The long story of SWIM methodology: from grapevine to personalised medicine

Paola Paci[1,2]
[1] *Department of Computer, Control and Management Engineering, Sapienza University of Rome, Rome, Italy*
[2] *Karolinska Institutet, 17177 Stockholm, Sweden*

***Abstract*—SWIM is a recently developed network-based tool that fulfils the criteria of the new quickly emerging field of Network Medicine in finding disease-associated genes, called switch genes. The phenotype-specific applications of SWIM are broad and include the identification of switch genes in grapevine berry maturation as well as in complex diseases, including but not limited to human cancers.**
Here, a brief summary of the promising results obtained by applying SWIM in different biological contexts is presented.

***Keywords*—Network Medicine, Network Theory, Disease genes.**

I. INTRODUCTION

Recently, I developed a new promising methodology, called SWIM (SWItch Miner), which integrates different network-based methods to analyse the correlation network arising from large-scale gene expression data [1]. Considering the topological properties of the nodes and assessing their functional roles according to their ability to convey information within and between modules in the network, SWIM identifies a small pool of genes (called switch genes) that are associated with intriguing patterns of molecular co-abundance and play a crucial role in the observed phenotype.

The phenotype-specific applications of SWIM are broad and include the identification of switch genes in grapevine berry maturation (*Vitis vinifera*) as well as in human cancers, including Glioblastoma (Fig. 1).

In viticulture, SWIM has been gainfully applied to the global gene expression atlas of grapevine in order to identify switch genes between immature and mature phase of the developmental program of grapevine [2]. In cancer research, SWIM network analysis has been gainfully applied to a large panel of TCGA (The Cancer Genome Atlas) cancer datasets in order to characterise disease etiologies and identify potential therapeutic targets [1]. SWIM has also been used to investigate glioblastoma multiforme (GBM) and to uncover new insights into the molecular mechanism determining the stem-like phenotype of glioblastoma cells [3], [4].

Recently, SWIM methodology has been successfully applied to the chronic obstructive pulmonary disease (COPD), a severe lung disease characterized by progressive and incompletely reversible airflow obstruction [5]. The results of this study would both support known pathways and provide evidence for novel pathways in COPD pathogenesis. In the last two years, SWIM has been applied within the framework of Network Medicine to study the interplay between switch genes and human diseases in the human interactome (i.e., the cellular network of all physical molecular interactions) [6].

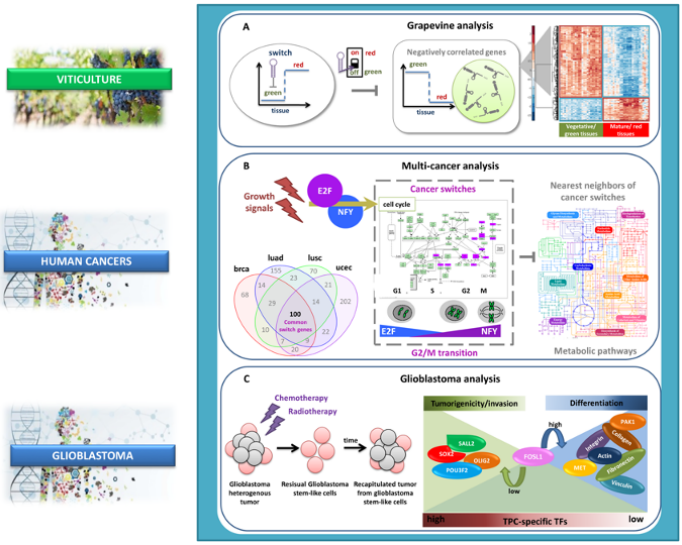

Fig. 1: Phenotype-specific applications of SWIM methodology.

In the following, a detailed description of the more recent applications of SWIM to complex diseases is provided.

