Combinatorial and Algorithmic Issues for Microarray Analysis

Carlos Cotta

Departamento de Lenguajes y Ciencias de la Computación

Universidad de Málaga, ETSI Informática (3.2.49)

Campus de Teatinos, 29071-Málaga, SPAIN

Email: ccottap@lcc.uma.es

Michael Langston

Department of Computer Science

University of Tennessee

203 Claxton Complex, 1122 Volunteer Boulevard

Knoxville, TN 37996-3450, USA

E-mail: langston@cs.utk.edu

Pablo Moscato

Newcastle Bioinformatics Initiative

School of Electrical Engineering and Computer Science

The University of Newcastle

Callaghan, NSW, 2308, AUSTRALIA

E-mail: Pablo.Moscato@newcastle.edu.au

August 16, 2005

1 Introduction

J. Craig Venter declared in 1998: *"We are now starting the Century of Biology"*. This is undoubtedly true and has been recognized before. Gregory Benford was already pointing this fact in 1995 when he also noted that Physics has dominated the 20th century, as Chemistry had probably dominated the 19th century. In his own words:

> *"And yet, far from the physics departments of the great campuses, a clarion call is sounding through our time, one that responds to hot-button environmental problems and that incorporates rapid advances in other laboratories: Biology has turned aggressively useful."*

Among the novel and revolutionary biotechnologies, microarrays have been evolving fast in the past ten years and are reshaping our understanding of biological systems as well as shaking the grounds of biomedical research. Microarrays allow to monitor the expression of thousands of genes at once. A single experiment allows to text for billions of individual hypotheses. A query using the PubMed website [1] shows that more than 11,010 entries have already either the word "microarray" or "DNA array". Almost all of these publications appeared in less than 10 years, approximately 70 percent of them appeared in print in the last two years. These high-throughput molecular assays generate immense datasets. These datasets have the potential to help us to understand biological systems in ways that are completely new. While huge promises are ritually proclaimed (personalized medicine, targeted therapies, genetic engineering for more efficient crops, etc.) [31], the challenges are equally enormous [42].

Not only biotechnologies are evolving fast, interestingly enough, combinatorial optimization has also turned to be an "aggressively useful" discipline. Many advances in exact algorithms, though worst-case exponential, are allowing increasingly larger instances to be solved to optimality. The development of *fixed-parameter algorithms* as a recognized subdiscipline of computational complexity, is aiming at a systematic development of data reduction methods that bound the search of optimal solutions. The development of metaheuristics starting with Simulated Annealing in 1983, followed by Tabu Search [33, 54] and Memetic Algorithms a few years later [45, 50], has allowed researchers to adapt *ad hoc* heuristics, and to develop very powerful stochastic algorithms for large-scale optimization problems.

[1] `http://www.ncbi.nlm.nih.gov/entrez/query.fcgi`

In spite of all the progress, sometimes combinatorial optimization methods have been naively applied in Bioinformatics, missing “the users real needs”. For a number of problems, the ultimate goal is not to find an optimal solution (minimum cardinality is naively sought after many times in the literature), but to uncover the underlying interaction genetic networks and the conserved aspects of biosystems (conserved by evolution, in different disease states or under different perturbations). Recognition of this strategic issue has led the authors of this manuscript to work on a number of NP-hard problems that provide a good mathematical formulations of the basic questions of biological interest.

In general, these research questions have a natural representation as combinatorial optimization problems on graphs. However, it is sometimes the case that a variant or a generalization, of a known combinatorial optimization problem, is better suited. In this chapter, we will present three illustrative examples of this research methdology drawn from the authors’ own experience in the analysis of microarray datasets. We aim at illustrating the benefits of combinatorial approaches by unifying the presentation with the underlying theme of three types of subgraph identification problems: *cliques, bicliques* and *hamiltonian paths*. In addition, another unifying theme is that the associated decision problems are all NP-complete. Another unifying theme is the power of data reduction in two of them, specially in the case of clique, where a *fixed-parameter algorithm* is of extreme importance and help us to calculate a particular *clique interaction graph*, of interest to uncover the active genetic pathways.

1.0.1 Fixed-Parameter Tractability.

The origins *fixed-parameter tractability* can be traced at least as far back as in 1988, to the works of Fellows and Langston. These authors have shown that, thanks to the Graph Minor Theorem, a variety of *parameterized versions* of NP-complete problems are tractable when a relevant input parameter is assumed to be fixed, i.e. independent of the instance size [29, 30]. A computational complexity class (FPT) was introduced (FPT) encompasses those parameterized problems for which there exists an algorithm that runs in $O(f(k)n^c)$, where n is the size of the instance, k is the input parameter, and c is a constant independent of both n and k [24].

A number of problems of interest in bioinformatics belong to this class. One of the most emblematic ones in the FPT class is Vertex Cover. In Vertex Cover the input is an undirected graph G with n vertices, and a parameter $k \leq n$. The decision problem is whether G contains a set C of k vertices that covers every edge in G (an edge is said to be covered if either one or both of its endpoints are in C).

Informally, *"parameterized complexity is a deal with the Devil of Intractability"* as it aims to confine the combinatorial explosion, often associated to NP-hard problems, to the parameter (which is assumed to be fixed). In practice, a certain "dialogue" develops between the computer scientist and the problem domain, where the aim is to find parameters that, if assumed to be fixed, render the basic problem tractable.

These parameterized problems seem to appear naturally in problems in the area of the Life Sciences. It is often the case that a certain *parameter* can be assumed to be fixed (or bounded by some constant independent of n for all the instances of interest). There are cases in which these parameters represent a structural property, or a property of the solutions sought.

2 Genetic Networks: a case for clique finding algorithms

It is clear that the structure of biological networks has a natural representation as a graph. As a consequence, algorithms for a number of problems that involve optimal subgraph detection become powerful tools for the investigation of biological function. In genetic networks any given gene may have different functions as its activity influences, and is influenced by a number of other genes [59]. A gene in one species may be very similar, at the sequence level, to another gene in some other species. As a consequence, the existence of some subgraphs among the different graphs that represent biological networks help to infer evolutionarily conserved modules of co-expressed genes [4, 7, 52]. This has recently led to approaches that aim to derive phylogenetic trees based on the detection of common subgraphs of metabolic pathways between taxa [38]. Biology, helped by graph-theoretic formulations, is moving from the study of single genes/proteins to "biosystem identification" of the basic common blocks of life.

At the core of this quest is the identification of sets of commonly existing putatively co-regulated genes. For the computer scientist, this basic problem can be formalized as a search for cliques on undirected weighted graphs (see [13] for an excellent review of this topic). For instance, we can make a one-to-one correspondence between genes and vertices, and a co-expression value between two genes is represented by the weight placed on an edge joining a pair of vertices. The inputs to clique are an undirected graph G with n vertices, and a parameter $k \leq n$. The decision problem is whether G contains a clique of size k, that is, a subgraph isomorphic to K_k, the complete graph on k vertices. Since clique is NP-complete there is no known algorithm for deciding clique that runs in time polynomial in the size of the input. Clique can be decided by generating and checking all $\binom{n}{k}$ of vertices selections. But this brute force approach requires $O(n^k)$ time, and is thus prohibitively slow, even for problem instances of only modest size.

There is an strategic advantage in formulating the problem of biological interest as a clique-finding quest. A vertex can be part of any number of *maximum* or *maximal* cliques. This also naturally matches that the gene product can be involved in more than one biological pathway. The problem of generating *all maximal cliques* is also a problem of biological relevance which, in turn, is sometimes related to the maximum clique size. In the context of microarray analysis, the approach here reported can be viewed as “fuzzy clustering” method, as a gene could be a member of different groups. Note that with the denomination of “clustering” several researchers encompass different problems which naturally lead to different approaches (see [8, 9, 10, 35, 37] for some of them). In general, there is always the “one gene to one cluster” relationship. The aim is to partition the gene set into disjoint subsets, so that the genes that correspond to the vertices within each subset have in common some chosen measure of similarity. The method here is novel in that respect.

