heterogeneous. These observation suggested that each core of trajectories was characterized by specific protein signatures*

associated with antigen receptor-mediated signaling and serine-threonine kinase activity, respectively (fig. 6). The functional annotation of cores 3 and 4 were more heterogeneous while core 5 clustered signaling trajectories that are clearly involved in immune response. An important conclusion from these results is that even if signaling trajectories share many proteins, our analysis revealed groups of trajectories that correspond to different functional families.

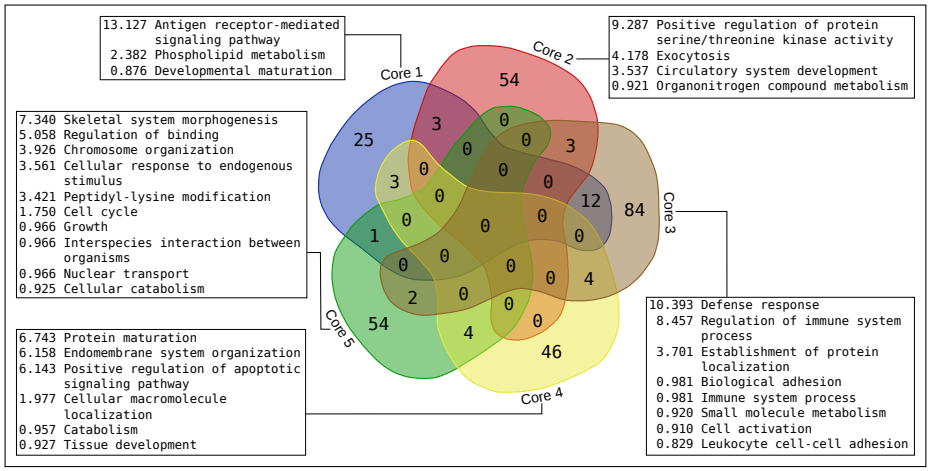

Fig. 6: *Gene ontology enrichment analysis. The lists of proteins and their respective score frequency from each Core are used for GSEA.The lists of enriched GO terms associated to biological processes are compared using Venn diagram and the score is uniqueness score of the GO-term calculated by REVIGO tool.*

Together our data demonstrate that our approach for clustering signaling trajectories based on their protein content is powerful to discriminate TGF-β-influenced networks. To illustrate the complexity of TGF-β-dependent signaling pathways, we compiled the trajectories from each core and the resulting networks were illustrated in figure 7.

3.4 Web visualization of TGF-β-influenced networks

To facilitate the exploration of the signaling trajectories clustered in each core, we developed a web interface:
`http://www.irisa.fr/dyliss/public/tgfbVisualization`

The interface is based on the Cytoscape JavaScript library (`http://js.cytoscape.org`). Nodes are proteins and their size is correlated to the number of trajectories involving this protein. Node color indicates the occurrence of the protein in the trajectories from a core. The occurrence is based on the zScore of protein frequency (blue for zScore<0 and red for zScore>0) and selection of the level of occurrence permits to filter information. The black circle nodes illustrate biological reactions (association, dissociation, phosphorylation,degradation, migration etc) as described in [2]. The black edges link proteins to the input or the output of a reaction, green edges link the protein that regulates positively the reaction and red edges link the protein that regulate negatively the reaction (fig. 7). Exploration of the graphs is facilitated by manually repositioning nodes and edge. The graph can be exported in JSON format.

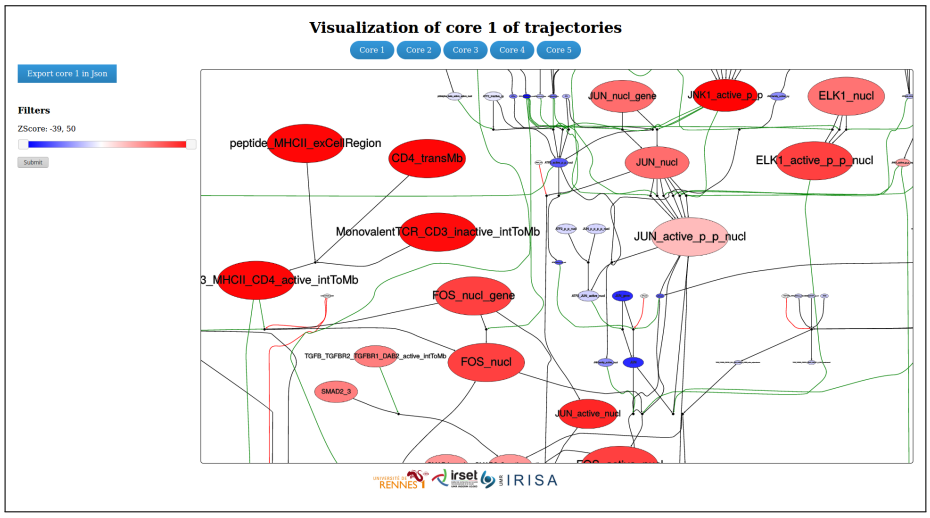

Fig. 7: *Screenshot of the Web visualization of core 1. A node is a bio-molecule, the node size corresponds to the number of trajectories involving this bio-molecule and the node color correspond to the representation of the bio-molecule in the core (blue the bio-molecule is under-represented and red it is over-represented).*