II. METHODS

SWIM is a freely downloadable network-based tool, developed both in MATLAB [1] and in R language [7], which predicts important (switch) genes that are strongly associated with drastic changes in cell phenotype. SWIM first computes the differentially expressed genes (DEGs) between two conditions of interest (e.g., normal state versus tumor state) and then builds a gene correlation network (GCN) by calculating correlations (positive and negative) between the expression profiles of each gene pair. Specifically, SWIM implements a hard thresholding approach to build a GCN where nodes are DEGs, and a link occurs if their expression profiles are highly correlated or anti-correlated (according to a defined threshold). Then, SWIM classifies each network hub (i.e., nodes with degree at least equal to 5 according to [8]) as date, party, or fight-club on the basis of the Average Pearson Correlation Coefficient (APCC) between its expression profile and that of its first nearest neighbours. Date hubs show a positive and mild APCC value; party hubs show a positive and high APCC value; fight-clubs hubs show a negative APCC

value. To assign a role to each node in the GCN, SWIM firstly searches for clusters (or modules) using the k-means algorithm and evaluates the quality of clusters by minimising the Sum of the Squared Error (SSE), depending on the distance of each object to its closest centroid. The position of an elbow (i.e., change of the slope) in the behaviour of the SSE as function of the number of clusters (scree plot) indicates the number of clusters to use. Then, SWIM draws the heat cartography map by evaluating two coordinates related to their intra- and inter-modular connections: the clusterphobic coefficient, which is a measure the fear of each node to be confined in its own cluster and measures the links of each node to nodes outside its own cluster; the within-module degree, which measures how "well-connected" each node is within its own cluster. In the heat cartography map, dots (i.e., nodes of the correlation network) are distributed across seven regions (R1 to R7) according to their clusterphobic coefficient (x-axis), and to their within-module degree (y-axis). Each node is coloured according to its APCC value. Nodes having much more external than internal links present high values of the clusterphobic coefficient and are called connectors, whereas high values of the within-module degree denote nodes that are hubs within their community and are called local hubs. Switch genes are defined as a subset of fight-club nodes with the following features: i) they are network connectors that mainly interact outside their own cluster (very high value of the clusterphobic coefficient); ii) they are not local hubs (very low value of the within-module degree); iii) they are mainly anti-correlated with their interaction partners (negative value of the APCC).

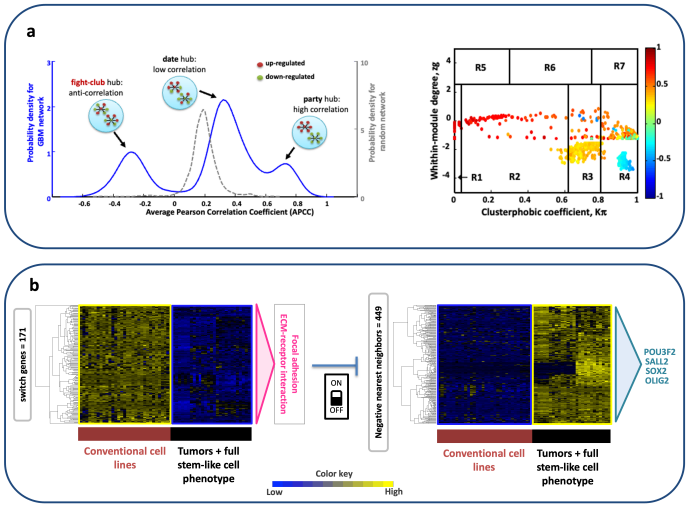

Fig. 2: SWIM application to the GBM dataset of [13]. (a) Left: probability distribution of APCC for hubs identified in the GBM correlation network (blue solid line) and in its randomized counterpart obtained by shuffling the edges but preserving the degree of each node (grey dashed line). Differently from the randomized case, the true APCC distribution shows a clear trimodal pattern where peaks correspond to previously reported hub categories (such as party and date hubs) but also to the new category of hubs, the fight-club hubs. Right: heat cartography maps where dots correspond to network nodes colored according of APCC. (b) Dendrogram and heat map of switch genes (left) and of their negative nearest neighbors (right) with their corresponding enriched pathways The expression profiles of the switch genes and their negative nearest neighbors are clustered according to genes (rows) and GBM cell lines (columns), using Pearson correlation distance as metrics. Heat map colors represent different expression levels increasing from blue to yellow.

