INVESTIGATING THE CHANGE OF THE HIERARCHICAL PATTERN OF GENE EXPRESSION IN THE NORMAL AND PARKINSON'S BRAIN USING A COMBINATORIAL OPTIMIZATION BASED UNSUPERVISED CLUSTERING METHOD

Mou'ath HOURANI[1], Pritha MAHATA[2], Ibrahiem M. M. El EMARY[3]

Previous works on Parkinson's disease (PD) mainly focused on genes differentially expressed between the anterior and the posterior sections of the brains of a normal mouse and the one with PD. However, no work has been done in finding a hierarchical pattern of gene expression between the different regions of a brain. Such a hierarchy is useful to locate genetic specializations within a normal brain, thus in analyzing how brain infirmities affect these specializations. We use a recently proposed method of robust hierarchical clustering using arithmetic-harmonic cut to construct the hierarchical relation between different regions of the brain. Then, we show how similar regions of the normal and PD brain differ in gene expressions, indicating a functional variation due to Parkinson's disease in a few high-level clusters of brain regions.

Keywords: Parkinson's disease, gene expression, combinatorial optimization, unsupervised clustering.

1. Introduction

Many countries are increasingly witnessing the effect of Parkinson's disease (*PD*) in a majority of their old population. It is a progressive neurodegenerative disease characterized by continual tremors, rigidity of the limbs, slowness of movement and difficulty with balance and coordination. Works from literature like [1] targeted mice for finding the genetic markers of *PD*. They created PD model by administration of toxic doses of methamphetamine (MA) to C57BL/6J mice [1]. At the doses used by [1], the mouse model of PD has been reported to have substantial loss (45%) tyrosine hydroxilase-positive

[1] Assistant Professor, Faculty of Information Technology, Al Ahliyya Amman University, Amman, Jordan, mouath.hourani@yahoo.com
[2] Research Fellow, School of Information Technology and Electrical Engineering, The University of Queensland, Queensland, Australia, Pritha.Mahata@newcastle.edu.au
[3] Associate Professor, Information Technology Deanship, King Abdulaziz University, Jeddah, Kingdom of Saudi Arabia, omary57@hotmail.com

dopaminergic cells in the substantia nigra (a part in the brain), as well as destruction of dopaminergic nerve terminals in the neostriatum.

We consider the microarray data provided by [1]. This data consists of 9, 000 genes and 80 experiments (40 experiments from the normal and 40 experiments from the MA-treated mouse). In fact, brains from both normal and MA-treated mice were divided into 40 voxels (cubic 3D image elements) by slicing each brain into 10 coronal sections and cutting each slice again into four voxels. The work of [1] reports that both normal and PD brains have striking lateral symmetry, i.e., both brains have similar expression in their left and right halves. Furthermore, they compared left (right) brain of the PD mouse with the left (right) brain of the normal one and found that there is a significant difference in expression between the genes repressed or induced in the MA brain.

However, the work in [1] mainly considers genes which distinguishes between the whole anterior (20 voxels) part from the posterior part (20 voxels) of the normal and the MA brains. This motivated us to concentrate on identifying parts of the normal brain which are genetically similar (with respect to the expressions of the genes) and analyze if they are also functionally similar. In this paper, we employ a recent combinatorial optimization based hierarchical clustering method [33] on the voxels of the normal brain. Notice that the use of this clustering algorithm over popular methods like agglomerative hierarchical clustering, *k*-means, etc for this purpose is motivated by the success of [33]'s clustering results on diverse datasets like the dataset of 84 Indo-European languages [35], National Cancer Institute's microarray data of 64 cancer cell-lines [36], etc. Then, we find the genetic signatures for such partitions of the normal brain, i.e., we select the genes which are differentially expressed in the two sides of each node of the tree. The strongly contrasting expressions of a huge number of genes in the first few partitions lead us to infer that such unsupervised partitioning really cluster regions of the brains which are genetically similar. Furthermore, this allows us to analyze how the genes in these signatures are affected in the similar voxels of the MA-treated mouse brain. This approach seems to be very logical in the sense that brain activity may not always depend on spatially co-located regions in the brain. To see deeper, we compare three significant clusters of voxels from the normal and the PD brain and show the genes which make these regions different in the two brains. One of the regions includes all voxels from the rostral cerebellum, a few from septo-striatum and diencephalon.

2. Methods

Our approach consists of employing a recently developed combinatorial optimization based method for hierarchical clustering to find a hierarchical structure within the 40 voxels normal mouse's brain. Then in order to explore the

relations that arise within the hierarchical tree structure, we use a statistical approach that selects genes supporting the hierarchy of the samples in the normal brain.

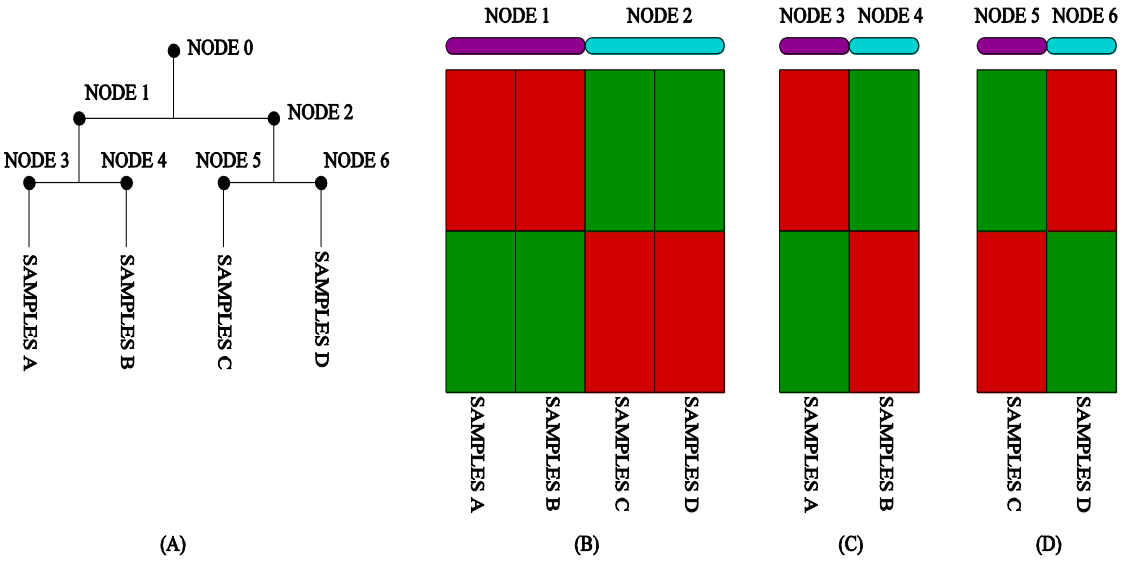

Fig.1 Example of the hierarchical genetic signature of a classification tree. (A) An example of hierarchical clustering of 4 sets of samples, given by SamplesA, SamplesB, SamplesC, and SamplesD. (B) Genes that explain the division of Node0 into Node1 and Node2. (C) Shows the genes distinguishing between the sets Node3 and Node4. (D) Shows the genes distinguishing between the sets Node5 and Node6.

