# Towards using Hopfield Networks for the Identification of Therapeutic Targets for Cancer

Olger Calderón-Achío*† and Francisco Siles Canales†
*Computer Science Postgraduate Program, School of Computer Engineering
Technological Institute of Costa Rica
San José, Costa Rica
†Pattern Recognition and Intelligent Systems Laboratory, School of Electrical Engineering
University of Costa Rica
San José, Costa Rica

***Abstract*—Cancer currently constitutes both a national and worldwide health problem for the human population. General scientific effort has been directed towards improving targeted and personalized therapy, which are characterized by making an intelligent use of the patient's genomic blueprint in order to make more informed treatment decisions. NIH's project, Genomic Data Commons (GDC), provides an openly available online data repository which stores great diversity of cancer related data from different cases (patients). This study leverages patterns found in gene expression data for the identification of potential therapeutic targets. Data was restricted to highly expressed genes from cases of breast invasive carcinoma. A Hopfield network (a type of recurrent neural network) was used for clustering purposes. Preliminary tests subdivided the cases into two clusters, where four highly expressed genes better characterized one cluster from the other. Some of these genes are reported in the literature, indeed, as biomarkers for breast cancer. These results suggest that attractor states of Hopfield networks might provide means for discovering or better understanding potential therapeutic targets when treating a particular cancer subtype.**

***Index Terms*—Cancer, Therapeutic Targets, Neural Networks, Machine Learning, Pattern Recognition, Bioinformatics.**

## I. Introduction

Cancer has been a mainstream human race health problem for years. It mainly causes an erratic behavior at the genetic level in the cells, which in turn triggers abnormal cell growth and proliferation through the tissues. Currently, it constitutes one of the main causes of death both nationally (Costa Rica) and globally. Current trends about mortality rates are really alarming [1], which justifies interdisciplinary and transdisciplinary efforts involving research fields like Computer Science, Electric Engineering, Mathematics, Statistics, Pharmacology, Biotechnology, Genetics, among many others; in favor of developing advances in the state of the art.

A fundamental idea present in state of the art literature is that cancer treatment, in general, becomes more effective if genetic and epigenetic features of patients are exploited for making more informed decisions. This is usually called "targeted therapy" [2], because it aims at disrupting specific molecular agents (certain proteins or miRNA for example) that are playing some regulatory role in cancer evolution/proliferation. Drugs need to target these agents which vary across patients and tumor types. These methods can be easily contrasted with traditional cytotoxic chemotherapy, which has proven to be more harmful and riskfull for the patient. The previous statement could be easily explained by citing some adverse effects of chemotherapeutic compounds like: myelosuppression (bone marrow suppression), mucositis (ulceration of the mucous membranes lining the digestive tract) and alopecia (hair loss). In favor of improving the "targeted" approach, data analysis and good prediction models are needed [3], which constitutes a current research trend in the field and it is the aim of this study.

There are several factors that also motivate and contribute to this research effort:

- Data science advancements and state of the art machine learning techniques.
- Latest generation genomic and sequencing techniques in molecular biology and public repositories of genomic information.
- Collaborative interdisciplinary and transdisciplinary efforts.

This paper describes some preliminary tests that illustrate the potential of the Hopfield networks [4] to model cancer as a dynamic system, in which its stable states (called attractors) might prove useful for detecting and understanding biomarker genes (which could also serve as therapeutic targets). In particular, the authors of [5], already explored the capacity of Hopfield networks for clustering data into cancer subtypes, whose performance was comparable to other algorithms like k means. However, GDC data was not used and the aim of the authors was not exactly at determining therapeutic targets.

In section 2, the methodology is described, focusing on aspects like data acquisition, treatment and the modeling approach to the problem. In section 3, some preliminary results are showed. In section 4, some other ideas that are worth exploring as future research are mentioned. Finally, in section 5, conclusions are shared.

