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]

contact author: idiaz@uniovi.es

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.

Universidad de Oviedo

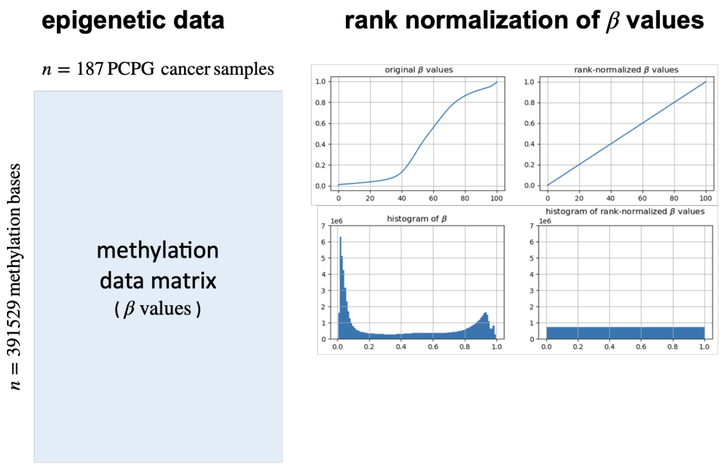
GSDPI grupo de supervisión, diagnóstico y descubrimiento del conocimiento en procesos de ingeniería

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.

Data preparation

- 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 [a].
- Prior to SOM training, the β values were transformed using *rank normalization* to obtain a uniform equalized histogram of methylation values.



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

Batch SOM and bootstrap training

- 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 $\mathbf{m}_i$.
- To overcome memory requirements, the prototypes can be updated at each epoch 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):

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

For a 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.

Conclusions

- 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 relating epigenetic and genetic data allows identifying connections between the dysregulated methylation landscapes and genetic signatures of PCPG.
- Our approach demonstrates the potential of SOM analyses to gain insights into the interplay of epigenetics and genetics in cancer, with potential 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.

Acknowledgements

GOBIERNO DE ESPAÑA MINISTERIO DE CIENCIA E INNOVACIÓN AGENCIA ESTATAL DE INVESTIGACIÓN

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`.

DNA methylation planes for 187 tumors

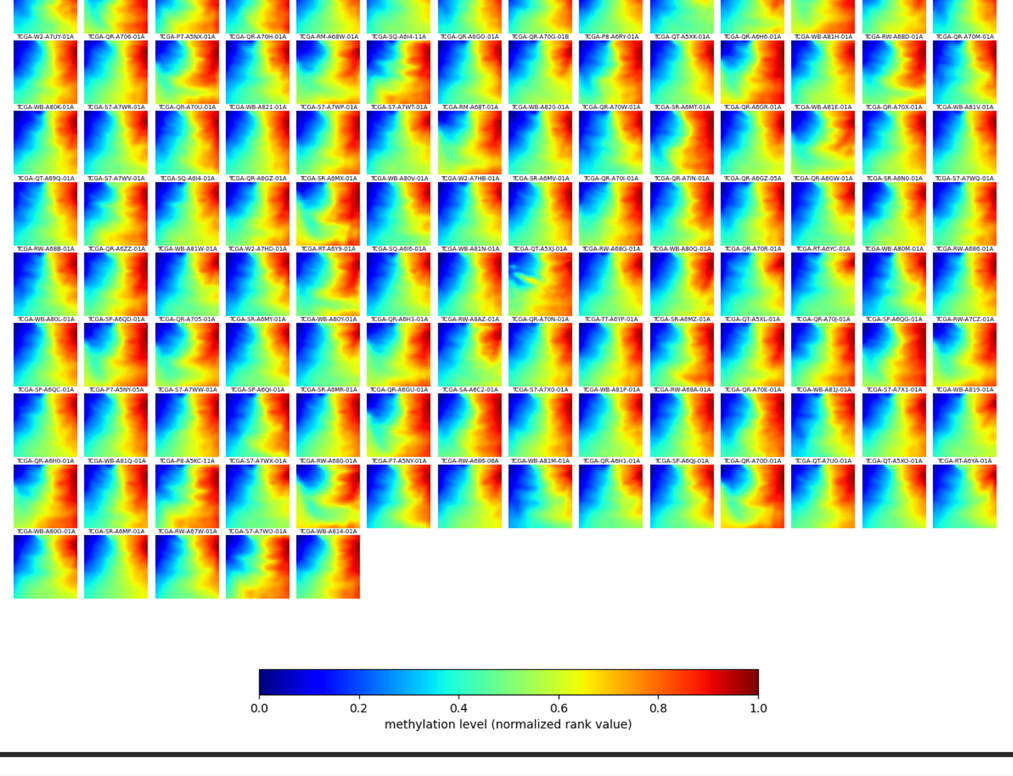
SOM of DNA methylation data

- 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[a], 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.