There are exceptions, however. New clustering techniques, for example those employing factor analysis that not require exclusive cluster membership for single genes [5]. Unfortunately, these tend to produce biologically uninterpretable factors without the incorporation of prior biological information [32]. The approach centered in clique finding does not have this requirement; a decomposition of a graph in a set of maximal cliques could be very informative, since cliques need not be vertex-disjoint.

2.1 Classical and parameterized complexity

Clique is not in class FPT unless a certain conjecture is not true and the $\mathcal{W}$ hierarchy collapses [24] (The $\mathcal{W}$ hierarchy, whose lowest level is FPT, can be viewed as a fixed-parameter analog of the polynomial hierarchy, whose lowest level is $\mathcal{P}$). However, Clique has as complementary dual problem, the Vertex

Cover problem which, although being $\mathcal{NP}$-complete, is in FPT. This fact can be exploited to develop an alternative algorithmic approach for Clique.

We first define what we understand with "complementary dual". If we denote $\overline{G}$, the complement of G. (by definition $\overline{G}$ has the same vertex set as G and all the edges present in G are absent in $\overline{G}$ and vice versa), then a vertex cover of size k in $\overline{G}$ turns out to be exactly the complement of a clique of size $n-k$ in G. If we regard these problems in their optimization versions, the search for a minimum vertex cover in $\overline{G}$ corresponds to the search of a desired maximum clique in G. This fact is very important since there have been great improvements in the fixed-parameter algorithmics for Vertex Cover in the past decade; the fastest known vertex cover algorithm runs in $O(1.2852^k + kn)$ time [16].

2.1.1 Kernelization and Branching

The approach to solve the vertex cover problem is separated in *kernelization* and *branching*. We start reducing an arbitrary input instance of Vertex Cover to a, hopefully, much smaller instance of the same problem (the *kernel*). Buss and Goldsmith [15] have shown that the size of the kernel is in $O(k^2)$. Kernel sizes in $O(k)$ can also be obtained at the expense of methods that rely on linear programming relaxation [40, 51], which tend to be slower in practice.

More recenlty, a new technique, termed *crown reduction*, was introduced for kernelization. A *crown* is an ordered pair (I, H) of subsets of vertices from G that satisfies the following criteria: *(1) $I \neq \emptyset$ is an independent set of G, (2) $H = N(I)$, and (3) there exists a matching M on the edges connecting I and H such that all elements of H are matched.* H is called the *head* of the crown. The *width* of the crown is $|H|$.

The following theorem is then central to this algorithmic approach:

Theorem[1] *Any graph G can be decomposed into a crown (I, H) for which H contains a minimum-size vertex cover of G and so that $|H| \leq 3k$. Moreover, the decomposition can be accomplished in $O(n^{\frac{5}{2}})$ time.*

Branching is applied after the kernel was obtained, in general using a binary tree search. Subtree searches can be spawned off at each level and can be concurrently explored [2]. Up to 64 processors have been used for an application in motif discovery [6]. Contrary to the current folkloric belief in combinatorial optimization, the method allows to solve to optimality large instances [3]. This has allowed to solve huge problems in genomics and proteomics. For large problems, and for particularly difficult subtrees, hardware accelerators have delivered average speedups in the neighborhood of 125 over software-only implementations [23].

2.2 Our final objective: the clique intersection graph

Our final objective is relatively close, but we need first to enumerate all maximal cliques of the graph we are studying. Note that, for a general graph, we can have up to $3^{n/3}$ maximal cliques. However, we expect that in this application the maximum clique size (k_{max}) is not too large. As a consequence, knowing the maximum clique size (which we obtain with the methodology we have just described), allows us to use other data reduction techniques which reduce the size of the graph. Assume that initially this value is $k = k_{max}$. The reduction step basically consists of removing all low-degree vertices with degree less than $k - 1$. This is safe, as these vertices can not be part of any k-clique. We iterate on the application of these rules, as vertices with degree more than $k - 1$ may now have a degree lower than k. When we can not reduce the graph any longer, all k-sets can be checked (to see if they are a k-clique or not) and we ennumerate them all. The whole process is repeated, but now with $k = k - 1$. We keep enumerating all maximal cliques until a specified minimum value, in this study it was $k = 3$.

To give some illustrative figures, for this study we analyzed a dataset with 6,830 genes. A threshold of 0.85 was chosen for the correlation between pairs and as a result we had to process a graph having only 2,281 vertices and 2,619 edges. The enumeration algorithm listed 355 cliques of size between 15 and 3 vertices. With this information, we computed the clique interaction graph defined as follows in this study. Vertices in the clique interaction graph have a one to one correspondence with cliques in the original graph (so we will have a total of 355 vertices). Two vertices in the clique interaction graph are connected with an edge if the two corresponding cliques share at least one vertex. The clique interaction graph is presented in Fig.1.4. In the next sections we will introduce the details of the computational experiment to which it corresponds.

3 A biclique-oriented approach

We have mentioned before how an approach based on clique finding and new techniques based on fixed-parameter tractability have been useful to identify highly correlated groups of genes. In some cases, however, the different samples belong to particular classes of interest to the biologist or the medical researcher. The samples can correspond to either particular clearly separated clinical conditions [57], or to different cellular processes [56, 26], to different parts of an organ (voxelization techniques) [47], different cancer types [34], prediction of tumor outcome [53], different cell lines [11], etc. Now the question is: *"Given that such a labelling on the samples is available, can we identify which is the set of genes that most likely explain the existence of these classes"?.*

As such, this is a generic problem that needs a precise formalization. Since in a typical microarray experiment the number of samples is usually much smaller than the number of genes, its is often the case that several high correlations exist between some genes and the labelling. As a consequence, minimization of the number of genes that can "explain" the labelling should be taken with some caution. It would be possible that we can find a small number of genes for which the following holds. For any two pairs of samples that have different labellings it is always true that there exist at least one gene which has a significantly different expression value. As a consequence, we need to find some new formalization of this problem that would give "robust genetic signatures". With "robust", we mean that the explanation should rely in the co-expression of many genes, as a way of avoiding individual spurious correlations that may dramatically influence the gene selection task. Towards this end, a very useful mathematical formalization has been introduced by Cotta, Sloper and Moscato, the ($(\alpha, \beta) - k-$Feature Set problem). With this, it has been possible to find genetic signatures for Alzheimer disease [47], the molecular classification of cancer [11], and the prediction of US presidential election results [49].

We will see that the $(\alpha, \beta) - k-$Feature Set Problem can be formalized as a problem of finding a certain type of subgraph in a bipartite graph. In addition, such a subgraph contains a biclique $K_{k',k}$ where k' is the minimum of the values α and β.

3.1 The $(\alpha, \beta) - k-$Feature Set Problem

We use the $(\alpha, \beta) - k-$Feature Set Problem as our mathematical formalization of the problem of interest since we aim to obtain *robust* genetic signatures of the different types of cancer. Robustness is obtained via some redundance in the genes/features that allow the discrimination. As a consequence, our genetic signatures guarantee that, if a feasible solution exist for the dataset of interest, at least α genes will help discriminate between any two samples of different classes. In addition, the genetic signature will have at least β genes with similar values between any two samples of the same class.