## II. Methodology

### A. Used Database

Given its current popularity, diversity and quality, The Cancer Genome Atlas (TCGA) database (now part of the

NIH's Genomic Data Commons project[1]) was chosen as the data source for this study. The TCGA project alone has been able to gather and organize data of around 11.000 patients involving 33 different tumor types (including 10 rare types of cancer). The project was funded by the US government and supervised by the National Cancer Institute's Center of Cancer Genomics. The total data is more or less 2.5 petabytes.

In summary, the GDC DB is a property oriented graph database which organizes a great diversity of files. Information could be classified as genomic, clinical or biospecimen depending of its nature. Genomic data categories include: sequencing data, copy number variation, DNA methylation, simple nucleotide variation, genomic profiling and transcriptome profiling. Each data category uses its own data file formats (for example .tsv for transcriptome profiling and .bam for aligned reads). Also each data category can be further subdivided into different data types. Transcriptome profiling has 3 subtypes: Gene Expression, Exon Expression and miRNA Expression. For this study only Gene Expression data was used.

### B. Chosen Cancer Cases

Given the results of [3] it could be argued that some cancer studies might be restricted to particular cancer types without loss of generality (because most tumors of the same organ tend to cluster together by their genome-wide characteristics). This gave this study the opportunity to concentrate on a single tumor type before trying to generalize results to other tumor types. Cases of breast cancer proved to be a good subset of the total data for use in this study. In general, it is one of the most studied cancer types and it also has relevance in the national context. GDC contains more than 1000 cases for breast invasive carcinoma.

### C. Used Data Type

Results in [5] suggest that gene expression data alone gives good results for characterizing cancer sub types. This can be intuitively argued, given that gene expression is the final result of complex interactions of gene regulatory networks. So overall it could be considered a good predictor. However, as indicated in [3] even better results could be obtained if expression data is used alongside other *-omics features like single number variation, mutation data and methylation values (all provided by the TCGA).

For the scope of this project it was convenient to initially work with expression data. Gene expression data is modeled as a matrix of numeric values where patient cases are observations (rows) and genes are features (columns). Let $M$ be a gene expression matrix, the value at row $i$, and column $j$ (labeled as $M_{ij}$) represents a relative value of how much gene $j$ is expressed in case $i$. Higher values indicate a higher protein synthesis rate for that gene. The data is high dimensional as there are thousands of genes that are profiled for gene expression values.

TABLE I
ARGUMENTS PASSED TO GDCQUERY FUNCTION

| Name of Parameter | Argument |
|---|---|
| project | "TCGA-BRCA" |
| data.category | "Transcriptome Profiling" |
| data.type | "Gene Expression Quantification" |
| workflow.type | "HTSeq - Counts" |

### D. Computational Platform and Fetching of the Data

The implementation platform for the Hopfield network's clustering using TCGA data was the R Statistical Computing Environment.

The software package **TCGABiolinks** [6] from the Bioconductor project[2] was chosen for accessing and fetching the data from the GDC/TCGA via its web API. Other two Bioconductor packages with similar functionality where also evaluated (**RTCGA** and **RTCGAToolbox**), however **TCGABiolinks** was favored in the end because of its wider use statistics and higher commit rate at its GitHub repository.

The fetching of the data was relatively simple. The arguments passed to the **GDCQuery** function are showed in Table I.

### E. Data Cleanup, Exploration and Preprocessing

After being loaded and prepared, the matrix of data contained the gene expression (numerical) value for 60488 genes across 1222 cases. Then some basic descriptive measures were calculated. In particular, the distribution of the means of expression values for each gene was calculated. The distribution was extremely right skewed. The majority of genes showed expression rates that were extremely tiny in comparison to other ones. Just to give an idea, the maximum value was 554437.8 and the median was at 2.3.

Given that the magnitude of the differences between some genes was so huge, it was opted to apply a $log$ transformation to the data.[3] After this transformation, the distribution of mean values for genes was still right skewed. However the differences between genes were a lot easier to visualize. See Figure 1.