This is illustrated using Fig. 1. In Fig. 1A, we show an example of hierarchical clustering of 4 sets of samples, given by {SamplesA, SamplesB, SamplesC, SamplesD}. We call the union of these sets as Node0. Similarly, Node1 (Node2 resp.) denotes the union of the sets SamplesA and SamplesB (SamplesC and SamplesD resp.). Finally, Node3, Node4, Node5 and Node6 refer to SamplesA, SamplesB, SampleC and SamplesD, respectively. Now we proceed to select relevant genes distinguishing the siblings in the tree. Fig. 1B shows genes that "explain" the division of Node0 into Node1 and Node2. We repeat this procedure over the new sets created. Thus Fig. 1C (Fig. 1D) shows the genes distinguishing between the sets Node3 and Node4 (Node5 and Node6 resp.). We stop when we cannot find significant evidences within the data set to continue with this procedure, or if there are too few samples to evaluate the quality of the results. This analysis gathers the information characterizing each group of samples (SamplesA, SamplesB, SamplesC and SamplesD). Also, at the nodes further from the leaves (say at Node0), it tells us which genes have similar expressions in

SamplesA and SamplesB (which are clustered together in Node1), while having an altogether different expressions in both *SamplesC* and *SamplesD* (which are together in a different partition, namely Node2). The data at each of the non-leaf nodes can be visualized as in Figs. 1B, C and D.

This methodology requires reliable cluster approaches and competent gene selection method. In Mahata et al. (2006) [33] one novel approach for the clustering was designed and tested using the Indo-European languages data set and NCI60 cancer data set. The above paper shows the robustness of the clustering method used here, when compared with other clustering methods previously reported in the literature and applied to the same data sets. In applying the algorithm, we follow the following steps:

- The algorithm used in this paper poses the clustering problem as a graph optimization problem. It uses a novel objective function that performs very well in diverse types of datasets.
- It starts with forming a distance matrix for a set of objects and computes a weighted graph in which vertices represent objects and edges are weighted by the distance between the corresponding vertices.
- Then the objective function tries to obtain a solution whose fitness is maximal and proportional to the sum of the weights on the edges between two sets of vertices, and to the sum of the reciprocals of the weights on the edges inside the sets. This is denoted as *arithmetic-harmonic cut*.
- The recursive application of such cuts generates a tree-based classification of the data.

While the primary concern was the classification of microarray data, the algorithm was also tested under two different datasets: (a) a dataset for 84 Indo-European languages, and (b) a dataset for 60 cancerous cell-lines (NCI60) to explain the robustness of the approach and validating it in different domains [33]. The detailed clustering approach is explained below.

2.1. Combinatorial Optimization based Unsupervised Hierarchical Clustering Method

We apply an unsupervised hierarchical clustering method to establish the hierarchical pattern of gene expression within the voxel scheme of a normal mouse's brain (gene expression data of 40 voxels from [1]). In Mahata et al. (2006) [33] one novel approach for the clustering was designed and tested using the Indo-European languages data set and NCI60 cancer data set. The above paper shows the robustness of the clustering method used here, when compared with other clustering methods previously reported in the literature and applied to the same data sets. In case of Indo-European languages, there are historical and

archeological evidences, which provide good expectation about the correct solution and more importantly it was easier to detect gross and medium size error. Again, in case of NCI 60 dataset, the labels of different samples were known a priori.

The method of [30] for bi-partitioning the vertices of a given graph, maximizing a given objective function (Equation 1) to yield maximum inter-cluster dissimilarity and minimum intra-cluster dissimilarity is called Arithmetic-Harmonic Cut. They use this method within a top-down procedure, which recursively bi-partitions the samples in the dataset. In the dataset under consideration in this paper, we stop bi-partitioning when we cannot find significant evidences within the data set to continue with this procedure (the expression levels of the two partitions are not having high contrast), or if there are too few samples to evaluate the quality of the results.

Formally, we create a complete, weighted graph $G(E, V, W)$ without self-loops where V is the set of vertices (corresponding to the voxels in our case). The weight of any edge e is a positive integer number (i.e. $w(e) > 0$) representing the distance or some measure of dissimilarity between a pair of vertices. In this case the weight of an edge (i, j) between the vertices i, j is given by the *Pearson correlation* based distance.

We consider that any partition of the set V of voxels into two non-empty subsets S and $V \setminus S$ also generates a partition of the set E of edges in two sets E_{in} and E_{out}. The set $E_{out} \subset E$ is the subset of edges that link two vertices of different sets, i.e., a vertex in S and a vertex in $V \setminus S$. Similarly, $E_{in} = E \setminus E_{out}$ is the set of edges connecting vertices within the same subset of voxels. The partition of our interest is defined as the one that maximizes the following objective function

$$F = \left(\sum_{e \in E_{out}} w(e)\right)\left(\sum_{e \in E_{in}} 1/w(e)\right) \quad (1)$$

It turns out that solving the above optimization problem algorithmically is APX-hard [27]. Thus, we use a meta-heuristic, so-called, *memetic* algorithm for solving AH-Cut. Memetic algorithms provide a population-based approach for heuristic search in optimization problems. Basically, they combine local search heuristics with crossover operators used in *Genetic Algorithms* [24]. The essence of our algorithm is similar to the work of Merz and Freisleben in [28] for Graph Bi-partitioning. The difference in our algorithm from that of [28] arises from the fact that we need to remove the constraint of equal partitioning of the graph. The method consists of three main procedures: (a) a differential greedy algorithm (a modification of the algorithm in [29]) for initialization of a set of solutions for AH-Cut, (b) a differential greedy crossover (a modification of the algorithm in

[28]) for evolution of the population, and (c) a variable neighbourhood local search (see [30]) to improve the newly generated solutions.

First of all, we use a ternary tree for population as in [25] and keep two solutions at each node of this tree. One solution is the best obtained so far at the node, called *pocket solution* and the other one is the *current* solution. Essentially, if we generate a current solution by recombination or local search which is better than the pocket solution, we swap this *current* solution with the *pocket* solution. Furthermore, each parent node of the tree must have better *pocket* solution than its children's *pocket* solution. Similar tree structures were previously advocated successfully in various combinatorially hard problems (e.g., see [31, 25, 32, 34]).

