ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

Analysis of DNA methylation patterns in cancer samples using SOM

Ignacio Díaz[1], José M. Enguita[1], Diego García[1], Abel A. Cuadrado[1], Nuria Valdés[2] and María D. Chiara[3] *

1- University of Oviedo - Dept. of Electrical Engineering
Edificio Torres Quevedo, módulo 2, Campus de Gijón 33204 - SPAIN

2- Department of Endocrinology and Nutrition, Hospital Universitario Cruces, Bilbao, Bizkaia. Biobizkaia, CIBERER, CIBERDEM, EndoERN

3- Institute of Sanitary Research of the Principado de Asturias
Hospital Universitario Central de Asturias, Oviedo 33011 - SPAIN.

Abstract. By leveraging the SOM algorithm and the extensive epigenomic data from TCGA, this work aims to suggest a valid approach to explore the relationships between epigenetic alterations and PCPG pathogenesis. Additionally, the methodological approach presented here lays the foundation for a potentially valuable analysis tool that can be applied to other cancer types and epigenetic research.

1 Introduction

Pheochromocytomas and paragangliomas (PCPG) are rare neuroendocrine tumors that arise from neural crest-derived chromaffin cells. These tumors can excessively secrete catecholamines, which can lead to hypertension, cardiac arrhythmias and other clinical symptoms. Despite advances in diagnosis and treatment, significant challenges remain in understanding the underlying biology of these tumors. Epigenetic disregulation, and specifically changes in the methylation patterns of DNA [1] have emerged as important molecular mechanisms involved in tumor generation. DNA methylation is a covalent modification that is often associated to gene silencing and may play a key role in activation or inhibition of signaling pathways involved in the development of different cancer types [2]. In this work we propose to analyze the methylation profiles of more than 390K valid CpG sites in 187 PCPG samples using the *self-organizing map* (SOM) algorithm to obtain 187 component planes that are visual epigenetic signatures of the PCPG tumors. We arranged these SOM planes spatially according to their methylation similarities, and labeled the tumors presenting mutations related to pseudohypoxia conditions, involved in PCPG development and progression. The results reveal cluster structure for these tumors, providing evidences of epigenetic mechanisms involved and suggesting our approach as a complementary analysis tool.

*This work is part of Grant PID2020-115401GB-I00 funded by MCIN/AEI/ 10.13039/501100011033. The results shown here are based upon data generated by the TCGA Research Network: `https://www.cancer.gov/tcga`.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

2 Methods

2.1 The SOM algorithm

The self-organizing map (SOM) [3] can be described as a nonlinear, smooth mapping of a high-dimensional input space onto a low-dimensional output space, typically meant for visualization. It consists of S units where each unit i is associated to an r-dimensional prototype vector $\mathbf{m}_i$ in the input space and a position vector on a low dimensional grid, $\mathbf{g}_i$, in the output space. The SOM algorithm trains the prototypes $\mathbf{m}_i$ to learn the distribution of the input data points $\mathbf{x}_k \in \mathbb{R}^r$, while preserving the topology defined by the $\mathbf{g}_i$. It can be divided in two stages: a *competitive* stage, where the closest prototype $\mathbf{m}_c$ to the input vector is obtained:

$$c \quad = \quad \arg\min_i \|\mathbf{x} - \mathbf{m}_i(t)\| \tag{1}$$

and a *cooperative* stage, where $\mathbf{m}_c$ but also its neighbor prototypes are adapted

$$\mathbf{m}_i(t+1) \quad = \quad \mathbf{m}_i(t) + \alpha(t) h_{ci}(t)[\mathbf{x} - \mathbf{m}_i(t)] \tag{2}$$

being $\alpha(t)$ the *learning rate* and $h_{ci}(t)$ a *neighborhood function* between units c and i, which states the amount of adaptation for units that are close to the winner in the topology defined by the $\mathbf{g}_i$ coordinates. A common choice is $h_{ij} = \exp\left[-d(\mathbf{g}_i, \mathbf{g}_j)/\sigma^2\right]$, that measures the gaussian neighborhood between units i and j in the output space, being $d(\cdot,\cdot)$ generally a L_2 or L_1 distance.