The next step was particularly important. It was decided to limit this preliminary study to only those highly expressed genes (with a mean value higher than 5). There were two reasons. The first one is that it can be hypothesized that, indeed, it is possible to characterize tumors in an effective way using only those genes that are highly expressed (even if there are also inactive genes that play a transcendental role in cancer development). The second reason is purely methodological, as the analysis of all the genes was posing time and resource limitations. To illustrate this, Hopfield networks need

[1] https://gdc.cancer.gov/

[2] Bioconductor brings a repository for packages specifically aimed at the analysis of biological data. See http://www.bioconductor.org/

[3] One unit is added in order to avoid $log(0)$ values, given that are genes that are not expressed at all.

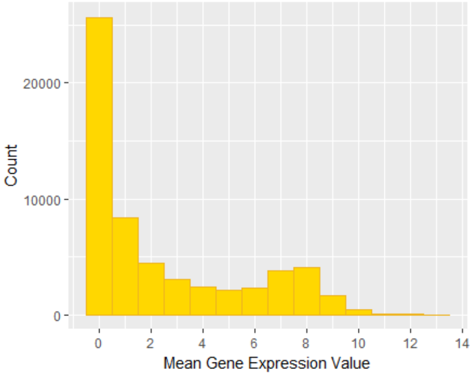

Fig. 1. Distribution of gene expression means for TCGA-BRCA cases after applying a *log* transformation to the data.

to allocate space for a matrix in $\Theta(n^2)$, where $n$ is the number of dimensions. Using all genes results approximately in 25Gb of memory consumption. For the immediate resources, it was unfeasible, even making use of packages for memory mapped files like **Bigmemory**.

Finally the values where centered and scaled in the same way that it is described on [5]. The resulting matrix contained only 8236 rows (genes) by 1222 columns (cases), making the analysis feasible for a mainstream computer.

### *F. Hopfield Networks*

A Hopfield network[4] is a type of recurrent neural network with $n$ neurons, where $n$ is also the number of dimensions of the data. Each neuron acts as an input and output unit, which results in a matrix $W$ of approximately $n^2$ weighted links representing the network (basically, a complete graph). The network can be trained using Hebbian learning with the following formula:

$$W = \frac{1}{m} \sum_{k}^{m} p_k p_k^T - I \tag{1}$$

Here, $m$ is the number of patterns used to train the network. The formula calculates the sum over the outer products of the pattern vectors $p_k$, then normalizes values and sets the diagonal with zero values. The result, the matrix $W$, contains the numerical values of the network links, where $W_{ij}$ is the weight of link connecting neuron $j$ to neuron $i$.

After being trained, the network can be used to "recall". This means that given a new input pattern $p$, the network can associate the pattern $p$ with the closest pattern $q$ that it remembers. This ability to reconstruct stored patterns from similar input is the reason Hopfield networks are considered a way of implementing associative memory.

The "recall" phase works as follows. The process is divided into different time steps $t$. At any time step $t'$, each neuron has a binary state $s_i \in \{-1, 1\}$, $i \in 1, \ldots, n$. The states for the next time step are calculated as follows:

$$s_i^{(t'+1)} = sgn\left(\sum_{j}^{n} W_{ij} s_i^{(t')}\right) \tag{2}$$

The $sgn$ is defined in the following way:

$$sgn(x) = \begin{cases} +1 & x > 0 \\ -1 & x \leq 0 \end{cases} \tag{3}$$

Given an input pattern $p = (p_1, \ldots, p_n)$ for the "recall" phase, the initial state of the network is calculated as $s^0 = (sgn(p_1), \ldots, sgn(p_n))$. Then, equation 2 is run for the $t$ iterations. It is guaranteed that if $t \to \infty$, then the state of the network will converge to a stable state with minimum energy. Describing the definition of the energy function is outside the scope of this article. However it should be noted that these "energy" and "stable state concepts" is what allows Hopfield networks to model the evolution of cancer as a dynamic system, where the stable states are where gene expression levels tend to converge. In [5] it is argued that stable states correspond to gene signatures that characterize different cancer subtypes. In this study we take a similar approach but using only high expressed genes to describe these signatures.