4 Discussion

Cell signaling networks are essential to life. They allow cells to sense and interpret microenvironment changes to provide adapted phenotypes such as differentiation, proliferation and apoptosis. As a result, disturbance or alteration of signaling networks have been associated with many diseases such as fibrosis and cancer. In particular, TGF-β plays major roles both in physiological and pathological processes through canonical and non canonical signaling pathways that cross-react with other pathways [13]. Understanding how signaling molecules combine to provide signaling trajectories is a prerequisite for future therapeutic strategies, however analyses of large signaling networks remain a challenging task.

While qualitative approaches are suited to large-scale networks, the analysis of numerous signaling trajectories remains difficult. Reduction methods focus on diminishing the size of large-scale boolean networks [25,18] or dividing methods in several sub-networks [26]. However, these methods typically consist in performing the reduction before the analysis, whereas for TGF-β we focused on an exhaustive analysis of the signaling network.

In addition to exhaustivity, the originality of our approach lies in analyzing the signaling trajectories according to their protein composition rather than the genes they influence. Our approach was motivated by the fact that signaling pathways share a large number of "modular domains" in various combinations [11]. These combinations support the functional diversity of signaling pathways.

These modular domains provide the underlying structure of the signaling trajectories. Our goal was to identify groups of similar trajectories. When considering two trajectories, the more modules they share, the more similar they are. There are many clustering methods (for example hierarchical, K-means, distribution-based, density-based) [8]. As we mentioned previously, a modular domain can be involved in multiple combinations, so their study required soft-clustering methods which allows clusters to overlap and share some elements. We selected shared nearest-neighbours (SNN) clustering, which have successfully been applied to handle the heterogeneity and large-scale of trajectories [5]. The Relevant Set Correlation method is further appropriate in that there is no need to define the neighborhood size. Likewise, our approach does not rely on a priori assumption on the number of clusters.

Relevant Set Correlation proved to be a robust clustering method for our dataset. All 64 combinations of parameter values generated clusters that systematically belonged to group 1 and group 2 and one of groups 3, 4 and 5. Half the simulations produced clusters that belonged to groups 3, 4 or 5. In figure 4, the analysis of the influence of the parameter values for groups 3, 4 and 5 showed that x_1 and x_3 had no influence on the groups, whereas pairs of values of x_2 were associated to different groups: the two lowest with group 5, the two highest with group 4 and a combination of the highest and the lowest with group 3. Surprisingly, the two intermediate values of x_2 (2000 and 3000) were markers of groups 4 and 5, for which they were associated with their closest extreme value,

whereas the lowest and highest values of x_2 were associated to group 3. This indicates that RSC produced either groups 3 and 5 for the low values of the range of the clusters' maximum size (x_2), or groups 3 and 4 for the high values. At this point, further analysis is required for determining either which of the low or high values are the more adapted to our dataset, or if groups 3, 4 and 5 are all biologically-relevant and we are facing a limitation of RSC. Overall, our study with the various combinations of parameter values showed that (1) because it is non-deterministic, performing multiple runs with the same parameter values is useful, (2) RSC is a robust clustering method for our dataset, (3) groups 1 and 2 were independent from the parameter values whereas groups 3, 4 and 5 were not, and (4) low values of clusters' maximum size produced clusters in groups 3 and 5, whereas high values produced clusters in groups 3 and 4. According to this observation, the over-represented proteins in trajectories from core 1 and 2 clearly discriminate the canonical pathways associated with TGF-β receptor-dependent cell response during injury and development (core 1) and the non canonical pathways involving all other kinase-dependent signaling (core 2), respectively. Together these two cores of clusters illustrated the so-called "Jekyll and Hyde" aspects of TGF-β in cancer [3].

Although it does not rely on a priori knowledge, our approach may be dependent on annotation bias. Since biological knowledge is by nature incomplete, some well studied signaling processes may be described in details in databases, whereas some lesser studied ones would be incompletely described, or with a coarser granularity (usually both). This would then result in a higher frequency of the well studied modules and give a misleading impression of being more important. It should be noted that this is an intrinsic bias of the data we rely on, and not of our analysis method. This bias should be taken into account by the experts when analyzing the results.

5 Conclusion

We proposed an exhaustive and without prior assumption soft-clustering-based method for identifying families of functionally-similar trajectories in signaling network. Among 15,934 trajectories involved in TGF-β signaling, our approach identified five groups of trajectories based on their molecular composition. The functional characterization of these groups revealed that each group is involved in different roles of TGF-β, which confirmed that our approach yields biologically-relevant results. The approach can be generalized to explore any large-scale biological pathways.