Cotta, Sloper and Moscato have introduced the $(\alpha, \beta) - k-$FEATURE SET Problem as a generalization of the $k-$FEATURE SET [21]. The problem is trivially $NP-$complete as Davies and Russell proved in 1994 that $k-$FEATURE SET is $NP-$complete [22] ($k-$FEATURE SET problem corresponds to an $\alpha = 1, \beta = 0$ $(\alpha, \beta) - k-$FEATURE SET. Formally:

$(\alpha, \beta) - k-$FEATURE SET

- **Instance**: A set of m examples $X = \{x^{(1)}, \ldots, x^{(m)}\}$, such that for all i, $x^{(i)} = \{x_1^{(i)}, x_2^{(i)}, \ldots, x_n^{(i)}, t^{(i)}\} \in \{0,1\}^{n+1}$, and three integers $k > 0$, and $\alpha, \beta \geq 0$.
- **Question:** Does there exist an $(\alpha, \beta) - k$-feature set S, $S \subseteq \{1, \cdots, n\}$, with $|S| \leq k$ and such that:
  - for all pairs of examples $i \neq j$, if $t^{(i)} \neq t^{(j)}$ there exists $S' = S'(i,j) \subseteq S$ such that $|S'| \geq \alpha$ and for all $l \in S'$, $x_l^{(i)} \neq x_l^{(j)}$?
  - for all pairs of examples $i \neq j$, if $t^{(i)} = t^{(j)}$ there exists $S' \subseteq S$ such that $|S'| \geq \beta$ and for all $l \in S'$, $x_l^{(i)} = x_l^{(j)}$?

where the set S' is not the same for all pairs of examples so we have written $S' = S'(i,j)$.

3.2 Parameterized intractability

We have seen that the the $NP-$completeness result for k-FEATURE SET implies that there currently exists no polynomial-time algorithm for this problem. We have seen before what happens with CLIQUE that is not in FPT but its parametric dual is. The first natural parameter to consider is the cardinality of the feature subset. Cotta and Moscato have proved the following result:

Theorem 1.1 *Unless $FPT = W[2]$, the $(\alpha, \beta) - k-$*FEATURE SET *problem is not fixed-parameter tractable for parameter k.*

The proof follows from the main result of [20] where it was proved that the $k-$FEATURE SET problem is $W[2]$-complete ($W[2]$ is a parameterized class comprising substantially harder problems than FPT). There exist a current widely believed conjecture in parameterized complexity, that $FPT \neq W[1]$. Also, since $W[1] \subseteq W[2]$ we do not expect that a fixed-parameter algorithm can easily be found for the $(\alpha, \beta) - k-$FEATURE SET problem. Unlike CLIQUE, in this case we can not rely on an FPT algorithm, so heuristic algorithms are a resonable alternative for this problem. However, even if the problem is not FPT, it may have powerful instance reduction rules that would shrink its size. These rules reduce the size of the problem instance and any solution of the original instance can be obtained from the solutions of the simplified instance, and vice versa. The application of reduction rules may turn large instances of $NP-$hard problems into trivial instances or small instances solvable by hand or by enumeration [58]. Next, we will present greedy heuristic coupled with reduction rules for this purpose.

3.3 Reduction rules for the $(\alpha, \beta) - k-$Feature Set Problem

We will explain these rules with the help of the RED-BLUE DOMINATING SET problem. We start considering the case of the $(1, 0) - k-$FEATURE SET problem. If I is an instance of this problem we can transform it to an instance of the RED-BLUE DOMINATING SET using the following procedure: we denote with $G(V_1 \cup V_2, E)$ a bipartite graph such that:

- there is a red vertex $g_i \in V_2$ for each feature/gene in I, i.e., $|V_2| = n$.
- there is a blue vertex $p_{jk} \in V_1$ for each pair of examples $x^{(j)}$ and $x^{(k)}$ such that $t^{(j)} \neq t^{(k)}$.
- there is an edge (g_i, p_{jk}) whenever $x_i^{(j)} \neq x_i^{(k)}$.

The reader can easily verify that I is a yes-instance if, and only if, there exists a red dominating set $D \subseteq V_2$ such that $|D| \leq k$ and it can be generalized to the $(\alpha, 0) - k-$FEATURE SET (requesting that D be $\alpha-$dominating, i.e., that at least α vertices in D dominate each vertex in V_1 [36]). The final generalization to the $(\alpha, \beta) - k-$FEATURE SET problem is easy from here: a tripartite graph $G(V_1 \cup V_2 \cup V_3, E)$ is constructed such that V_1, V_2, and the edges among vertices in them are as described before, and

- there is a blue vertex $c_{jk} \in V_3$ for each pair of examples $x^{(j)}$ and $x^{(k)}$ such that $t^{(j)} = t^{(k)}$.
- there is an edge (g_i, c_{jk}) whenever $x_i^{(j)} = x_i^{(k)}$.

then an instance I would be a yes-instance if, and only if, $D \subseteq V_2$ $\alpha-$dominates V_1, $\beta-$dominates V_3, and $|D| \leq k$.

We associate an auxiliary integer variable r_v with each vertex $v \in V_1 \cup V_3$; such that, initially, $r_p = \alpha$ for each $p \in V_1$, and $r_c = \beta$ for each $c \in V_3$; let $G(v) = \{g \in V_2 \mid (g, v) \in E\}$ be the set of vertices (genes) dominating vertex $v \in V_1 \cup V_3$; conversely, let $N(g) = \{v \in V_1 \cup V_3 \mid (g, v) \in E\}$ be the vertices in $V_1 \cup V_3$ dominated by gene $g \in V_2$. We apply then three basic rules for this problem following:

R1. For each $v \in V_1 \cup V_3$ such that $r_v = |G(v)|$ do

i. For each $g \in G(v)$, mark g as belonging to the solution.

ii. Delete v from G.

R2. For each $v \in V_1 \cup V_3$ such that $r_v \leq 0$ delete v from G.

R3. For each $v_1, v_2 \in V_1 \cup V_3$, $v_1 \neq v_2$ such that $r_{v_1} \geq r_{v_2}$ and $G(v_1) \subseteq G(v_2)$, delete v_2 from G.

If a gene is marked, or a vertex is deleted, the following actions are taken:

Gene marking [g]: For each $v \in N(g)$ do

i. $r_v \leftarrow r_v - 1$.

ii. $G(v) \leftarrow G(v) \setminus \{g\}$.

Vertex deleting [v]: For each $g \in G(v)$ do $N(g) \leftarrow N(g) \setminus \{v\}$

These rules greatly simplify the original instance by marking genes that are bound to appear in the final solution, and removing *subsumed* vertices, i.e., vertices that will be dominated for sure upon domination of another vertex. The application of these rules is interleaved until the the graph cannot be further simplified.

3.4 Discretization of numeric values

In data mining, an important problem is to determine, given numeric value information, a reasonable discretization. We note that the $(\alpha, \beta) - k-$FEATURE SET Problem was defined as having a boolean input matrix. This said, we need to find, for each gene a threshold value that dicotomizes the expression. For this study, we have used to methods, one proposed by Fayyad and Irani [28] and another one in which an Evolutionary Search strategy is applied to find a large biclique and employs the reduction rules described above [21]. The methods give similar, but different results, and we currently use them as complementary approaches to retrieve many relevant genes [47].

4 A hamiltonian path-motivated approach for gene ordering

The genetic signatures found required to be presented in a meaningful way. A number of approaches for the problem of ordering gene expression patterns have been based on combinatorial optimization. Most of the time, since the number of genes to be ordered can be large (several thousands), the researchers have resorted to some heuristic, as the basic problem of finding a hamiltonian path of minimum weight is NP-hard. As a consequence, a number of heuristic and metaheuristic algorithms have been developed, possibly *Self Organizing Maps* (SOMs) is one of the most used. Implementations of SOMs have found their way into some commercial packages for microarray data analysis. In addition, some software packages (both commercial and on the public domain) use some form of hierarchical clustering and *ad hoc* heuristics for the final ordering of the leaves of the dendogram that represents the final clustering. Under some special, but still quite practical conditions, the optimal arrangement can be solved in polynomial time [12].

However, it has been recently recognized that this type of approach may lead, in some cases, to results that do not entirely satisfy the Life Sciences researchers. Gene members of the same functional group are scattered in those orderings. In [19], we have introduced an alternative objective function to optimize. If a hierarchical clustering is not given as extra constraint, this leads to a problem that is NP-hard, as it contains the minimum weight hamiltonian path problem as a special case.