2.2 Dealing with large number of samples

The SOM training algorithm involves the computation of distances from the n input samples to the S prototypes. When n is very large —in methylation data n may be in the order of 10^5 CpG sites— the SOM algorithm becomes computationally unaffordable. In this work, the *batch* version, more stable and computationally efficient was used to obtain the prototypes $\mathbf{m}_i$:

$$c(k) = \arg\min_i \|\mathbf{x}(k) - \mathbf{m}_i(t)\|, \qquad \mathbf{m}_i(t+1) \quad = \quad \frac{\sum_k h_{c(k)i}(t) \cdot \mathbf{x}(k)}{\sum_k h_{c(k)i}(t)} \tag{3}$$

To overcome memory requirements, the prototypes can be updated at each epoch using (3) for *batches* of a smaller size n_b, by randomly sampling with replacement from the original dataset, and then averaged with an exponentially weighted moving average (EWMA):

$$\mathbf{m}'_i(t) = \frac{\sum_{k=1}^{n_b} h_{c(k)i}(t) \cdot \mathbf{x}(k)}{\sum_{k=1}^{n_b} h_{c(k)i}(t)} \qquad \mathbf{m}_i(t+1) = \lambda \mathbf{m}_i(t) + (1-\lambda)\mathbf{m}'_i(t) \tag{4}$$

For sufficiently long number of epochs, this bootstrap approach accurately approximates the input data distribution and yields a stable convergence allowing to trade memory demand for iterations in large data samples.

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

3 Results

3.1 The SOM for epigenetics data analysis

The methylation data under analysis involves a large dataset $X \in \mathbb{R}^{n,m}$ with $n = 391529$ methylation levels (β values) of CpG sites, and $m = 187$ PCPG samples from the TCGA database [1]. Prior to SOM training, the β values were transformed using *rank normalization* [4] to obtain a uniform equalized histogram of methylation values. Inspired in [5], the training of the SOM is done shifting the usual role of samples and attributes, so the bases are considered as samples and the tumors are considered as attributes. We trained a 50×50 SOM[2], resulting in $S = 2500$ codebooks $\mathbf{m}_i$, with 187 methylation values each. Each codebook can be seen as a "prototype CpG base" that is indeed an *aggregation* representing a cluster of CpG sites with similar methylation patterns.

3.2 Component planes as epigenetic signatures

The k-th tumor, in turn, may be represented by a *component plane*, in this case a 50×50 image obtained assigning a color according to a color scale for the k-th component of the 2500 prototypes $\mathbf{m}_i$, at the positions of the grid nodes $\mathbf{g}_i$ in the 2D visualization space. The resulting component planes, shown in Fig. 1, are composed of the aggregated methylation levels of the prototypes for the 187 tumors each resulting, thereby, in an "epigenetic signature" of a tumor composed of 2500 methylation values. Since the number of units $S = 2500$ is much smaller than the number of CpG sites, $n = 391529$, this is a form of dimensionality reduction achieved through aggregation. The epigenomic signature we get on each component plane is a smoothed portrait of the methylation activity of a tumor, averaging out detailed variations and preserving the main trends.

Interestingly, the regions in the planes represent sets of CpG sites —and the genes to which they belong— with similar methylations across the m tumors. Thus, they can also be interpreted as *epigenetic maps* of gene locations according to their methylation activity. As a matter of sample, in Fig. 2, the black points represent the location of CpG sites from protocadherin genes. This fact can be used to compare and analyze genes in terms of epigenetic activity.

3.3 Visualization of tumors by similar epigenetic behavior

The m component planes can be treated as feature vectors describing the tumor samples. Using a dimensionality reduction algorithm, such as the t-SNE algorithm [6] we can display the tumors spatially organized in terms of similarity of their component planes, as shown in Fig. 2. After testing several perplexity values we finally considered a low value ($p = 5$) due to the relatively small total sample size (187), and also to reveal detail of clusters with a small number of

[1] GDC TCGA Pheochromocytoma & Paraganglioma (PCPG); Illumina Human Methylation 450 DNA methylation (available at Xenabrowser `https://xenabrowser.net/datapages/`)

[2] Full code and experiment parameters to reproduce the results available in `https://github.com/gsdpi/SOM-DNA-Methylation`