For implementation purposes, an example code for the course "Foundations of Neural and Cognitive Modeling" of the University of Amsterdam was used as base[5]. Some minor modifications were made. These include substituting uses of base **matrix** R objects with **Matrix** S4 objects, more suited for high computation and efficient representation. Also, the outer product of equation 1 is calculated with the function **tcrossprod**, which overall is more efficient for this case.

The tests performed consisted on first training a Hopfield network using all cases of breast invasive carcinoma. Then the "recall" phase was effectuated for each pattern individually using a maximum of 100000 time steps or iterations. Numbers less than this proved to be insufficient for the majority of cases to converge to a stable state.

## III. Results

The tests effectuated for the study are preliminary in nature, and lack a formal experimental design. However they proved to be useful for gaining some insight into the potential of the methods described previously.

After the tests, 384 cases converged to a particular attractor state $a_1$, 448 cases converged to a second attractor $a_2$, and the rest of the 387 cases did not converge to a stable state (even more than 1000000 iterations are needed for that). From the look of the patterns it can be seen that those 387 cases that did not converge, after some more iterations, would have probably converged to either $a_1$ or $a_2$.

One kind of unexpected result was that $a_1$ and $a_2$ were vectors complementary to each other. That is, those components equal to +1 in $a_1$ were equal to -1 in $a_2$ and vice-versa. Like opposite poles. It remains to be confirmed if this

[4] For a comprehensive introduction to the Hopfield model and other associative models, please refer to [7].

[5] http://www.illc.uva.nl/LaCo/clas/fncm13/assignments/computerlab-week4/.

TABLE II
POTENTIALLY DETECTED THERAPEUTIC TARGETS

| Gene Ensembl Code | Relation to Cancer |
|---|---|
| GSTP1 | Displays antiapoptotic activity by interacting with c-*Jun* $NH_2$-terminal kinase, a key regulator of apoptosis. |
| CRYAB | Downregulated for certain clinical luminal B subtype breast tumors. |
| KRT17 | Blood-Based biomarker for metastatic breast cancer. |
| PPP1R1B | Upregulated when breast tumor cells are co-cultured with ADSCs cells. |

behavior is expected when only two attractors are generated by the network.

A quick inspection to these 2 attractor states revealed that in either case, there were 4 components that differed from the rest (8233). In the case of $a_1$, 8233 components converged to -1 and 4 to +1. In the case of $a_2$, 4 converged to -1 and 8233 to +1.

A possible interpretation is that those 4 characteristic components (that correspond to 4 genes) are good discriminators in order to cluster the cases into two groups. The first group of cases tend to evolve the gene expression values for those genes towards very high values, while in the second group this is not the case. It would require further experimentation and feedback from experts in the biological subject to see if that interpretation is completely valid.

The 4 distinctive genes are coded as GSTP1, CRYAB, KRT17, PPP1R1B in the Ensembl database. Each one of these is mentioned in the literature to have showed an altered behavior in the context of breast cancer [8]–[11] (see Table II), which in turn suggests them as possible candidates for being biomarkers or therapeutic targets. From these, GSTP1 and KRT17 appear to show major relevance. These results show the potential of using Hopfield networks stable states to infer useful information that might be used in the development of new targeted drugs for other less studied cancer types.

## IV. FUTURE WORK

This preliminary study was restricted strictly to cases of breast invasive carcinoma from The Cancer Genome Atlas. In order to better assess the effectiveness of the tested method, other cancer types could be explored.

Also there was a cutoff value ($mean \geq 5$) for working only with high expressed genes. The value 5 was chosen conveniently because it made the study feasible for the computing resources at hand. Having more computing resources could expand the study to consider a higher number of genes o even the complete genome, however, it might be wiser to opt for a better selection of features (e.g, considering genes with higher expression variance or genes that code for transcription factors).