The input is an integer matrix of gene expression values $G = g_{ij}, 1 \leq i \leq n, 1 \leq j \leq m$, where n is the number of genes, m is the number of samples, and g_{ij} represents the level of activity of gene i under condition j. We are also given a function that allows us, given any two patterns, to compute the degree of dissimilarity between them. We need to find a permutation of the genes' names $\pi = (\pi_1, \pi_2, ..., \pi_n)$, such that the genes with the most similar expression patterns are close to each other in the sought permutation.

Now, the task is to find the permutation that minimizes the following objective function

$$TotalCost(\pi) = \sum_{l=1}^{n} \sum_{i=\max(l-w_s,1)}^{\min(l+w_s,n)} (w_s - |l-i| + 1).D[\pi_l, \pi_i] \tag{1.1}$$

where the window size is $2w_s + 1$ (the number of genes involved in each partial distance calculation) and $D[\pi_l, \pi_i]$ represents the measure of dissimilarity between π_l and π_i. For this chapter, the parameter w_s was fixed to $\lfloor 0.01n \rfloor$ (see Ref. [18] for the influence of that this parameter in the final solution). This objective function was also recently adopted by T. Conrad in his award-winning paper on the visualization and analysis of metabolic pathways [17].

During the last decade, several combinatorial optimization problems for finding an optimal permutation have been addressed with memetic algorithms [46, 44, 14]. We also use this metaheuristic to address this problem. In addition, memetic algorithms have been introduced with the motivation of obtaining an almost linear speed-up when parallelized [43] due to its inherent asychnronism and low inter-processor communication requirements. In Ref. [48], it has been shown that this algorithm is very robust to individual noise measurements and was used to order genetic signatures of Alzheimer's disease [47]. They have also been applied to cancer's genetic signatures [11]. The next section shows an illustrative example (Fig. 1.1) of its performance and a comparison with some of the most used methods available on the public domain.

5 Computational experiments and results

We present results on the application of these three techniques using a microarray dataset of a number of cell-lines originating from different cancers. To ensure the reproducibility of our techniques, we have chosen to work with cell-lines and a public domain dataset called NCI60. The original dataset and a clustering analysis was introduced by Ross *et al.* [55]. In addition, we will show how a memetic algorithm, using a similarity measure between pairs of genes (or pairs of samples), is able to obtain permutations of the rows and colums such that the final layout is highly correlated and highlights the major common groups.

In Fig. 1.1, we present the results of three different algorithms for ordering microarray data. We have used an image to illustrate their main characteristics and the memetic algorithm is later used to order the genetic signatures of Figs. 1.2 and 1.3. Fig. 1.1.a. shows the original image that contains 489 rows and

971 colums of grey-scale pixel values. The rows and colums are randomly permuted to obtain Fig. 1.1.b., illustrating the task we have on real data. We present results of two of the best algorithms for analysing microarray datasets that are available on the public domain. Fig. 1.1.c., shows the results of a hierarchical clustering algorithm *European Bioinformatics Initiative* (EBI) as part of the *Expression Profiler* software tool [2]. Fig. 1.1.d., proposed by Eisen *et al.* [25], a hierarchical clustering algorithm that also performs the ordering of the genes [3]. Finally Fig. 1.1.e. shows the result of our memetic algorithm [19], and in the three cases we have used the same algorithms to order both the rows and columns. In the rest of the chapter, we will only use the memetic algorithm to order the genetic signatures shown in all the other figures.

The original NCI60 dataset has 64 samples from 60 cell lines (i.e. two cell lines have three samples each in the set). A total of 9,703 human cDNAs have been spotted on glass microscope slides; the cDNAs thus included around 8,000 different genes. We have worked with the dataset that corresponds to Fig. 2 of Ref. [55], which helps to illustrate the power of our combinatorial approach. Again, for the purpose of illustration of the technique, we have selected only a subset of the samples that corresponds to four types of cancer: Melanoma (SK-MEL-5, M-14, SK-MEL-28, UACC-257, MALME-3M, UACC-62, SK-MEL-2A), Leukaemia (RPMI-8226, K562, K562, K562, HL-60, MOLT-4, CCRF-CEM, SR), Colon (HCT-116, SW-620, HCT-15, KM12, HCC-2998, COLO205, HT-29) and Renal (A498, RXF-393, a786-0, CAKI-1, ACHN, UO-31, TK-10). This means that we have excluded for the purpose of this study cell lines LOXIMVI (Melanoma), as well as SN12C and SNB-75 (both Renal). The reason is that they seem to have, overall, a very different gene expression pattern than the others from the same class. While the reason of removing for consideration was only done to help illustrate better the power of the basic technique (providing very distinctive genetic signatures), other issues should be considered. For instance, have these cell lines remained with molecular characteristic of their parent tumours ? Again, for the purpose of the illustration case, we would not include them in this study.

The first question that we would like to address could be informally phrased as: *Which are the genes that are a genetic signature of colon cancer ?*. An analogous question can be asked for the three other different types. We realize that this basic question is implicit in the analysis of [55] and is also implicit in several

[2] `http://ep.ebi.ac.uk/EP/EPCLUST/`

[3] `http://rana.lbl.gov/EisenSoftware.htm`

other analysis. In [55], an attempt has been made to identify "clusters" of genes that are related to a given type of cancer. Unfortunately, the authors only used the information given by the clustering algorithm. This has lead them to identify genetic signatures containing only the highly-expressed genes.

Fig.1.2 shows the genetic signatures of the Renal, Melanoma, Colon and Leukaemia [55] cell lines listed above. Figs.1.2.a-d, correspond to an different $(\alpha, \beta) - k$-feature sets obtained. All these genetic signatures have been obtained using a methdology first employed in [11]. Initially, an $(\alpha_{max}, \beta = 0) - k$-feature set is obtained, where α_{max} is the maximum obtainable discrimination that can be guaranteed for all pairs of samples. This means that there exist at least a pair of samples that belong to different classes (Renal vs. non-Renal) such that we can only find α_{max} differentially expressed genes. For the Renal vs. non-Renal case, we have found $\alpha_{max} = 768$. The parameter β is set to zero, thus not considering the within-class similarity. We have then found a $(768, \beta = 0) - k$-feature set with the objective of minimizing the number of genes in the signature (k). We found it and requires only 1,073 genes. We then proceed trying to increase the within-class similarity of our genetic signatures without incrementing the number of genes. We stop when we obtain a maximum value of β such that if we increase it by at least one unit, we can not obtain a genetic signatures with the optimal value of 1,073 (obtained when we aimed to find the minimum cardinality $(768, \beta = 0) - k$-feature set). Fig.1.2.a shows the result: a genetic signature for Renal cancer (relative to the other three types), that corresponds to a $(\alpha_{max} = 768, \beta = 655), k_{opt} = 1,073)$ feature set (where the genes are the features in this case). Analogously, Figs.1.2.b-d correspond to the genetic signatures of Melanoma $(\alpha_{max} = 714, \beta = 673, k_{opt} = 985)$, Colon $(\alpha_{max} = 358, \beta = 277, k_{opt} = 521)$, and Leukaemia $(\alpha_{max} = 814, \beta = 743, k_{opt} = 1,253)$, respectively.

In total, the union of the four genetic signatures has 3,832 genes, but only 2,998 are different. Fig.1.2.e finally displays those 2,998 genes. In all cases, the order of the genes was found with the memetic algorithm allowing to identify different groups of up and down regulated genes. When the union of the four signatures is displayed as a whole, the within-class differences of the different tumors start to become evident. A clear example is given by Leukaemia's cell lines RPMI-8226 and SR, Colon's HCT-116, and Melanoma's SK-MEL-5.