III. Results and Discussions

A. Glioblastoma

Glioblastoma is the most aggressive and frequent brain tumour, with a median survival time of 12–15 months from diagnosis [9], [10], [11]. The mortality rate is extremely high with the 5-years survival rate achieved for only 5% of patients. This tumour is resistant to the standard therapies like radio and chemotherapy. Its aggressiveness is due to the presence of cancer stem-like cells that sustain tumour growth and are hence named "tumour fuel". Cancer stem-like cells are cancer cells that have characteristics typical of normal stem cells: i) self-renewal that is the ability to maintain their undifferentiated state; ii) potency that is the ability to differentiate into specialised cell types. Cancer stem-like cells are resistant to many conventional cancer therapies and cause relapse and metastasis by giving rise to new tumours [12]. Thus, targeting cancer stem-like cells could pave the way for new therapeutic strategies.

A recent study identified 19 neurodevelopmental transcription factors (TFs) that are selectively expressed in glioblastoma stem-like cells to maintain their stem-like phenotype and prevent differentiation [14]. A subset of only four of them (named 4-core TFs), SOX2, OLIG2, POU3F2, and SALL2, has been shown to be sufficient to fully reprogram differentiated cells into glioblastoma stem-like cells [14].

In order to identify switch genes related to the stem-like phenotype, SWIM was applied to glioblastoma dataset of [14] and then the further dataset of [13] was used to validate the results [3], [4]. In Fig. 2 the results obtained by applying SWIM to the GBM dataset obtained in [13] are shown. The APCC behaviour (blue solid line) shows a clear peak for negative values indicating the fight-club hubs (Fig. 2a left). The APCC randomised counterpart (grey dashed line), obtained by shuffling the edges but preserving the degree of each node, indicates that fight-club hubs are not a random event. A subset of the fight-club hubs falling in the region of connectors (R4) of the heat cartography map corresponds to switch genes (Fig. 2a right, blue dots). The expression profiles (z-score normalised) of the switch genes (Fig. 2b left) and their negative nearest neighbours (Fig. 2b right) indicates that switch genes are all up-regulated in differentiated cells, and their activation strongly correlates with the inhibition of their first network interactors, including the 4-core TFs. By performing a functional enrichment analysis, switch genes were found to be involved in cell-cell communication pathway. Thus, while their activation could promote differentiation and restrain tumour growth, their repression could promote tumour invasiveness due to the loss of cell-cell adhesion.

Among the common switch genes obtained by running

SWIM on the two GBM datasets of [13] and [14], there is FOSL1. It is up-regulated in differentiated glioblastoma cells and this up-regulation highly correlates with the over-expression of genes involved in cell-cell communications (Fig. 3 top left/middle). It is down-regulated in stem-like cells and this down-regulation highly correlates with the up-regulation of the 4-core of TFs (Fig. 3 top right). In order to investigate possible co-regulation of the 4-core of TFs, their promoter regions were inspected to search for enriched motifs and they were found to harbour a consensus binding site for FOSL1 (Fig. 3 bottom right).

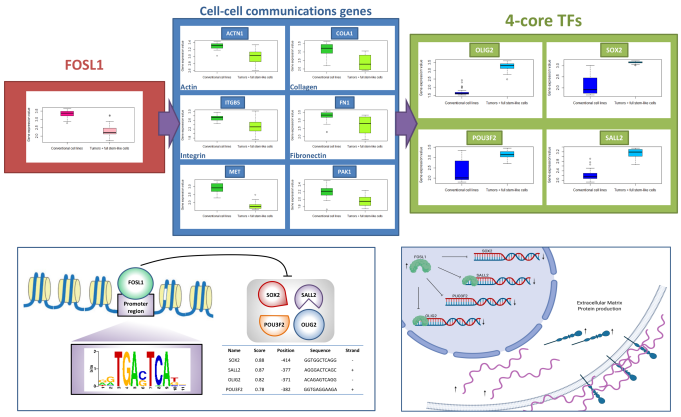

Fig. 3: FOSL1 mode of action. Upper panel: from left to right, boxplot of the expression of FOSL1, its positive and negative nearest neighbours in the GBM correlation network, in conventional GBM cell lines with respect to cancer stem-like cells in GBM dataset of [13]. Lower panel: logo plot for the statistically significant enriched motif found by using JASPAR to analyse the promotore regions of the 4-core of TFs (left). This motif corresponds to a consensus binding site for FOSL1. A schematic representation of the FOSL1 mode of actions: it acts as putative repressor of the 4-core of TFs and its activation reduces the cells' ability to generate aggregates increasing the extracellular matrix component.