First we initialize the population by using a variant of the differential greedy algorithm used in [29] for the GRAPH BI-PARTITIONING problem. This scheme has a bottom-up approach. It first randomly chooses two vertices v, v' from the set V and puts one in S and the other in $\bar{S}$. For each of the remaining vertices u in $V := V \setminus \{v, v'\}$, we compute the sum of the weights from the vertex to the set S and $\bar{S}$ and call this distance $dist(S, u)$ and $dist(\bar{S}, u)$ respectively. These vertices are then sorted in an increasing order according to the metric $dist(\bar{S}, u) - dist(S, u)$ in list L. Unless the set V is empty, we choose whether a vertex will be inserted in set S or in $\bar{S}$ by tossing a coin. If the chosen set is S, we pop a vertex u from the bottom of the list L and update $S := S \cup \{u\}$. We also update the list of remaining vertices according to the function *dist* and the current S and $\bar{S}$. Also, we update the set $V := V \setminus \{u\}$. The case for including a vertex in the set $\bar{S}$ is similar, where a vertex u is popped from the top of the list L.

After initializing, we keep on doing the following unless there are no more changes in the best solution. We randomly choose a parent and a child solution from the ternary tree and crossover the parent's pocket solution with one of its child's current solution. All vertices that are contained in the same set for both the parents are included in the same set in the offspring. Then both sets are filled according to a differential greedy recombination method similar to that in the initialization. In this case, the starting set S ($\bar{S}$) for the offspring is given by the intersection of the set S ($\bar{S}$ resp.) from both the parents. This crossover method takes care of the diversity in the population.

To better the quality of the obtained solution, we also employ the *variable-neighborhood search (VNS)*, first proposed by Hansen and Mladenovic [26] for the local search in the neighborhood of the new offspring. Contrary to other local search methods, VNS allows enlargement of the neighborhood structure along the search.

Finally, whenever the population stagnates, we keep the best solution and re-initialize the rest of solutions in the set and run the above process again for certain number of generations (say, 30).

For the small sized problems (graphs containing less than 25 vertices), we used backtracking. Notice that even though backtracking gives us an optimal solution, a memetic algorithm may not.

2. 2. Feature Selection

The next step is to find the genetic signature for each non-leaf node of the hierarchical tree obtained from the above clustering method. This method is aimed to find the optimal set of genes which enforces inter-class discrimination and intra-class similarity.

Given a set of voxels divided into two disjoint subsets $S1$ and $S2$, we define the function (often called *target*) t that maps voxel $v \in S1 \cup S2$ into the set $\{1; 2\}$ such that $t(v) = i$ if $v \in Si$ where $i \in \{1; 2\}$. Briefly, $t(v)$ gives the set to which v belongs. Now, the feature selection method is composed of two phases:

Phase 1 - Minimum feature selection. In this phase, a linear integer model is built and solved to discover the minimum number of genes, k, that are necessary such that all pairs of voxels $(i; j)$ from different sets (i.e. with $t(i) \neq t(j)$) have at least α dichotomies. Also, all pairs of samples $(i; z)$ from the same class (i.e., with $t(i) = t(z)$), have at least $ß$ similarities.

Phase 2 - Maximum cover gene selection. As there might be multiple solutions with k genes from phase 1, we construct and solve a new linear integer model to find the set of k genes that maximizes the contrast between different classes, and also maximizes the similarities within classes. To do this, phase 2 keeps the $(\alpha, ß)$ constraints plus one extra constraint of k genes to ensure that we are looking only to the solutions obtained from phase 1.

It is clear that to use this method, we need to define when a given gene g keeps similarities within one class or not. In this case, entropy is applied as in [3] over each gene in the data set.

Basically, the Fayyad and Irani (1993) [3] method finds thresholds for a gene's expression, which best divide the distinct classes, according to a minimum description length criterion. As we are interested in dichotomies, a modified version of their algorithm was used to find only one threshold r [3].

Therefore, a gene g makes a dichotomy between voxels i and j if and only if, the threshold r for gene g separates the expression value g_i and g_j, i.e.

$g_i < r < g_j$ or $g_j < r < g_i$

$$(g_i \leq r \wedge g_j > r) \vee (g_i > r \wedge g_j \leq r) \quad (2)$$

The above phases comprise optimization problems that are not likely to be fixed-parameter tractable as shown in [4]. However, the use of a standard integer-programming formulation as in [37] in conjunction with an IP solver (ILOG CPLEX 9.1) enabled us to solve this problem to optimality.

3. Experiments

We give a hierarchical classification tree for the normal brain based on the genes expressions using the above mentioned divisive hierarchical clustering method. The grouping obtained is confirmed by applying Principal Component Analysis (PCA) over the entire data set. We produce a set of genetic signatures based on the hierarchy obtained, which identifies each group, and also characterize their hierarchical relationship. The provided genetic signatures reveal the profile of similarities between some non-connected regions of the brain. The identified signatures for the hierarchical classification are further examined over a mouse brain with Parkinson's disease. We show that the genetic specializations in the normal brain, pointed by the signatures, change considerably in the specimen with Parkinson's disease. Furthermore, we compare the regions of the normal and diseased brains according to the hierarchy, and give genetic signatures that highlight significant changes due to the Parkinson's disease. In this work, we show how the known regions of the affected PD brain, i.e., cerebellum and striatum and the relatively unexplored regions are affected. The hypothesis is that the whole brain is slowly affected in PD. It is also highly likely that the effect observed in this model of PD is a result of chemically induced disease.

4. Normal Brain

4.1 Classification Tree of the Voxels

The hierarchical clustering using the Arithmetic-Harmonic Cut (AHC) (described in Section 2.1) on the voxels from a normal mouse brain yields the classification tree shown in Fig. 2.

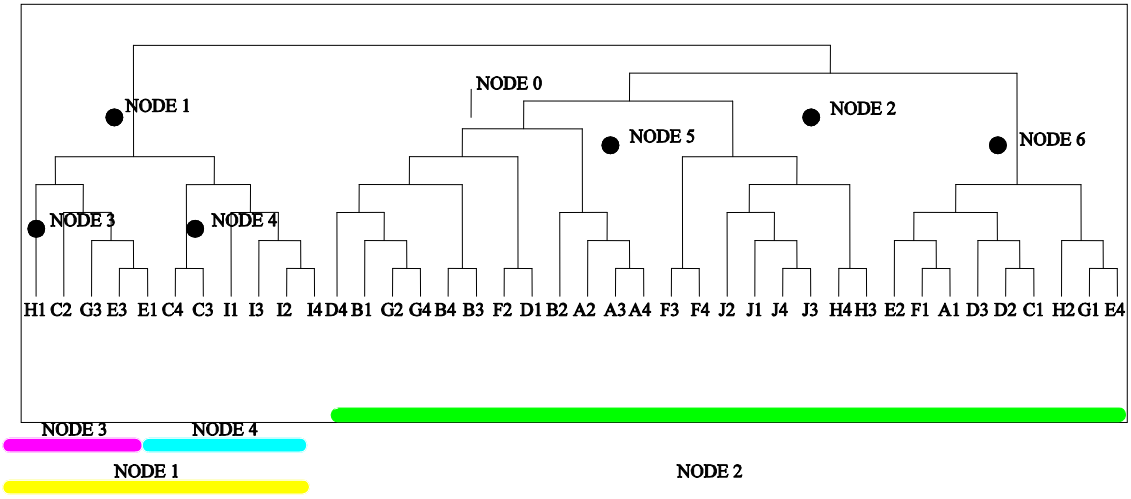

Fig.2 Hierarchical Classification for the normal mouse brain. After applying AHC on Node1, all voxels from section *I* are clustered together in Node4.