6 Conclusions

We have shown how a combinatorial optimization approach for the problem of pattern recognition in microarray data helps to provide useful solutions to classify hundreds of genes involved in a disease. These approaches are complementary to statistical methdologies which, in turn, can benefit from the extraordinary performance of these methods to organize the data and extract interesting hypothesis for further testing and validation.

We have used publicly available data, to ensure reproducibility and for illustrative purposes. We have selected the NCI60 dataset, since it has been available since 2000 and some researchers have regarded it as "uninformative" in the past. Our results seem to indicate that this label may be related to the inadequacy of previous methodologies rather than something intrinsic to this dataset. We have shown how a combination of powerful metaheuristics and exact algorithms allow to find genetic signatures for some of the major cancer groups in the dataset.

If the genetic signatures that we have found correspond to characteristic of the tumour types *in vivo*, they may have several uses. At the very least, they can help in determining the true origin of a metastases without obvious primary. However, possibly the most important role of this type of analysis is to provide a molecular classification of cancer which is novel and independent from traditional clinical taxonomies. Finally, if this classification correlated well with the characteristics *in vivo*, they may have a central role in personalized medicine. It could then be possible to link patients with the most appropriate tumour chemotherapy, a dreamed scenario which may be closer than we imagine.

Acknowledgement

P.M. would like to thank Alexandre Mendes, Pritha Mahata, and Wagner Costa for their invaluable help and discussions during the preparation of this manuscript. P.M. also wants to thank Profs. Rodney Scott, John Forbes and Alan Coates for discussions and their comments.

References

[1] F. N. Abu-Khzam, R. L. Collins, M. R. Fellows, M. A. Langston, W. H. Suters, and C. T. Symons. Kernelization algorithms for the vertex cover problem: Theory and experiments. In *Proceedings, Workshop on Algorithm Engineering and Experiments (ALENEX)*, 2004.

[2] F. N. Abu-Khzam, M. A. Langston, and P. Shanbhag. Scalable parallel algorithms for difficult combinatorial problems: A case study in optimization. In *Proceedings, International Conference on Parallel and Distributed Computing and Systems (PDCS)*, pages 563–568, 2003.

[3] F. N. Abu-Khzam, M. A. Langston, P. Shanbhag, and C. T. Symons. Scalable parallel algorithms for FPT problems. Technical Report UT-CS-04-524, Department of Computer Science, University of Tennessee, 2004.

[4] U. Alon. Biological networks: the tinkerer as an engineer. *Science*, 301:1866–1867, 2003.

[5] O. Alter, P. O. Brown, and D. Botstein. Singular value decomposition for genome-wide expression data processing and modeling. *Proceedings of the National Academy of Sciences*, 97:10101–10106, 2000.

[6] N. E. Baldwin, R. L. Collins, M. A. Langston, M. R. Leuze, C. T. Symons, and B. H. Voy. High performance computational tools for motif discovery. In *Proceedings, IEEE International Workshop on High Performance Computational Biology (HiCOMB)*, 2004.

[7] A-L Barabási and Z. N. Oltvai. Network biology: Understanding the cell's functional organization. *Nature Reviews Genetics*, 5:101–113, 2004.

[8] A. Bellaachia, D. Portnoy, Y. Chen, and A. G. Elkahloun. E-CAST: A data mining algorithm for gene expression data. In *Workshop on Data Mining in Bioinformatics*, pages 49–54, 2002.

[9] A. Ben-Dor, L. Bruhn, N. Friedman, I. Nachman, M. Schummer, and Z. Yakhini. Tissue classification with gene expression profiles. *Journal of Computational Biology*, pages 54–64, 2000.

[10] A. Ben-Dor, R. Shamir, and Z. Yakhini. Clustering gene expression patterns. *Journal of Computational Biology*, 6(3/4):281–297, 1999.

[11] R. Berretta, A. Mendes, and P. Moscato. Integer programming models and algorithms for molecular classification of cancer from microarray data. In Estivill-Castro [27], pages 361–370.

[12] T. Biedl, B. Brejova, E.D. Demaine, A.M. Hamel, and T. Vinar. Optimal arrangement of leaves in the tree representing hierarchical clustering of gene expression data. Technical Report 2001-14, University of Waterloo, Canada, 2001.

[13] I. Bomze, M. Budinich, P. Pardalos, and M. Pelillo. The maximum clique problem. In D.-Z. Du and P. M. Pardalos, editors, *Handbook of Combinatorial Optimization*, volume 4. Kluwer Academic Publishers, 1999.

[14] L.S. Buriol, P.M. França, and P. Moscato. A new memetic algorithm for the asymmetric traveling salesman problem. *Journal of Heuristics*, 10(5):483–506, 2004.

[15] J. F. Buss and J. Goldsmith. Nondeterminism within $\mathcal{P}$. *SIAM Journal on Computing*, 22:560–572, 1993.

[16] J. Chen, I. Kanj, and W. Jia. Vertex cover: further observations and further improvements. *Journal of Algorithms*, 41:280–301, 2001.

[17] T. Conrad. New approaches for visualizing and analysing metabolic pathways. In *2nd Annual Australian Undergraduate Students' Computing Conference*, pages 48–58, Melbourne, Victoria, Australia, December 2004. Australian Undergraduate Students' Computing Conference. Best Conference Paper Award.

[18] C. Cotta, A. Mendes, V. Garcia, P. Franca, and P. Moscato. Applying memetic algorithms to the analysis of microarray data. In *Proceedings of EvoBIO2003 - 1st European Workshop on Evolutionary Bioinformatics, Lecture Notes in Computer Science n. 2611*, pages 22–32, 2003.

[19] C. Cotta, A. Mendes, V. Garcia, P. França, and P. Moscato. Applying memetic algorithms to the analysis of microarray data. In G. Raidl et al., editor, *EvoBIO2003 - 1st European Workshop on Evolutionary Bioinformatics, Lecture Notes in Computer Science n. 2611*, volume 2611 of *Lecture Notes in Computer Science*, pages 22–32. Springer-Verlag, Berlin, 2003.

[20] C. Cotta and P. Moscato. The k-Feature Set problem is $W[2]$-complete. *Journal of Computer and Systems Science*, 67(4):686–690, 2003.

[21] C. Cotta, C. Sloper, and P. Moscato. Evolutionary search of thresholds for robust feature set selection: Application to the analysis of microarray data. In G.R. Raidl, S. Cagnoni, J. Branke, D. Corne, R. Drechsler, Y. Jin, C.G. Johnson, P. Machado, E. Marchiori, F. Rothlauf, G.D. Smith, and G. Squillero, editors, *EvoWorkshops*, volume 3005 of *Lecture Notes in Computer Science*, pages 21–30. Springer, 2004.

[22] S. Davies and S. Russell. NP-completeness of searches for smallest possible feature sets. In R. Greiner and D. Subramanian, editors, *AAAI Symposium on Intelligent Relevance*, pages 41–43, New Orleans, 1994. AAAI Press.

[23] M. Dorai, D. W. Bouldin, M. A. Langston, and G. D. Peterson. FPGA-based solutions for the branching phase of fixed-parameter tractable computations. Manuscript, 2004.

[24] R. G. Downey and M. R. Fellows. *Parameterized Complexity.* Springer-Verlag, 1999.

[25] M.B. Eisen, P.T. Spellman, P.O. Brown, and D. Botstein. Cluster analysis and display of genome-wide expression patterns. In *Proceedings of the National Academy of Sciences of the USA 95*, pages 14863–14868, 1998.

[26] R. Elkon, C. Linhart, R. Sharan, R. Shamir, and Y. Shiloh. Genome-wide in silico identification of transcriptional regulators controlling the cell cycle in human cells. *Genome Research*, 13(5):773–780, 2003.

[27] V. Estivill-Castro, editor. *Computer Science 2005, Twenty-Eighth Australasian Computer Science Conference (ACSC2005), Newcastle, NSW, Australia, January/February 2005*, volume 38 of *CRPIT*. Australian Computer Society, 2005.