Altogether these findings suggest FOSL1 as possible therapeutic biomarker of glioblastoma, which could promote the differentiation of cancer stem-like cells by repressing the 4-core TFs. This hypothesis has been partially experimentally validated in [15], where the NTERA-2 and HEK293T cells where selected for an in-vitro study to investigate the role of FOSL1 in the reprogramming mechanisms (Fig. 3 bottom right). The two cell lines where transfected with a constitutive FOSL1 cDNA plasmid. This study showed that FOSL1 i) directly regulates the 4-core of TFs binding their promoter regions and reducing their expression; ii) is involved in the deregulation of several stemness markers; iii) reduces the cells' ability to generate aggregates increasing the extracellular matrix component FN1. Although further experiments are necessary, these findings support the hypothesis that FOSL1 may reprograms the stemness by regulating the 4-core TFs.

This result could have a significant impact on personalized healthcare, since promoting differentiation and thus restraining tumor growth may support rational, personalized planning of disease prevention or treatment.

B. COPD

COPD is a heterogeneous and complex syndrome influenced by both genetic and environmental determinants, and is one of the main causes of morbidity and mortality worldwide.

By applying SWIM on COPD [5], the correlation network turned out to be formed by three well-characterised modules (data not shown, see [5]): i) one (module 3) populated by switch genes, all up-regulated in COPD cases and involved in COPD-related pathways, like B cell receptor signalling pathway; ii) one (module 1) populated by negative interactors of switch genes, down-regulated in COPD cases, including well-known GWAS genes like AGER and CAVIN1; iii) one (module 2) populated by well-recognised immune signature genes, all up-regulated in COPD cases. Switch genes appear to form localised connected subnetworks displaying an intriguingly common pattern of up-regulation in COPD cases compared with controls. A more sophisticated analysis revealed that they were not only topologically related, but also functionally relevant to the observed phenotype as witnessed by their enrichment in the regulation of inflammatory and immune responses.

In order to demonstrate the disease specificity of switch genes, SWIM was applied on another COPD dataset and on the acute respiratory distress syndrome (ARDS), another severe lung disease with an inflammatory component. The two lists of COPD switch genes were found to form overlapping modules in the human interactome that are topologically separated with the ARDS switch genes (data not shown, see [5]). This observation demonstrates that even though different diseases can share similar endophenotypes, the molecular network determinants responsible for them are disease-specific.

Interestingly, ARDS switch genes were different from COPD switch genes, but the major pathways affected in the two diseases were similar, emphasising that different diseases often have common underlying mechanisms and share intermediate endophenotypes (convergent phenotypes) [16].

C. Network Medicine

Network Medicine is a new emerging paradigm in medicine, where disease proteins are assumed not to be randomly scattered, but agglomerate in specific regions of the molecular interactome, suggesting the existence of specific disease network modules for each disease [17], [18], [19].

To quantify the interplay between switch genes and human diseases in the human interactome, the results obtained by the pan-cancer [1] and COPD [5] SWIM-based analysis were complemented with the application of SWIM tool on two cardiac disorders (i.e., ischemic and non-ischemic cardiomyopathy) and on Alzheimer's disease (AD) [6]. Switch genes associated with specific disorders were found to be not randomly scattered but they form localised connected subnetworks (Fig. 4a). These subnetworks overlap between similar diseases (like cancers or cardiac disorders) and are situated in different neighbourhoods for pathologically distinct phenotypes (like AD and COPD), showing a direct relation between the pathobiological similarity of diseases and their