The first partition at Node0 of the tree (i.e. Node1 consists of voxels H1, C2, G3, E3, E1, C4, C3, I1, I3, I2 and I4). Amongst these, (I1, I2, I3 and I4) belong to the *rostral cerebellum*. The cerebellum is a very important structure in motor movement and motor-vestibular memory and learning. It is responsible for processing sensory information and providing coordinated, smooth movements of the skeletal muscular system. After the application of the AHC on Node 1, again we see that all voxels from the section *I* are clustered together in Node4 (see Figure 2). The voxels C2, C3, C4 in Node1 are from *septo-striatum*. Striatum is best known for its role in the planning and modulation of movement pathways. Both cerebellum and striatum are mentioned in the literature for affecting the brain of a MA-treated mouse. We also note that the slices E and G correspond to *septo-diencephalon* and *caudal diencephalon* respectively. The *Thalamus* in the diencephalon part is responsible for coordination and regulation of all functional activity of the cortex. Finally, the slice H corresponds to the *caudal mesoencephalon* region of the brain.

Note that after unsupervised clustering of the voxels, even spatially disconnected voxels in the mouse brain are clustered together, depending on the similarity of their expressions.

4.2. Genetic Signatures

In Fig. 3a, we consider the genes which best distinguish between the voxels in Node1 and Node2, while keeping intra-cluster similarity. We employ (α,ß)-feature selection (See Section 2.2) for this purpose. Using $\alpha = 1{,}001$, $ß = 881$, we found 1,657 genes that give us the best signature for Node0, splitting Node1 from Node2.

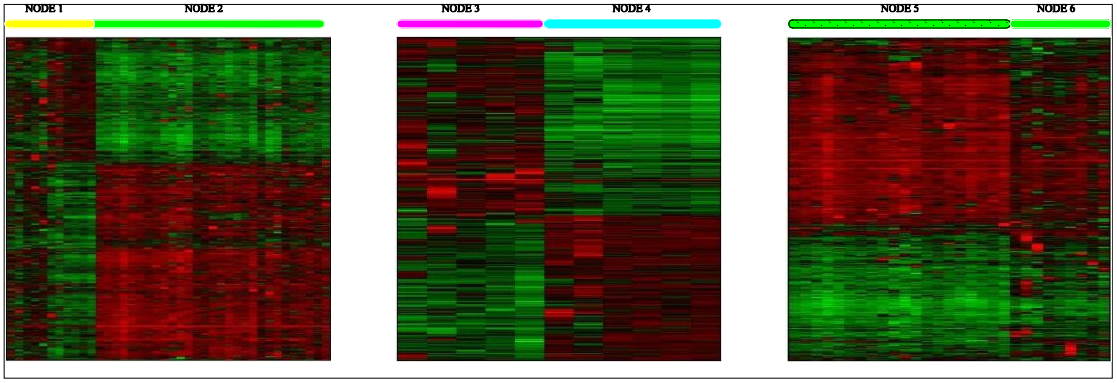

Fig. 3 .Genetic Signatures for the normal mouse brain: (A) Signature at Node0 distinguishing the voxels in Node1 and Node2. (B) Signature at Node1 distinguishing the voxels in Node3 and Node4. (C) Signature at Node2 distinguishing the voxels in Node5 and Node6

This signature shows that the voxels in Node1 have clearly distinguished expression for all the genes selected. This is a clear marker of a normal brain and is also the evidence why the disconnected voxels like C3 and C4 voxels are together with the cerebellum ones (I1, I2, I3 and I4).

Similarly, we consider the division at Node1 and Node2 in Fig. 3b and Fig. 3c respectively. We found *1,783* genes with $\alpha = 1{,}415$; $ß = 1{,}403$ at Node1 and *398* genes with $\alpha = 221$, $ß = 221$ at Node2 respectively. The signature at Node1 is again very clear in distinguishing voxels from Node3 and Node4. Notice that voxels from the cerebellum are all in Node4. However, the signature at Node2 is rather *homogeneous* (very uniform, the voxels are not changing much). This means that the voxels in Node2 are not too dissimilar in terms of gene expression. Therefore, we did not divide Node2 further. If this group is splitted, no distinguishing signature can be observed to support further sub-classification. Same reasoning applies to Node3 and Node4.

We use Wilcoxon-Mann-Whitney test to find the *50* best genes with lowest *p*-values from each of the above signatures of Node0 and Node1 and in Section 4.3, we describe our findings for the obtained genes.

Principal Component Analysis

Principal Component Analysis (PCA) reduces dimensionality while keeping the maximum possible variance of the original data. We performed an PCA analysis on the *7,035* genes for the normal brain and show the results in Fig. 4. The green, blue and the red points turn out to belong to Node3, Node4 and Node2 of Fig. 2 respectively. Notice that the voxels of Node3 and Node2 are not fully separated. In Figs. 4 and 5, we show again that the PCA analysis on *1,657* genes obtained by the feature selection at Node0 (shown in Fig. 3(A)). In this case, the clusters are much better separated.

4.3. Genes

First, we consider the *50* genes with best *p*-values from the signature at Node0 (shown in Fig. 3a). For all these genes, *p*-values are lower than 4×10^{-6}. We discuss the known functions of some of the obtained genes in the following. The gene MRP111 is involved in protein biosynthesis. Now, it is known that the abnormal accumulation of substrates due to loss of Parkin function may be the cause of neurodegeneration in parkin-related Parkinsonism.

In [14], p38 (a key structural component of the macromolecular aminoacyl-tRNA synthetase (ARS) complex involved in protein biosynthesis) was identified to be playing a role in the pathogenesis of PD. The p38 subunit of the aminoacyl-tRNA synthetase complex is a Parkin substrate, thus protein

biosynthesis and neurodegeneration are linked. Also, the genes STX7 and CSNK2A1 are found already in [6] to be in the Parkinson's disease pathway. Genes USP19, SENP5 and K1K12 are involved in Proteolysis.