[28] U.M. Fayyad and K.B. Irani. Multi-interval discretization of continuous-valued attributes for classification learning. In *IJCAI*, pages 1022–1029, 1993.

[29] M. R. Fellows and M. A. Langston. Nonconstructive tools for proving polynomial-time decidability. *Journal of the ACM*, 35:727–739, 1988.

[30] M. R. Fellows and M. A. Langston. On search, decision and the efficiency of polynomial-time algorithms. *Journal of Computer and Systems Sciences*, 49:769–779, 1994.

[31] D. Gershon. Microarray technology - an array of opportunities. *Nature*, 416(6883):885, 2002.

[32] M. Girolami and R. Breitling. Biologically valid linear factor models of gene expression. *Bioinformatics*, 2004. to appear.

[33] F. Glover and M. Laguna. *Tabu Search.* Kluwer Academic Publishers, Norwell, Massachusetts, 1997.

[34] T.R. Golub, D.K. Slonim, P. Tamayo, C. Huard, M. Gaasenbeek, J.P. Mesirov, H. Coller, M.L. Loh, J.R. Downing, M.A. Caligiuri, C.D. Bloomfield, and E.S. Lander. Molecular classification of cancer: Class discovery and class prediction by gene expression monitoring. *Science*, 286:531–537, 1999.

[35] P. Hansen and B. Jaumard. Cluster analysis and mathematical programming. *Mathematical Programming*, 79(1-3):191–215, 1997.

[36] J. Harant, A. Pruchnewski, and M. Voigt. On dominating sets and independent sets of graphs. *Combinatorics, Probability and Computing*, 8:547–553, 1999.

[37] E. Hartuv, A. Schmitt, J. Lange, S. Meier-Ewert, H. Lehrachs, and R. Shamir. An algorithm for clustering cDNAs for gene expression analysis. In *RECOMB*, pages 188–197, 1999.

[38] M. Heymans and A.K. Singh. Deriving phylogenetic trees from the similarity analysis of metabolic pathways. In *ISMB (Supplement of Bioinformatics)*, pages 138–146, 2003.

[39] S. Hor, H. Pirzer, L. Dumoutier, F. Bauerand, S. Wittmann, H. Sticht, J.C. Renauld, R. de Waal Malefyt, and H. Fickenscher. The T-cell lymphokine interleukin-26 targets epithelial cells through the interleukin-20 receptor 1 and interleukin-10 receptor 2 chains. *J. Biol. Chem.*, 279(32):3343–3351, Aug 2004.

[40] S. Khuller. The vertex cover problem. *ACM SIGACT News*, 33:31–33, June 2002.

[41] J.M. Kim, H.Y. Sohn, S.Y. Yoon, J.O. Yang, J.H. Kim, K.S. Song, S.M. Rho, H.S. Yoo, Y.S. Kim, J.G. Kim, and N.S. Kim. Identification of gastric cancer-related genes using a cdna microarray containing

novel expressed sequence tags expressed in gastric cancer cells. *Clinical Cancer Research*, 11:473–482, January 2005.

[42] R. Kothapalli, S. Yoder, S. Mane, and T. Loughran. Microarray results: How accurate are they? *BMC Bioinformatics*, 3(22):1–10, 2002.

[43] A. Mendes, C. Cotta, V. Garcia, P. França, and P. Moscato. Gene ordering in microarray data using parallel memetic algorithms. In *ICPP Workshops*, pages 604–611. IEEE Computer Society, 2005.

[44] P. Merz and B. Freisleben. Memetic algorithms for the traveling salesman problem. *Complex Systems*, 13(4):297–345, 2001.

[45] P. Moscato. On Evolution, Search, Optimization, Genetic Algorithms and Martial Arts: Towards Memetic Algorithms. Technical Report Caltech Concurrent Computation Program, Technical Report. 826, California Institute of Technology, Pasadena, California, USA, 1989.

[46] P. Moscato. An introduction to population approaches for optimization and hierarchical objective functions: a discussion on the role of tabu search. *Annals of Operations Research*, 41:85–121, 1993.

[47] P. Moscato, R. Berretta, M. Hourani, A. Mendes, and C. Cotta. Genes related with alzheimer's disease: A comparison of evolutionary search, statistical and integer programming approaches. In F. Rothlauf, J. Branke, S. Cagnoni, D.W. Corne, R. Drechsler, Y. Jin, P. Machado, E. Marchiori, J. Romero, G.D. Smith, and G. Squillero, editors, *EvoWorkshops*, volume 3449 of *Lecture Notes in Computer Science*, pages 84–94. Springer, 2005.

[48] P. Moscato, R. Berretta, and A. Mendes. A new memetic algorithm for ordering datasets: Applications in microarray analysis. In *to appear in Proceedings of the sixth Metaheuristics International Conference*, 2005.

[49] P. Moscato, L. Mathieson, A. Mendes, and R. Berretta. The electronic primaries: Predicting the u.s. presidency using feature selection with safe data reduction. In Estivill-Castro [27], pages 371–380.

[50] P. Moscato and M. G. Norman. A 'memetic' approach for the traveling salesman problem implementation of a computational ecology for combinatorial optimization on message-passing systems. In

M. Valero, E. Onate, M. Jane, J. L. Larriba, and B. Suarez, editors, *Parallel Computing and Transputer Applications*, pages 177–186. IOS Press, Amsterdam, 1992.

[51] G. L. Nemhauser and L. E. Trotter. Vertex packings: Structural properties and algorithms. *Mathematical Programming*, 8:232–248, 1975.

[52] Z. N. Oltvai and A.-L. Barabási. Systems biology. Life's complexity pyramid. *Science*, 298:763–764, 2002.

[53] S.L. Pomeroy, P. Tamayo, M. Gaasenbeek, L.M. Sturla, M. Angelo, M.E. McLaughlin, J.Y.H. Kim, L.C. Goumnerova, P.M. Black, C. Lau, J.C. Allen, D. Zagzag, J.M. Olson, T. Curran, C. Wetmore, J.A. Biegel, T. Poggio, S. Mukherjee, R. Rifkin, A. Califano, G. Stolovitzky, D.N. Louis, J.P. Mesirov, E.S. Lander, and T.R. Golub. Prediction of central nervous system embryonal tumour outcome based on gene expression. *Nature*, 415(6870):436–442, 2002.

[54] T. Riaz, W. Yi, and K.B. Li. A tabu search algorithm for post-processing multiple sequence alignment. *J. Bioinform. Comput. Biol.*, 3(1):145–56, 2005.

[55] Douglas T. Ross, Uwe Scherf, Michael B. Eisen, Charles M. Perou, Christian Rees, Paul Spellman, Vishwanath Iyer, Stefanie S. Jeffrey, Matt Van de Rijn, Mark Waltham, Alexander Pergamenschikov, Jeffrey C.F. Lee, Deval Lashkari, Dari Shalon, Timothy G. Myers, John N. Weinstein, David Botstein, and Patrick O. Brown. Systematic variation in gene expression patterns in human cancer cell lines. *Nature Genetics*, 24:227– 235, 2000.

[56] P.T. Spellman, G. Sherlock, M.Q. Zhang, V.R. Iyer, K. Anders, M.B. Eisen, P.O. Brown, D. Botstein, and B. Futcher. Comprehensive identification of cell cycle-regulated genes of the yeast saccharomyces cerevisiae by microarray hybridization. *Mol. Biol. Cell*, 9:3273–3297, 1998.

[57] L.J. van't Veer, H.Y. Dai, M.J. van de Vijver, Y.D.D He, A.A.M. Hart, M. Mao, H.L. Peterse, K. van der Kooy, M.J. Marton, A.T. Witteveen, G.J. Schreiber, R.M. Kerkhoven, C. Roberts, P.S. Linsley, R. Bernards, and S.H.Friend. Gene expression profiling predicts clinical outcome of breast cancer. *Nature*, 415(6871):530–536, 2002.