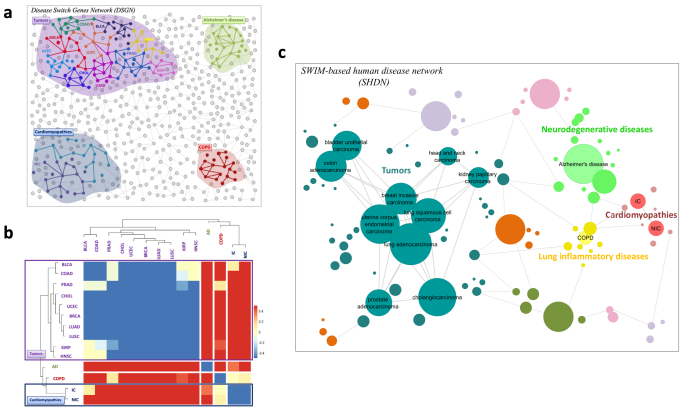

Fig. 4: SWIM and Network Medicine. (a) Switch genes mapping on the human interactome. (b) Hierarchical clustering where the network-based separation measure was used as a distance metric: blue color corresponds to overlapping modules, red color corresponds to non-overlapping modules. (c) SHDN where nodes are the disorders and the size of each node is proportional to the number of switch genes involved in the corresponding disorder. Nodes are coloured based on the disorder class to which they belong. Labeled nodes correspond to the diseases analysed, while unlabelled nodes are artificial.

relative distance in the human interactome (Fig. 4a). These results were confirmed by the hierarchical clustering where two main clusters were found: one including all tumour datasets and one including the two cardiomyopathies along with AD and COPD datasets as isolated branches (Fig. 4b). Finally, the first SWIM-informed Human Disease Network (SHDN) was built (Fig. 4c), where nodes correspond to distinct disorders and a link occurs between two diseases if they share a substantial number of switch genes. Clustering of nodes of similar color (denoting the same disease class) means that similar pathophenotypes have a higher probability of sharing switch genes than do pathophenotypes that belong to different disease classes.

These findings support the hypothesis that SWIM-based correlation network analysis can serve as a useful tool for efficient screening of potentially new disease gene associations. When integrated with an interactome-based network analysis, it not only identifies novel candidate disease genes, but also may offer testable hypotheses by which to elucidate the molecular underpinnings of human disease and reveal commonalities between seemingly unrelated diseases.

Acknowledgement

This work was partially funded by BiBiNet project (grant n: H35F21000430002) within the POR-Lazio FESR 2014-2020, by PRIN 2017 (grant n: 20178L3P38), and by Progetto di Ricerca di Ateneo 2021 of Sapienza University of Rome (grant n: RM12117A34663A2C).

References

[1] P. Paci, T. Colombo, G. Fiscon, A. Gurtner, G. Pavesi, and L. Farina, "SWIM: a computational tool to unveiling crucial nodes in complex biological networks," *Scientific Reports*, vol. 7, p. srep44797, Mar. 2017.

[2] M. C. Palumbo, S. Zenoni, M. Fasoli, M. Massonnet, L. Farina, F. Castiglione, M. Pezzotti, and P. Paci, "Integrated Network Analysis Identifies Fight-Club Nodes as a Class of Hubs Encompassing Key Putative Switch Genes That Induce Major Transcriptome Reprogramming during Grapevine Development," *The Plant Cell Online*, p. tpc.114.133710, Dec. 2014.

[3] G. Fiscon, F. Conte, and P. Paci, "SWIM tool application to expression data of glioblastoma stem-like cell lines, corresponding primary tumors and conventional glioma cell lines," *BMC bioinformatics*, vol. 19, p. 436, Nov. 2018.

[4] G. Fiscon, F. Conte, V. Licursi, S. Nasi, and P. Paci, "Computational identification of specific genes for glioblastoma stem-like cells identity," *Scientific Reports*, vol. 8, p. 7769, May 2018.

[5] P. Paci, G. Fiscon, F. Conte, V. Licursi, J. Morrow, C. Hersh, M. Cho, P. Castaldi, K. Glass, E. K. Silverman, and L. Farina, "Integrated transcriptomic correlation network analysis identifies COPD molecular determinants," *Scientific Reports*, vol. 10, pp. 1–18, Feb. 2020. Number: 1 Publisher: Nature Publishing Group.