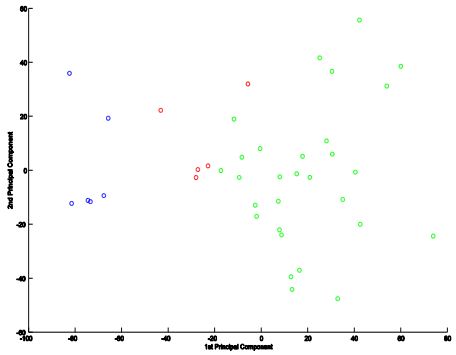

Fig. 4 PCA analysis on (a) total 7,035 genes.

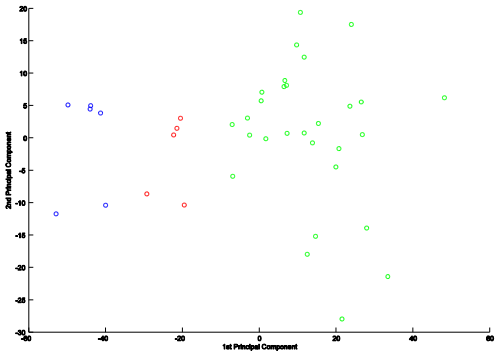

Fig. 5 1,657 genes obtained by the feature selection at Node0 (shown in Fig. 3(A))

In [7], failure of ubiquitin-mediated proteolysis is suggested to be central in the pathogenesis of neurodegenerative disorders like Alzheimer's and Parkinson's diseases. The gene IKBKB has a biological function of protein phosphorylation using protein phosphate, which represents the final pathway in the action of transmitters and hormones at neuronal level [17]. Also, there is a clear link between protein phosphorylation and Parkinson's disease, since it regulates NMDA receptors, contributing to the pathogenesis of motor dysfunction in PD subjects [9]. The gene PAPSS2 is involved in amino acid biosynthesis. The study in [10] shows that the biosynthesis of amino acids reduces the metabolic activity of neurons in Parkinsonism. The genes HOXC5 and TIEG1 are involved in mRNA transcription regulation. A recent report [11] indicated that D1, a dopamine receptor (used in mRNA transcription regulation) changes after chronic levodopa (L-dopa) treatment, linking it with Parkinson's disease. Also, the gene TNC is involved in Neurogenesis [12] and is present in our signature. The gene NDUFA7 which belongs to the Alzheimer disease pathway [13] also plays a significant role in this partitioning of the normal brain. Then, the gene MllT2H is an oncogene involved in cell cycle regulation. There is a strong connection between cell cycle and neurodegenerative diseases, especially Alzheimer's [12]. However, MllT2H also seems to play a good role in separating Node1 from Node2.

Similarly, we consider some of the *50* genes with best *p*-values (at most 7×10^{-3}), which distinguish between Node3 and Node4. It turns out that the genes

CMYA4 and SEPT4 are already known to be in the Parkinson disease pathway. The genes MRPS14, GSPT1 and RPS5 have functions in protein biosynthesis and are mentioned in [14] in relation to PD. We also found DDX50 in this signature, which is involved in nucleoside, nucleotide and nucleic acid metabolism. Interestingly, we found the gene TNFAIP8 in this signature. The gene TNFAIP8 belongs to the Huntington's disease pathway, a CNS condition characterized by the damage of the nerve cells in the *basal ganglia* and cerebral cortex. This gene is also involved in the influence of altered expression of complexins in the modulation of neurotransmitter releases [15] and morphological variations of striatal medium spiny neurons [16]. The genes BIRC2, MTHFD1, STK11 were also selected and they are active in the apoptosis signaling pathway, amino acid biosynthesis [10], and protein phosphorylation [17][9] respectively and the three of them were previously related to PD. The gene RAP1A is part of the G-protein signaling pathway, which plays a variety of roles in numerous neuronal functions. Also, the interaction between G protein-coupled receptor kinases (GRKs) and *ß*-arrestins regulates physiological responsiveness to psychostimulants, indicating a potential involvement in brain disorders, such as addiction, Parkinson's disease and schizophrenia [18]. The genes JMJD3 and MEF2B also take part in mRNA transcription regulation and are related to PD in [11]. Finally, the gene CATNA1 appears in the signature which was found in Alzheimer disease pathway [13]. The signature at Node2 is not very distinctive in the sense that there are no major two groups present in the voxels at Node2. Therefore, the signature at Node2 of the normal brain is omitted here.

5. Comparison between Normal and Parkinson Brain

Next we compare the level of the expressions between the normal brain and MA-treated mouse brain for the genes in the signature at Node0, Node1 and Node2 of the normal brain. In Fig. 6a, we consider the genetic signature for Node0 of the normal brain and show their expressions at the corresponding voxels in the MA-treated mouse brain. The genes presented in the picture are in the same order as in Fig. 3. It is evident from this picture that the voxels in Node1 of the the MA-treated mouse brain turns out to have expressions very similar to those for voxels in Node2. This means that the genes which were differentially expressed in Node1 and Node2 of the normal brain, are now similarly expressed in the voxels at both nodes, yielding a rather homogenous signature. The same phenomenon is also present at the voxels in Node1 of the MA-treated mouse brain. The genes which are differentially in Node3 and Node4 have now similar expressions in the MA-treated brain. Also, notice that the voxels in Node2 of the MA-treated brain are also affected, even though to a smaller extent. The genes which were differentially expressed (even though weakly) in Node5 and Node6 (Node5 and

Node6 originate from Node2) have now a more homogenous expression in both nodes of the MA-treated brain. This may generate a hypothesize for Parkinson's disease, even though initial effects can only be observed in the cerebellum, it is probably that genetic affects over other parts of the brain can be measured before other phenotype changes appear.

6. Genetic Signature for Node-wise Comparison of Normal and MA-Treated Brains

In the previous section, we showed that the expressions at different voxels of MA-treated brain tend to be very similar in contrast to those for a normal brain. Now, we seek a signature to find a small number of genes which are differentially expressed at each of the nodes Node1, Node2, Node3 and Node4 in the normal and MA-treated brain. This will allow us to look at each of the voxels in the corresponding nodes and pin-point genes which are affecting each of these nodes in PD. In Figs. 6a-d we show genetic signatures where the voxels of the normal brain at each of the nodes Node1, Node2, Node3 and Node4 are compared against those in the MA-treated brain respectively. In Figs. 6e and 6f, we also show the union of the genes present in Figs. 6a to 6d in the whole normal brain and in the whole MA-treated brain.

For further functional analysis of the above nodewise comparison, we select a set of *50* genes with best *p*-values in the following for each of the above cases in Figs. 6a-d. For Node1, we show in Fig. 6 the expressions of *1,888* genes from feature selection with $\alpha = 1{,}223;\ ß = 1{,}223$. We select *50* of the above genes with maximum *p*-value as 8×10^{-11}. Similarly for Node2, we show *1,375* genes in Figs. 6(b) with $\alpha = ß = 664$ and the maximum *p*-value of the selected *50* genes from this is 10^{-8}. Also, for Node3 and Node4, we get *1,763* and *2,691* genes respectively. For Node3, $\alpha = 1{,}435;\ ß = 1{,}403$ and the maximum *p*-value taken is 10^{-2} and for Node4, we have $\alpha = 2{,}691;\ ß = 2{,}260$ and the maximum *p*-value of the selected genes is 4×10^{-3}.