[58] K. Weihe. Covering trains by stations or the power of data reduction. In R. Battiti and A.A. Bertossi, editors, *Proceedings of Algorithms and Experiments (*ALEX 98*)*, pages 1–8, Trento, Italy, 1998.

[59] L.F. Wu, T.R. Hughes, A.P. Davierwala, M.D. Robinson, R. Stoughton, and S.J. Altschuler. Large-scale prediction of saccharomyces cerevisiae gene function using overlapping transcriptional clusters. *Nature Genetics*, 31(3):255–265, 2002.

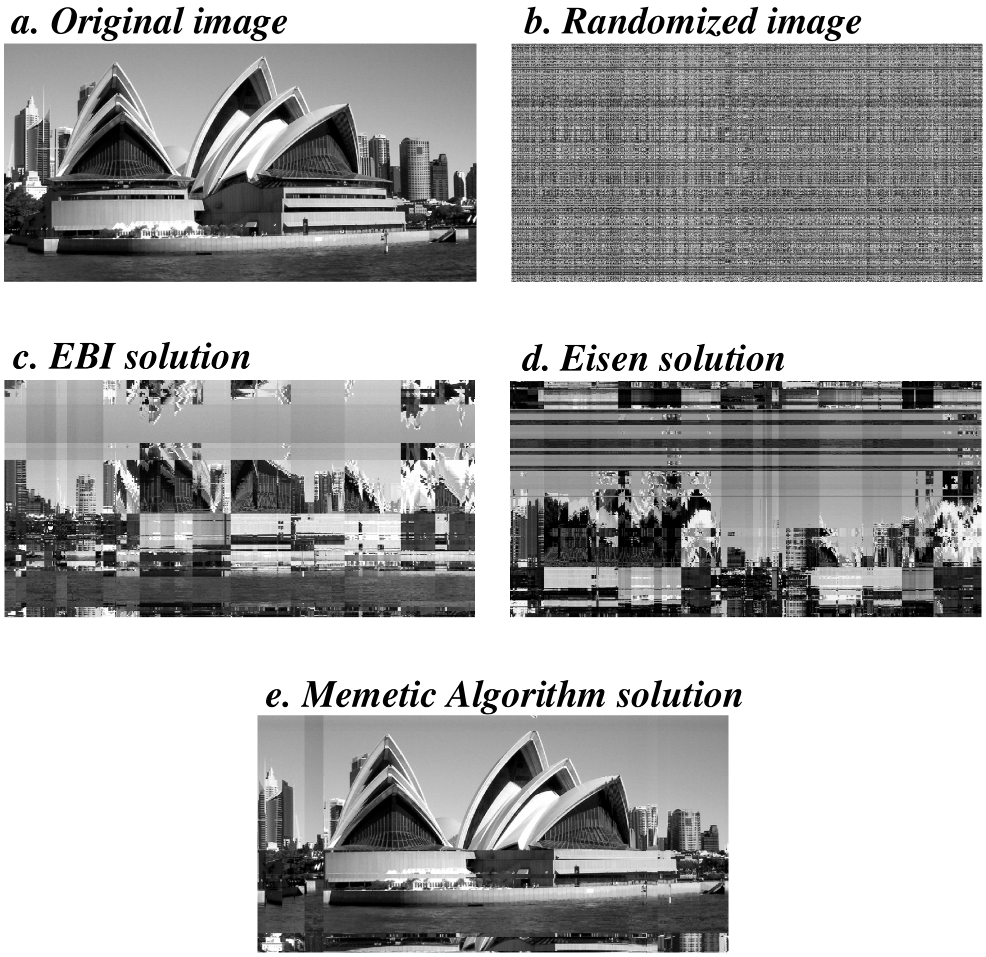

Figure 1.1: Opera House based images. (a) The original image, containing 489 rows and 971 columns and no noise; (b) a random permutation of rows and columns; (c) the solution from EBI hierarchical clustering; (d) the solution from Eisen's hierarchical clustering; (e) the memetic algorithm solution.

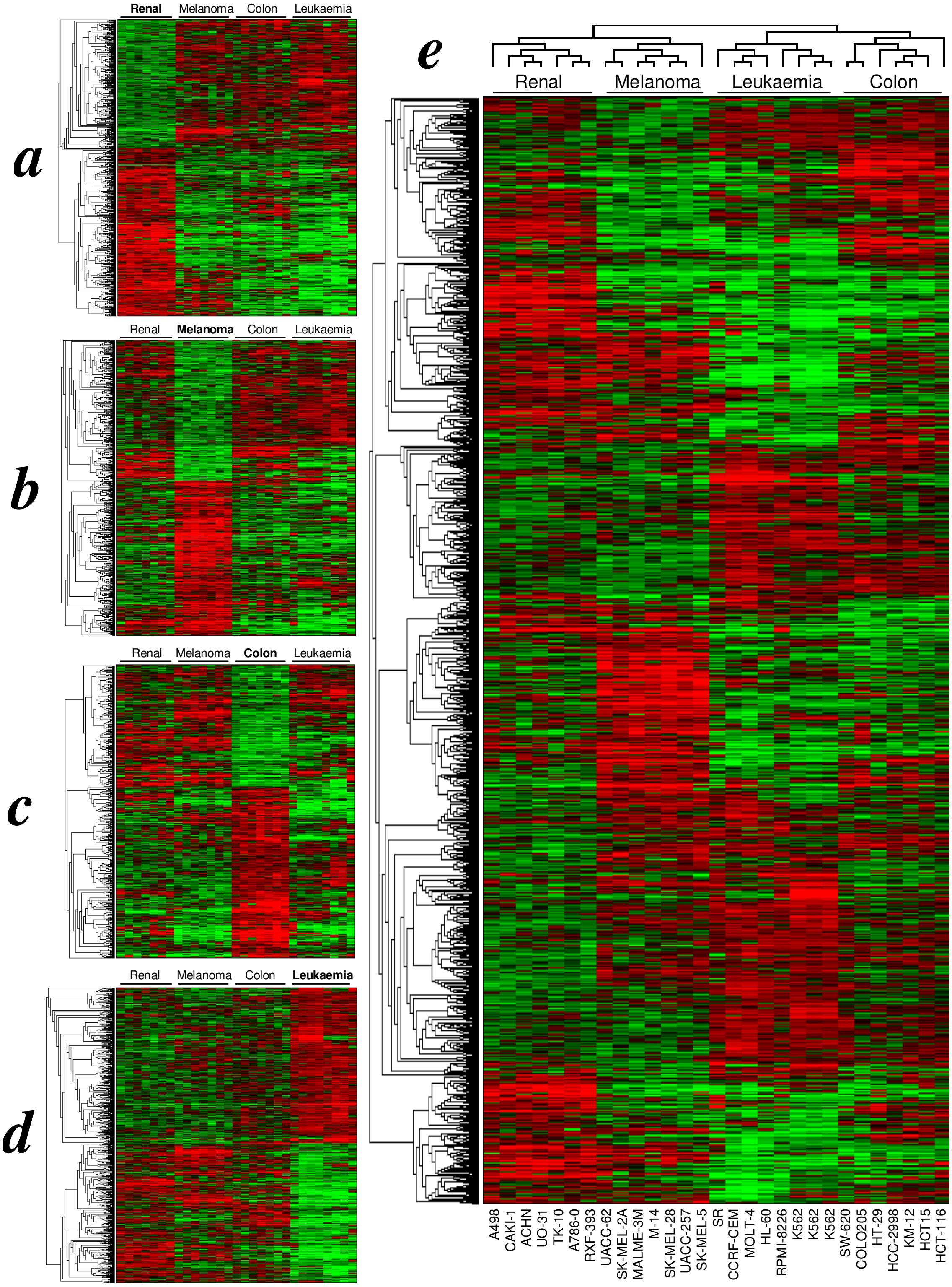

Figure 1.2: Signatures of the four types of cancer: (a) Renal, (b) Melanoma, (c) Colon and (d) Leukaemia. The image on the right (e) is the union of the four sets on the left and contains the profiles of 2,998 genes in 29 cell lines.