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

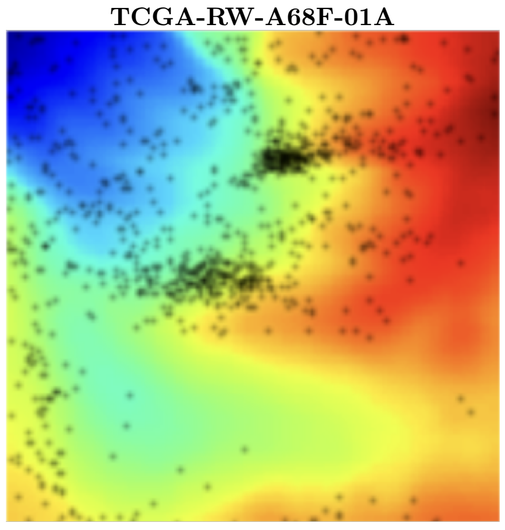
Methylation component planes for the 187 PCPG tumors

Blue tones reveal low methylation ($\beta \approx 0$) and red tones represent high methylation ($\beta \approx 1$).

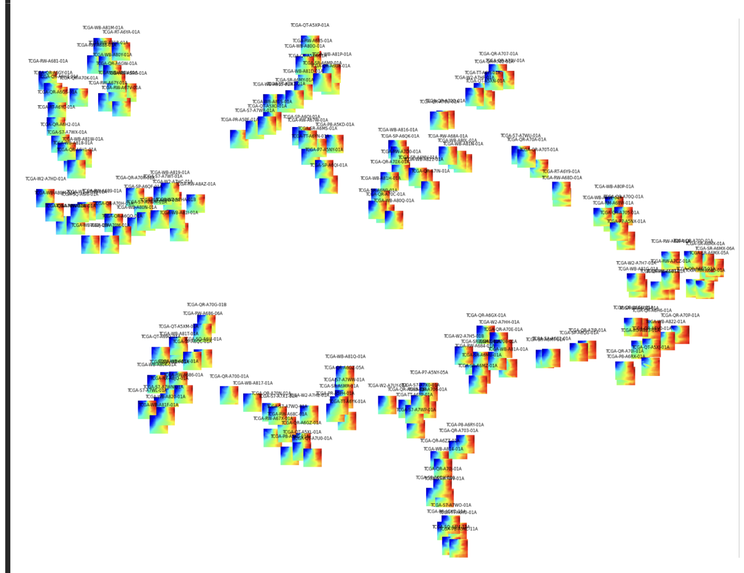


Interpretation of component planes

- Component planes are composed of the aggregated methylation levels of the prototypes for the 187 tumors. Each component plane is an epigenetic signature of the tumor, composed of 2500 (50×50) representative methylation values.
- The 2500 values summarize the overall amount of n=391529 methylation values of all CpG sites. This is a form of dimensionality reduction through aggregations.
- Regions in the planes represent CpG sites with similar methylation values across the 187 tumors. Since each CpG site belongs to a gene, the areas spanned by each gene can be displayed in the component plane.
- In this case, black points represent the locations of CpG sites from protocadherine genes, resulting in recognizable patterns potentially deserving analysis. This can be used to compare and analyze genes in terms of epigenetic activity.

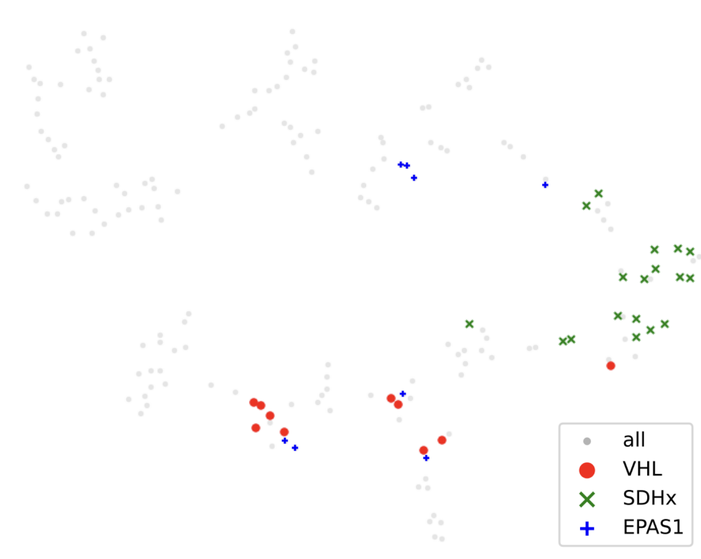


t-SNE of tumor epigenetic signatures



t-SNE map based on DNA methylation

- The m component planes can be treated as feature vectors describing the tumor samples.
- Using the t-SNE algorithm we can display the tumors spatially organized in terms of similarity of their component planes.
- The t-SNE arranges the tumors into groups with similar methylation patterns, which can be visually confirmed by the similar component planes observed within each group.
- 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.



Mutations related to hypoxia-inducible factor (HIF)

- The locations in the t-SNE map of tumors with mutations in the VHL, SDHx, and EPAS1 genes, provide 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 methylated areas, while VHL and EPAS1 lay together in areas with intermediate methylation.
- 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 reveals a connection between pseudohypoxia patwhays and DNA methylation patterns.