In the following, we consider the voxels in the nodes Node1, Node2, Node3 and Node4 for the normal and the MA-treated brains. In each case, we indicate some of the important genes from the above sets of *50* genes which distinguish a node in the normal brain from that in the MA-treated brain.

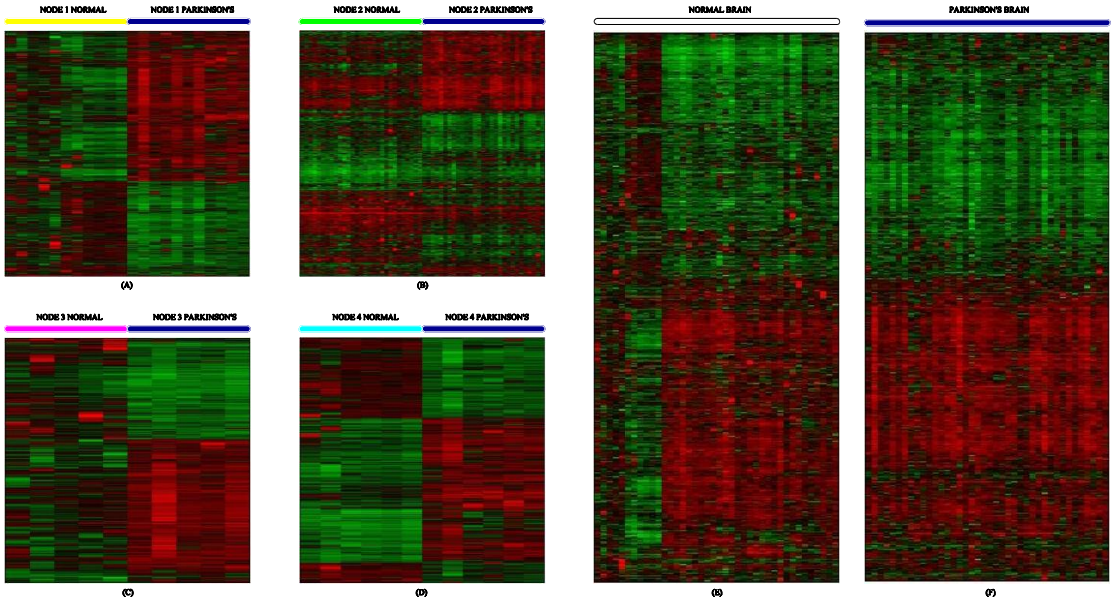

Fig. 6. Comparing different partitions of the normal and the MA-treated mouse brain. Genetic signature distinguishing (a) Node1, (b) Node2, (c) Node3, (d) Node4 of both brains, union of all the aforementioned genes in (e) normal brain and (f) MA-treated brain

6.1. Genes: Node1

We list down the genes which were significantly differentially expressed in the two sets of voxels (normal and PD) from Node1 and were mentioned in literature for their functions related to brain or in relation to some neurodegenerative diseases. We found the genes PINK1 [19] and CSNK2A1 [6], both of which are already mentioned in the literature to be PD-related. Notice that CSNK2A1 was already found to be different in Node1 and Node2. It seems that this gene expression in PD brain is more like that in Node2.We also found PPP1R9B which was also related to Schizophrenia [20]. This is very significant as it is widely accepted that neurodegenerative diseases in general share similar action mechanisms and also have genes in common. We also found GSK3B which is related to both Alzheimer's disease and Schizophrenia [13]. We obtained another gene, MTA2, which promotes the deacetylation of p53, modulating cell growth arrest and apoptosis. Recent studies have shown a relation between p53 and brain degeneration [21]. The gene BC003885 is involved in Protein biosynthesis and is also reported in [14] for playing a role in PD. The genes PTPN9, CDK9 and PRPF4B are involved in Protein phosphorylation, as

mentioned in Section 4.3. The genes RGS5 and PRKAR1B are part of G-protein signaling pathway [18] and are found in this case. Again, USP19 which was distinguishing Node1 and Node2 of the normal brain appears in the comparison between the Node1 of the two brains. There were also two genes, PHGDH and CTBP1, which are known to be active in Amino acid biosynthesis and affecting PD [10]. Finally, the genes DDX21, SlC35C1 and RBM28 are involved in nucleoside, nucleotide and nucleic acid metabolism and considering their functions, they may be related to Parkinson's disease.

6.2. Genes: Node2

We already mentioned that even though in a smaller amount, the voxels in Node2 are also affected in the PD brain. Here we mention some of the few genes which we found in the 50-gene set. We found the genes MAPK6 and GSK3B, which also appear in Alzheimer's disease pathway and also have been linked to Schizophrenia [13]. Notice that the gene GSK3B also played an important role in distinguishing the normal brain's Node1 and Node2. The gene SNX27 is again related to G-protein signaling pathway and known to affect PD [18]. We also found a new gene SlC1A2 in Node2 comparison, which is in the ionotropic glutamate receptor pathway (iGluR). IGluRs are responsible for fast, reliable neurotransmission in the vertebrate central nervous system, being essential for learning and memory. There is also another gene MRPS25 which plays a role in protein biosynthesis and is related to PD [14]. The genes NCOR1, E2F7, GFI1B, and ZFP113 take part in mRNA transcription regulation and are linked to PD in [11]. We also found the gene CD1D1 has an intriguing function called T-cell mediated immunity. It has been known for many years that there are individuals with Parkinson's disease who have alterations in their immune system. Actually, PD patients exhibit a lower frequency of infections and cancer, suggesting a stimulation of the immune system [22]. The genes STK11 and STK35 participate in protein phosphorylation and are related to PD as mentioned before [17][9]. The gene MTHFD1 is concerned with amino acid biosynthesis and appears to affect PD [10] as discussed earlier. Here, we notice that the genes STK11 and MTHFD1 also appears in the signature of the normal brain for distinguishing Node3 and Node4.

The genes ZBP1 and CDH13 are oncogenes involved in cell cycle regulation. There is a strong connection between cell cycle and neurodegenerative diseases, especially Alzheimer's [12]. However, this gene is also having differences in the Node2 of the normal and PD brain. The gene LPHN2 is a gene appearing in G-protein mediated signaling and is again found to be related to PD as mentioned earlier [18]. The gene HSD3B4 participates in cholesterol metabolism, which is a very generic biological function. However, differences in

movement of cholesterol between different cellular compartments of the CNS and across the blood-brain barrier to the plasma were detected in mice with one form of neurodegenerative disease (Niemann-Pick type C). The gene PCBD takes part in Pterin metabolism. Pterin cofactors are required for tyrosine hydroxylase (TH) activity, which converts tyrosine into L-dopa. The final conversion of L-dopa to dopamine is controlled by an enzyme called dopadecarboxylase. The relation between pterin metabolism and Parkinson's is well-known, since fibroblasts supplemented with pterin cofactors were found to produce L-dopa [23].