[6] P. Paci, G. Fiscon, F. Conte, R.-S. Wang, L. Farina, and J. Loscalzo, "Gene co-expression in the interactome: moving from correlation toward causation via an integrated approach to disease module discovery," *npj Systems Biology and Applications*, vol. 7, pp. 1–11, Jan. 2021. Number: 1 Publisher: Nature Publishing Group.

[7] P. Paci and G. Fiscon, "SWIMmeR: an R-based software to unveiling crucial nodes in complex biological networks," *Bioinformatics*, vol. 38, pp. 586–588, Jan. 2022.

[8] J.-D. J. Han, N. Bertin, T. Hao, D. S. Goldberg, G. F. Berriz, L. V. Zhang, D. Dupuy, A. J. Walhout, M. E. Cusick, F. P. Roth, and others, "Evidence for dynamically organized modularity in the yeast protein–protein interaction network," *Nature*, vol. 430, no. 6995, pp. 88–93, 2004.

[9] M. Jansen, S. Yip, and D. N. Louis, "Molecular pathology in adult gliomas: diagnostic, prognostic, and predictive markers," *The Lancet Neurology*, vol. 9, no. 7, pp. 717–726, 2010.

[10] R. M. Young, A. Jamshidi, G. Davis, and J. H. Sherman, "Current trends in the surgical management and treatment of adult glioblastoma," *Annals of translational medicine*, vol. 3, no. 9, 2015.

[11] K. Anjum, B. I. Shagufta, S. Q. Abbas, S. Patel, I. Khan, S. A. A. Shah, N. Akhter, and S. S. U. Hassan, "Current status and future therapeutic perspectives of glioblastoma multiforme (GBM) therapy: A review," *Biomed Pharmacother*, vol. 92, pp. 681–689, Aug. 2017.

[12] R. C. Gimple, S. Bhargava, D. Dixit, and J. N. Rich, "Glioblastoma stem cells: lessons from the tumor hierarchy in a lethal cancer," *Genes & Development*, vol. 33, no. 11-12, pp. 591–609, 2019.

[13] A. Schulte, H. S. Günther, H. S. Phillips, D. Kemming, T. Martens, S. Kharbanda, R. H. Soriano, Z. Modrusan, S. Zapf, M. Westphal, and K. Lamszus, "A distinct subset of glioma cell lines with stem cell-like properties reflects the transcriptional phenotype of glioblastomas and overexpresses CXCR4 as therapeutic target," *Glia*, vol. 59, pp. 590–602, Apr. 2011.

[14] M. L. Suva, E. Rheinbay, S. M. Gillespie, A. P. Patel, H. Wakimoto, S. D. Rabkin, N. Riggi, A. S. Chi, D. P. Cahill, B. V. Nahed, and others, "Reconstructing and reprogramming the tumor-propagating potential of glioblastoma stem-like cells," *Cell*, vol. 157, no. 3, pp. 580–594, 2014.

[15] V. Pecce, A. Verrienti, G. Fiscon, M. Sponziello, F. Conte, C. Durante, L. Farina, S. Filetti, and P. Paci, "FOSL1 in the stemness mechanism: An in vitro study," *Scientific Reports*, 2021.

[16] S. D. Ghiassian, J. Menche, D. I. Chasman, F. Giulianini, R. Wang, P. Ricchiuto, M. Aikawa, H. Iwata, C. Müller, T. Zeller, A. Sharma, P. Wild, K. Lackner, S. Singh, P. M. Ridker, S. Blankenberg, A.-L. Barabási, and J. Loscalzo, "Endophenotype Network Models: Common Core of Complex Diseases," *Scientific Reports*, vol. 6, pp. 1–13, June 2016.

[17] A.-L. Barabási, N. Gulbahce, and J. Loscalzo, "Network medicine: a network-based approach to human disease," *Nature Reviews. Genetics*, vol. 12, pp. 56–68, Jan. 2011.

[18] M. Caldera, P. Buphamalai, F. Müller, and J. Menche, "Interactome-based approaches to human disease," *Current Opinion in Systems Biology*, vol. 3, pp. 88–94, June 2017.

[19] A. R. Sonawane, S. T. Weiss, K. Glass, and A. Sharma, "Network Medicine in the Age of Biomedical Big Data," *Frontiers in Genetics*, vol. 10, Apr. 2019.