There is also great interest on exploring the properties of the Hopfield network attractors for modeling cancer as a dynamic system. Could it be possible to infer new Hopfield network attractors stable states which are non-cancerous? If it is possible, then which genes should be specifically targeted to maximize the probabilities of a certain state of the network to converge to a non-cancerous state? These questions will be probably explored in a future thesis research project.

## V. CONCLUSIONS

This study lays at the intersection of several fields: machine learning, statistics and molecular biology. It is clearly illustrated how interdisciplinary and transdisciplinary efforts can aid in the comprehension, prevention and treatment of cancer. The Hopfield network models cancer as a dynamic system on which gene expression levels tend to converge to stable states. These properties could aid us on the identification of potential therapeutic targets, which could be used for improving treatments. Also it should be kept in mind that these models, if proven good in silico, should be tested in vitro as a next step. The long term objective is to improve the quality of life of persons in a noticeable way.

## REFERENCES

[1] *Plan Nacional para la Prevención y Control del Cáncer 2011 - 2017*, 2012.

[2] C. Sawyers, “Targeted cancer therapy.” *Nature*, vol. 432, no. 7015, pp. 294–297, 2004.

[3] E. Taskesen, S. M. H. Huisman, A. Mahfouz, J. H. Krijthe, J. De Ridder, A. Van De Stolpe, E. Van Den Akker, W. Verheagh, and M. J. T. Reinders, “Pan-cancer subtyping in a 2D- map shows substructures that are driven by specific combinations of molecular characteristics,” *Nature Publishing Group*, 2016.

[4] J. J. Hopfield, “Neural networks and physical systems with emergent collective computational abilities,” *Proceedings of the national academy of sciences*, vol. 79, no. 8, pp. 2554–2558, 1982.

[5] S. R. Maetschke and M. A. Ragan, “Characterizing cancer subtypes as attractors of Hopfield networks,” *Bioinformatics*, vol. 30, no. 9, pp. 1273–1279, 2014.

[6] A. Colaprico, T. C. Silva, C. Olsen, L. Garofano, C. Cava, D. Garolini, T. S. Sabedot, T. M. Malta, S. M. Pagnotta, I. Castiglioni *et al.*, “Tcgabiolinks: an r/bioconductor package for integrative analysis of tcga data,” *Nucleic acids research*, p. gkv1507, 2015.

[7] R. Rojas, *Neural networks: a systematic introduction*. Springer Science & Business Media, 2013.

[8] L. Federici, C. L. Sterzo, S. Pezzola, A. Di Matteo, F. Scaloni, G. Federici, and A. M. Caccuri, “Structural basis for the binding of the anticancer compound 6-(7-nitro-2, 1, 3-benzoxadiazol-4-ylthio) hexanol to human glutathione s-transferases,” *Cancer research*, vol. 69, no. 20, pp. 8025–8034, 2009.

[9] M. P. Goetz, K. R. Kalari, V. J. Suman, A. M. Moyer, J. Yu, D. W. Visscher, T. J. Dockter, P. T. Vedell, J. P. Sinnwell, X. Tang *et al.*, “Tumor sequencing and patient-derived xenografts in the neoadjuvant treatment of breast cancer,” *JNCI: Journal of the National Cancer Institute*, vol. 109, no. 7, 2017.

[10] A. Berghuis, H. Koffijberg, J. Prakash, L. W. Terstappen, and M. J. IJzerman, “Detecting blood-based biomarkers in metastatic breast cancer: A systematic review of their current status and clinical utility,” *International journal of molecular sciences*, vol. 18, no. 2, p. 363, 2017.

[11] E. Koellensperger, L.-C. Bonnert, I. Zoernig, F. Marmé, S. Sandmann, G. Germann, F. Gramley, and U. Leimer, “The impact of human adipose tissue-derived stem cells on breast cancer cells: implications for cell-assisted lipotransfers in breast reconstruction,” *Stem Cell Research & Therapy*, vol. 8, no. 1, p. 121, 2017.