6.3. Genes: Node3

Next we discuss the significant genes which differ in Node3 of the two brains. We found the gene BTK which is involved in Protein phosphorylation and is mentioned in relation to PD in [17][9]. Then the gene ADAM12 and MEll1are required for proteolysis and it is known that failure of this process can trigger neurodegenerative diseases like PD [7]. The gene CDC2A is found the p53 pathway and recent studies have shown a relation between p53 and brain degeneration [21]. Also, the gene JARID2 is involved in mRNA transcription regulation and neurogenesis and is related to PD [11]. The gene CANT1 has role in nucleoside, nucleotide and nucleic acid metabolism and is affected in Node2. The gene PRKAR1A appears in G-protein signaling pathway and is mentioned in connection with PD by [18]. Again, the gene TNFAIP8, which appears in Huntington disease pathway and has already appeared in the signature of Node1 distinguishing Node3 and Node4 of the normal brain, gets affected in the Node3 of the PD brain.

6.4. Genes: Node4

Now, we describe the genes which are affected in Node4 (containing cerebellum). In the signature where Node3 and Node4 of the normal brain were compared, we found two genes, CMYA4 and CATNA1, which again appear in the comparison of two sets of Node4 voxels. The genes UBE2J1 and TRIM11 are involved in proteolysis and are known to affect neurodegenerative diseases [7]. We also found the following genes which participate in mRNA transcription regulation and are known to affect PD brain [11]: JMJD3, CHD1l, PHTF2 and MEF2B. The genes related to protein phosphorylation (IlKAP, PPM1B, CDK6 and ITK) are also affected in Node4 of the PD brain [17][9]. The ITK gene has also the T-cell mediated immunity biological function [22]. Again, CDK6 is an oncogene involved in cell cycle regulation and is connected to Alzheimer's disease [12]. In Node4 of the PD brain, we also found changes in the genes ADK, which is responsible for nucleoside, nucleotide and nucleic acid metabolism;

FGF15 which causes neurogenesis; and GSPT1 which takes part in protein biosynthesis (known to be related to PD [14]).

7. Conclusions

In this paper, we applied a new unsupervised clustering on the microarray data for the voxels in a normal mouse brain and obtained the corresponding genetic signature distinguishing the two sets of voxels of the normal mouse brain during first few bi-partitioning of this dataset. We also compared the similar clusters of normal and MA-treated mouse brains using the genes. Finally, we showed some of the genes which were significantly affected in each of the clusters of the MA-treated brain. In this work, we did not only show how the known regions of the affected PD brain, i.e., cerebellum and striatum are affected, but also showed how the relatively unexplored regions were affected. Our hypothesis is that probably the whole brain is slowly affected by PD.

REFERENCES

[1]. *V. Brown, A. Ossadtchi, A. Khan, S. Cherry, R. Leahy and S. Smith*, "High-throughput imaging of brain gene expression" in Genome Research, **vol**. 12, 2002, pp. 244-254.

[2]. *O. Troyanskaya, M. Cantor, G. Sherlock, P. Brown, T. Hastie, R. Tibshirani, D. Botstein and R. Altman*, "Missing value estimation methods for DNA microarrays" in Bioinformatics, **vol**. 17, 2001, pp. 520-525.

[3]. *U. Fayyad and K. Irani*, "Multi-interval discretization of continuous-valued attributes for classification learning" in Proc. of the 13th International Joint Conference on Artificial Intelligence, 1993, pp. 1022-1029.

[4]. *C. Cotta and P. Moscato*, "The K-feature set problem is W[2]-complete" in Journal of Computer and System Sciences, **vol**. 67, 2003, pp. 686-690.

[5]. *O. Corti, C. Hampe, H. Koutnikova, F. Darios, S. Jacquier, A. Prigent, C. Robinson, L. Pradier, M. Ruberg, M. Mirande, E. Hirsch, T. Rooney, A. Fournier and A. Brice*, "The p38 subunit of the aminoacyl-tRNA synthetase complex is a parkin substrate: linking protein biosynthesis and neurodegeneration" in Human Molecular Genetics, **vol**. 12, 2003, pp. 1427-1437.

[6]. *G. Lee, M. Tanaka, K. Park, S. Lee, Y. Kim, E. Junn and M. Mouradian*, "Casein kinase II-mediated phosphorylation regulates alpha-synuclein/synphilin-1 interaction and inclusion body formation" in The Journal of Biological Chemistry, **vol**. 279, 2004, pp. 6834-6839.

[7]. *R. Layfield, J. Cavey and J. Lowe*, "Role of ubiquitin-mediated proteolysis in the pathogenesis of neurodegenerative disorders" in Ageing Research Reviews, **vol**. 2, 2003, pp. 343-356.

[8]. *M. Magnoni, S. Govoni, F. Battaini and M. Trabucchi*, "The aging brain: protein phosphorylation as a target of changes in neuronal function" in Life Sciences, **vol**. 48, 1991, pp. 373-385.

[9]. *J. D. Oh and T. N. Chase*, "Striatal mechanisms and pathogenesis of parkinsonian signs and motor complications" in Annals of Neurology, **vol**. 47, 2000, pp. S122-S129.

[10]. *E. Grundig, W. Mayer and F. Gerstenbrand*, "Biosynthesis of amino acids from glucose in the central nervous system in the parkinson syndrome" in Arch Psychiatry Nervenkr, **vol**. 233, 1983, pp. 397-408

[11]. *I. Aubert, C. Guigoni, K. Hakansson, Q. Li, S.Dovero, N. Barthe, B. Bioulac, C. Gross, G. Fisone, B. Bloch and E. Bezard*, "Increased D1 dopamine receptor signaling in levodopa-induced dyskinesia" in Annals of Neurology, **vol**. 57, 2005, pp. 17-26.

[12]. *A. Raina, X. Zhu, C. Rottkamp, M. Monteiro, A. Takeda and M. Smith*, "Cyclin toward dementia: Cell cycle abnormalities and abortive oncogenesis in Alzheimer disease" in Journal of Neuroscience Research, **vol**. 61, 2000, pp. 128-133.

[13]. *E. S. Emamian, D. Hall, M. J. Birnbaum, M. Karayiorgou and J. A. Gogos*, "Convergent evidence for impaired AKT1-GSK3beta signaling in schizophrenia" in Nature Genetics, **vol**. 36, 2004, pp. 131-137.