References

1. Aldridge, B.B., Burke, J.M., Lauffenburger, D.A., Sorger, P.K.: Physicochemical modelling of cell signalling pathways. Nature cell biology 8(11), 1195–1203 (2006)
2. Andrieux, G., Le Borgne, M., Théret, N.: An integrative modeling framework reveals plasticity of tgf-β signaling. BMC systems biology 8(1), 1 (2014)

3. Bierie, B., Moses, H.L.: Tumour microenvironment: Tgfβ: the molecular jekyll and hyde of cancer. Nature Reviews Cancer 6(7), 506–520 (2006)
4. ElKalaawy, N., Wassal, A.: Methodologies for the modeling and simulation of biochemical networks, illustrated for signal transduction pathways: A primer. Biosystems 129, 1–18 (2015)
5. Hamzaoui, A., Joly, A., Boujemaa, N.: Multi-source shared nearest neighbours for multi-modal image clustering. Multimedia Tools and Applications 51(2), 479–503 (2011)
6. Houle, M.E.: The relevant-set correlation model for data clustering. Statistical Analysis and Data Mining 1(3), 157–176 (2008)
7. Ikushima, H., Miyazono, K.: Biology of transforming growth factor-β signaling. Current pharmaceutical biotechnology 12(12), 2099–2107 (2011)
8. Joshi, A., Kaur, R.: A review: Comparative study of various clustering techniques in data mining. International Journal of Advanced Research in Computer Science and Software Engineering 3(3) (2013)
9. Kashtan, N., Alon, U.: Spontaneous evolution of modularity and network motifs. Proceedings of the National Academy of Sciences of the United States of America 102(39), 13773–13778 (2005)
10. Kestler, H.A., Wawra, C., Kracher, B., Kühl, M.: Network modeling of signal transduction: establishing the global view. Bioessays 30(11-12), 1110–1125 (2008)
11. Lim, W.A.: Designing customized cell signalling circuits. Nature reviews Molecular cell biology 11(6), 393–403 (2010)
12. Luo, K.: Signaling cross talk between tgf-β/smad and other signaling pathways. Cold Spring Harbor Perspectives in Biology 9(1), a022137 (2017)
13. Massagué, J.: Tgfβ signalling in context. Nature reviews Molecular cell biology 13(10), 616–630 (2012)
14. Mu, Y., Gudey, S.K., Landström, M.: Non-smad signaling pathways. Cell and tissue research 347(1), 11–20 (2012)
15. Peisajovich, S.G., Garbarino, J.E., Wei, P., Lim, W.A.: Rapid diversification of cell signaling phenotypes by modular domain recombination. Science 328(5976), 368–372 (2010)
16. Rauzy, A.: Guarded transition systems: a new states/events formalism for reliability studies. Proceedings of the Institution of Mechanical Engineers, Part O: Journal of Risk and Reliability 222(4), 495–505 (2008)
17. Saadatpour, A., Albert, R.: Discrete dynamic modeling of signal transduction networks. Computational Modeling of Signaling Networks pp. 255–272 (2012)
18. Saadatpour, A., Albert, R., Reluga, T.C.: A reduction method for boolean network models proven to conserve attractors. SIAM Journal on Applied Dynamical Systems 12(4), 1997–2011 (2013)
19. Samaga, R., Klamt, S.: Modeling approaches for qualitative and semi-quantitative analysis of cellular signaling networks. Cell communication and signaling 11(1), 1 (2013)
20. Schaefer, C.F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., Buetow, K.H.: Pid: the pathway interaction database. Nucleic acids research 37(suppl 1), D674–D679 (2009)
21. Scott, J.D., Pawson, T.: Cell signaling in space and time: where proteins come together and when they're apart. Science 326(5957), 1220–1224 (2009)
22. Subramanian, A., Tamayo, P., Mootha, V.K., Mukherjee, S., Ebert, B.L., Gillette, M.A., Paulovich, A., Pomeroy, S.L., Golub, T.R., Lander, E.S., et al.: Gene set

enrichment analysis: a knowledge-based approach for interpreting genome-wide expression profiles. Proceedings of the National Academy of Sciences 102(43), 15545–15550 (2005)
23. Supek, F., Bošnjak, M., Škunca, N., Šmuc, T.: Revigo summarizes and visualizes long lists of gene ontology terms. PloS one 6(7), e21800 (2011)
24. Tian, M., Neil, J.R., Schiemann, W.P.: Transforming growth factor-β and the hallmarks of cancer. Cellular signalling 23(6), 951–962 (2011)
25. Zañudo, J.G., Albert, R.: An effective network reduction approach to find the dynamical repertoire of discrete dynamic networks. Chaos: An Interdisciplinary Journal of Nonlinear Science 23(2), 025111 (2013)
26. Zhao, Y., Kim, J., Filippone, M.: Aggregation algorithm towards large-scale boolean network analysis. IEEE Transactions on Automatic Control 58(8), 1976–1985 (2013)
27. Zi, Z., Chapnick, D.A., Liu, X.: Dynamics of tgf-β/smad signaling. FEBS letters 586(14), 1921–1928 (2012)