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

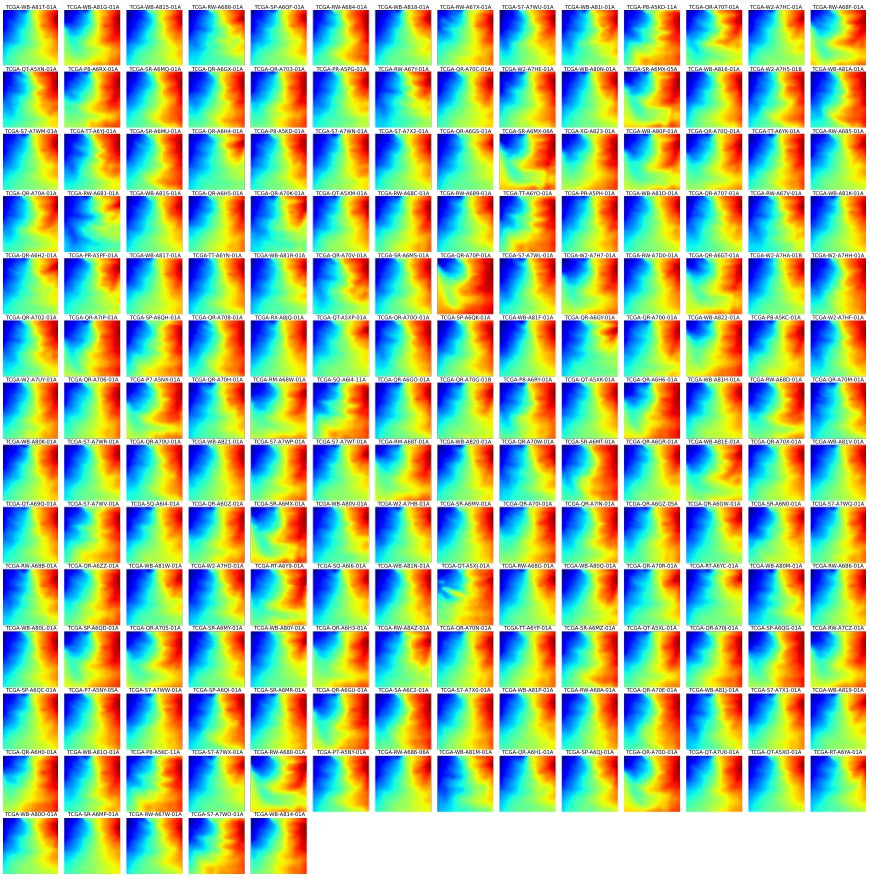

Fig. 1: Methylation component planes for the 187 PCPG tumors. Blue tones reveal low methylation ($\beta \approx 0$) and red tones represent high methylation ($\beta \approx 1$).

tumors. As shown in Fig. 2, the t-SNE method arranges the tumors into groups with similar methylation patterns, which can be visually confirmed by the similar component planes observed within each group. Noting that the proportion of red areas (high methylation) over blue areas (low methylation) in the component plane of a tumor sample is related to the overall level of methylation, a global structure is also found in the map according to the overall methylation levels, with a gradual distribution from low-methylated tumors on the left, to highly methylated tumors on the right.

Also, the location in the t-SNE map of tumors with mutations in the VHL, SDHx, and EPAS1 genes, shown in Fig.3, provides interesting insights. Mutations in these genes disrupt the normal regulation of the hypoxia-inducible factor (HIF) pathway, leading to pseudohypoxic conditions that promote tumor growth, angiogenesis, and progression of PCPG. SDHx appear grouped on highly