[14]. *O. Corti, C. Hampe, H. Koutnikova, F. Darios, S. Jacquier, A. Prigent, J. C. Robinson, L. Pradier, M. Ruberg, M. Mirande, E. Hirsch, T. Rooney, A. Fournier and A. Brice*, "The p38 subunit of the aminoacyl-tRNA synthetase complex is a parkin substrate: linking protein biosynthesis and neurodegeneration", in Human Molecular Genetics, **vol**. 12, 2004, pp.1427-1437.

[15]. *W. Freeman and A. J. Morton*, "Differential messenger RNA expression of complexions in mouse brain", in Brain Research Bulletin, **vol**. 63, 2004, pp. 33-44.

[16]. *D. J. Surmeier and N. Spruston*, "Peering into the dendritic machinery of striatal medium spiny neurons" in Neuron, **vol**.44, 2004, pp. 401-402.

[17]. *M. Magnoni, S. Govoni, F. Battaini, and M.T rabucchi*, "The aging brain: protein phosphorylation as a target of changes in neuronal function" in Life Science, **vol**. 48, 1991, pp. 373-385.

[18]. *R. R. Gainetdinov, R. T. Premont, L. M. Bohn, R. J. Lefkowitz and M. G. Caron*, "*Desensitization of G protein-coupled receptors and neuronal functions*", in Annual Review of Neuroscience, **vol**. 27, 2004, pp. 107-144.

[19]. *Y. Hatano, Y. Li, K. Sato, S. Asakawa, Y. Yamamura, H. Tomiyama, H. Yoshino, M. Asahina, S. Kobayashi, S. Hassin-Baer, C. S. Lu, A. R. Ng, R. L. Rosales, N. Shimizu, T. Toda, Y. Mizuno and N. Hattori*, "Novel PINK1 mutations in early-onset parkinsonism" in Annals of Neurology, **vol**. 56, 2004, pp. 424-427.

[20]. *A. Law, C. Weickert, T. Hyde, J. Kleinman and P. Harrison*, "Reduced spinophilin but not microtubule-associated protein 2 expression in the hippocampal formation in schizophrenia and mood disorders: Molecular evidence for a pathology of dendritic spines" in American Psychiatric Association, **vol**.161, 2004, pp. 1848-1855.

[21]. *W. B. Jacobs, G. S. Walsh and F. D. Miller*, "Neuronal survival and p73/p63/p53: a family affair" in Neuroscientist, vol. 10, 2004, pp.443-455.

[22]. *A. Czlonkowska, I. Kurkowska-Jastrzebska, A. Czlonkowski and S. D.Peter*, "Immune processes in the pathogenesis of Parkinsons disease a potential role for microglia and nitric oxide" in Medical Science Monitor, **vol**. 8, 2002, pp.RA165-RA177.

[23]. *J. A. Wolf, L. J. Fisher, L. Xu, H. A. Jinnah, P. J. Langlais, P. M. Iuvone, K. L. O'Malley, M. B. Rosenberg, S. Shimohama and T. Friedmann*, "Grafting fibroblasts genetically modified to produce L-dopa in a rat model of Parkinson disease" in Proceedings of the National Academy of Sciences, **vol.** 86, 1989, pp. 9011-9014.

[24]. *D. Whitley*, "A genetic algorithm tutorial", in Statistics and Computing, **vol**. 4, 1994, pp. 65-85.

[25]. *L. Buriol, P. M .Franca and P. Moscato*, "A New Memetic Algorithm for the Asymmetric Traveling Salesman Problem" in Journal of Heuristics, **vol**. 10, 2004, pp. 483-506.

[26]. *P. Hansen, N. Mladenovic*, "Variable Neighbourhood Search: Principles and Applications", in European Journal of Operational Research, vol. 24, 2001, pp. 449-467.

[27]. *P. Mahata*, "Exploratory Consensus of Hierarchical Clusterings for Melanoma and Breast Cancer**",** in IEEE/ACM Transactions on Computational Biology and Bioinformatics, **vol**. 99, 2008.

[28]. *P. Merz and B. Freisleben*, "Fitness Landscapes, Memetic Algorithms, and Greedy Operators for Graph Bipartitioning", in Evolutionary Computation, **vol**. 8, 2000, pp. 61-91.

[29]. *B. R. Battiti, and A. Bertossi*, "Differential Greedy for the 0-1 Equicut Problem" in: Du D, Pardalos P (eds) Proc. of DIMACS Workshop on Network Design: Connectivity and Facilities Location. American Mathematical Society, 1997, pp. 3—21.

[30]. *P. Festa, P. Pardalos, M. Resende and C. Ribeiro*, "Randomized heuristics for the MAX-CUT problem" in Optimization Methods and Software, **vol**. 7, 2002, pp. 1033-1058

[31]. *R. Berretta, C. Cotta and P. Moscato*, "Enhancing the performance of memetic algorithms by using a matching-based recombination algorithm", in: Resende M, de Sousa NJ, Viana A (eds) Metaheuristics: computer decision-making. Kluwer Academic Publishers Norwell MA USA, 2004.

[32]. *A. Mendes, C. Cotta, V. Garcia, P. Franca and P. Moscato*, "Parallel memetic algorithms for gene ordering in microarray data" in Proc. of the 2005 International Conference on Parallel Processing Workshops (ICPPW'05), **vol**. 14, 2005, pp. 604 – 611.

[33]. *P. Mahata, W. Costa, Cotta C and P. Moscato*, "Hierarchical Clustering, Languages, and Cancer", in: Rothlauf F and *et al.* (eds) Lecture Notes in Computer Scienc (Applications of

Evolutionary Computing: EvoWorkshops 2006: EvoBIO.). Budapest Hungary, **vol**. 3907, 2006, pp. 67-78.

[34]. *J. Brandt and W. Hein*, Artificial Intelligence in Theory and Practice, in: Bramer M (eds) IFIP AI. IFIP. Springer, Berlin Heidelberg New York, 2001.

[35]. *I. Dyen, J. B. Kruskal and P. Black*, "An Indoeuropean classification: A lexicostatistical experiment" in Transactions of the American Philosophical Society, **vol**. 82, 1992, pp. 1-132.

[36]. *D. Ross, U. Scherf, M. Eisen, C. Perou, C. Rees, P. Spellman, V. Iyer, S. Jeffrey, M. Rijn, M. Waltham, A. Pergamenschikov, J. Lee, D. Lashkari, D. Shalon, T. Myers, J. Weinstein, D. Botstein and P. Brown*, "Systematic variation in gene expression patterns in human cancer cell lines", in Nature Genetics, **vol**. 24, 2000, pp. 227-235.

[37]. *R. Berretta, A. Mendes and P. Moscato*, "Integer Programming Models and Algorithms for Molecular Classification of Cancer from Microarray Data" in: Proceedings of the Twenty-eighth Australasian conference on Computer Science, **vol**. 38, 2005, pp. 361 - 370.