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

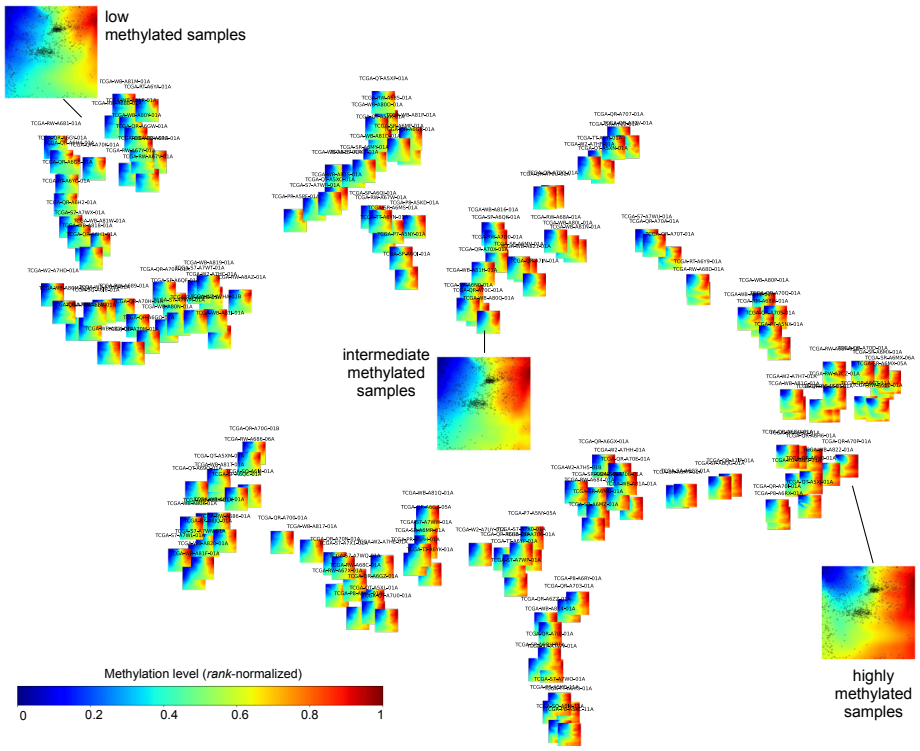

Fig. 2: *t*-SNE map of methylation patterns. Methylation levels grow from left to right.

methylated areas, while VHL and EPAS1 lay together in areas with intermediate methylation. It is known that VHL and EPAS1 are directly involved in the HIF signaling pathway, while SDHx mutations indirectly affect HIF by loss of function of SDH genes. This connection between pseudohypoxia patwhays and methylation patterns, suggests that the proposed approach can be a complementary way of analysis.

4 Conclusions

In this paper we have proposed using SOM to visualize and reduce the dimensionality of methylation data from PCPG tumors. The SOM component planes act as methylation signatures that revealed relationships between the tumors' epigenetic patterns and key genetic mutations like VHL, SDHx, and EPAS1. This SOM-based approach integrating epigenetic and genetic data allows identifying connections between the dysregulated methylation landscapes and genetic signatures of PCPG. It demonstrates the potential of SOM analyses to gain insights into the interplay of epigenetics and genetics in cancer, with potential

ESANN 2024 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium) and online event, 9-11 October 2024, i6doc.com publ., ISBN 978-2-87587-090-2. Available from http://www.i6doc.com/en/.

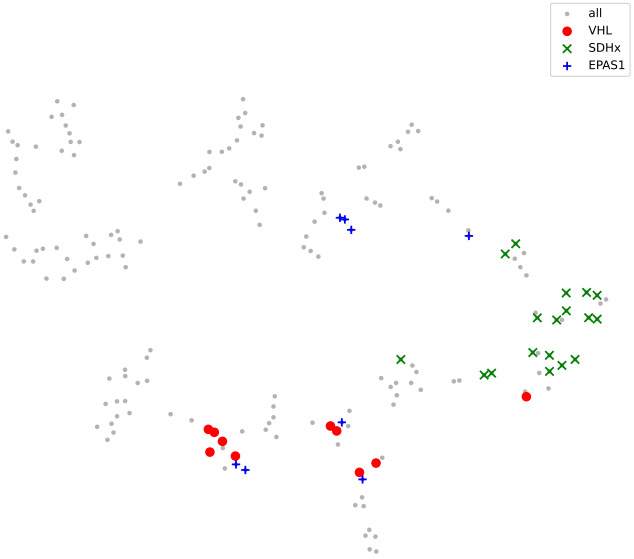

Fig. 3: Location of VHL, SDHx, and EPAS1 in the t-SNE map of methylation patterns

applications in biomarker discovery and personalized treatment development. The possibility to represent tumors with mutations or other phenotypes on epigenetic behavior maps with the proposed approach can help in elucidating PCPG molecular heterogeneity and subtypes, guiding targeted therapies.

References

[1] Peter A Jones and Daiya Takai. The role of dna methylation in mammalian epigenetics. *Science*, 293(5532):1068–1070, 2001.

[2] Shikhar Sharma, Theresa K. Kelly, and Peter A. Jones. Epigenetics in cancer. *Carcinogenesis*, 31(1):27–36, 09 2009.

[3] Teuvo Kohonen. The self-organizing map. *Proceedings of the IEEE*, 78(9):1464–1480, 1990.

[4] A. Szabo, K. Boucher, W.L. Carroll, L.B. Klebanov, A.D. Tsodikov, and A.Y. Yakovlev. Variable selection and pattern recognition with gene expression data generated by the microarray technology. *Mathematical Biosciences*, 176(1):71–98, 2002.

[5] Maria Nikoghosyan, Maria Schmidt, Kristina Margaryan, Henry Loeffler-Wirth, Arsen Arakelyan, and Hans Binder. Sommelier—intuitive visualization of the topology of grapevine genome landscapes using artificial neural networks. *Genes*, 11(7):817, 2020.

[6] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.