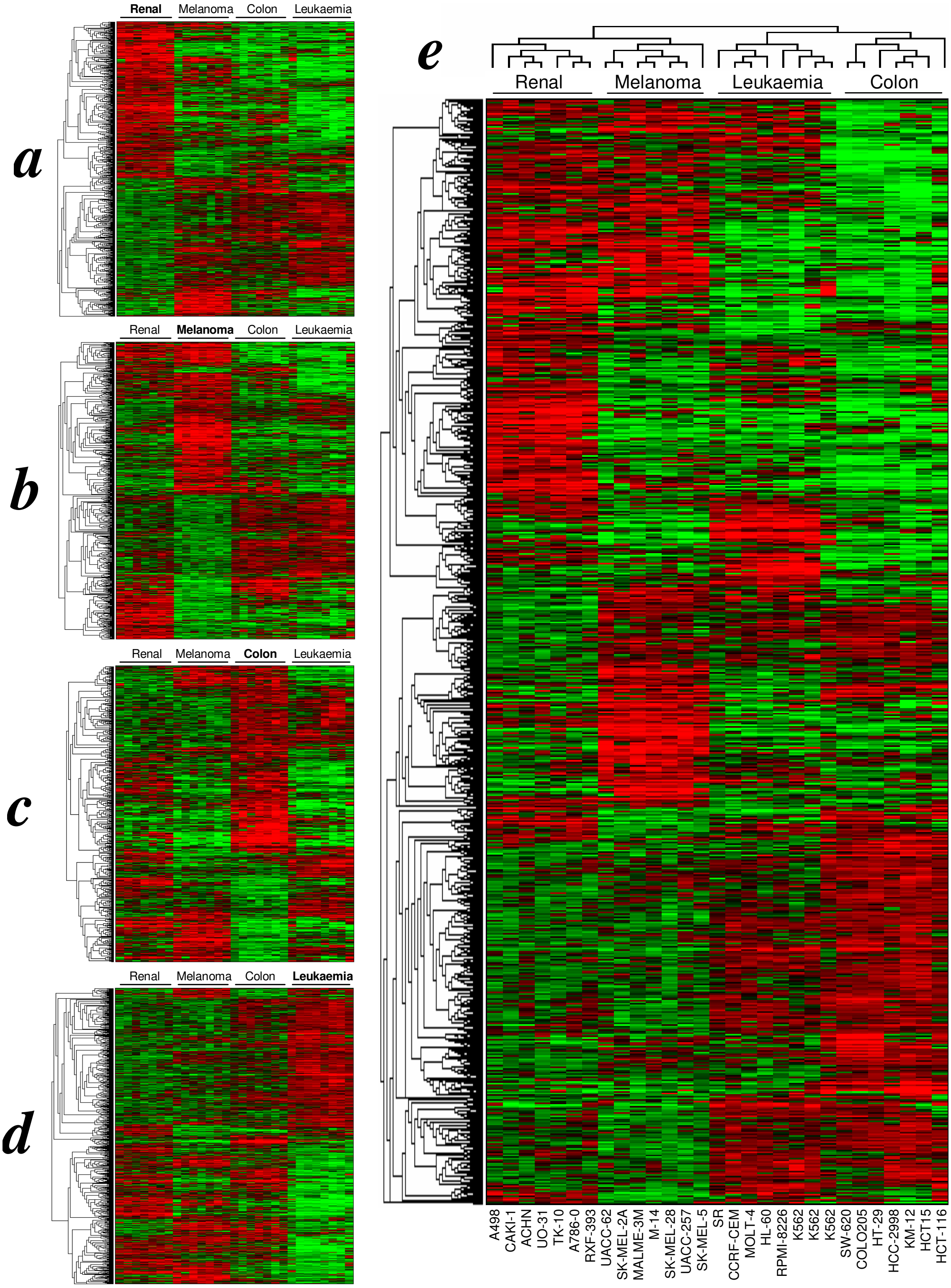

Figure 1.3: The genetic signatures of the four types of cancer found with the Evolutionary Search (ES) heuristic introduced in [21]. They discriminate Renal (a), Melanoma (b), Colon (c), and Leukaemia (d). Their union (2,259 genes) is shown in (e). The signatures have 1,120, 1,035, 556 and 1,255 genes respectively. The ES has an advantage for particular values of (α, β) where exact searches are too time consuming. In this case it shows comparatively similar results to the exact algorithm used for Fig. 1.2.

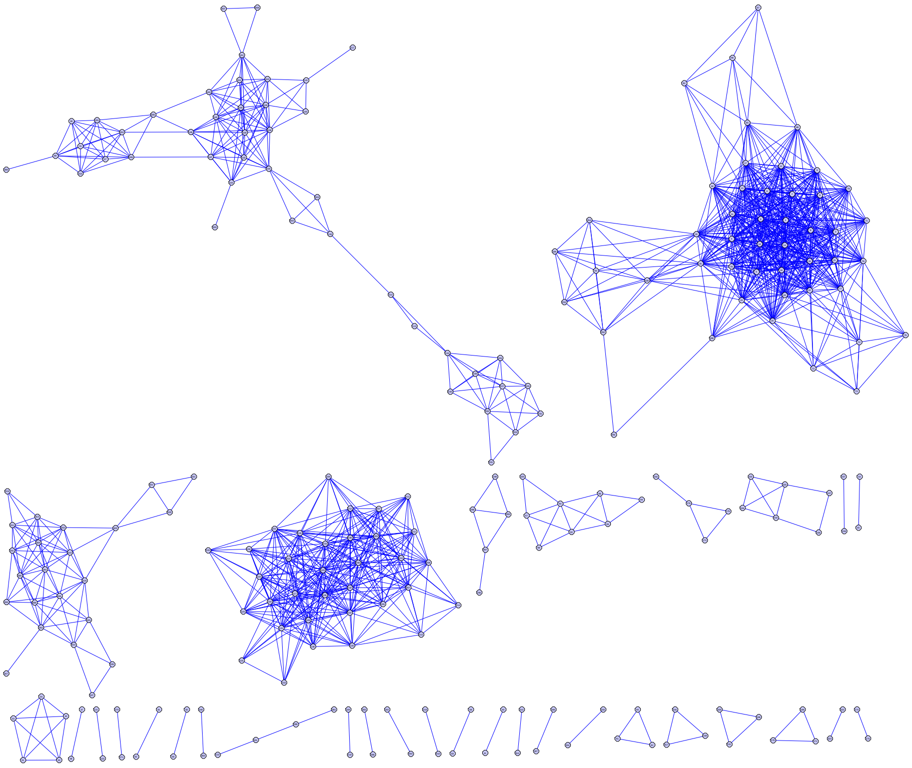

Figure 1.4: The Clique Intersection Graph obtained in this study. A graph is first constructed with 6,830 vertices (in a one to one corespondence with all the different genes in the original microarray). Edges in this graph links pairs of genes that have a correlation greater than 0.85 or smaller than -0.85. We then calculated its clique intersection graph, shown in this figure. This graph, in conjunction with the genetic signatures found with the $(\alpha,\beta)-k$-feature set method allows to identificate differential pathways associated with the disease. For instance, the K_5 at the bottom-left corner represents a set of cliques entirely composed of genes present in the Colon genetic signature shown in Fig. 1.2.c. The RPS16 gene (Ribosomal Protein S16) is present in all five cliques in the original graph, it is highly expresed in five cell lines, significantly less but still expressed in HCT-116 and underexpressed in HCT-15, matching recent reports [41]). Another gene common to all cliques is IL20RA, which encodes for receptor for interleukin 20 (IL20), a cytokine that may be involved in epidermal function. IL20RA is higly expressed in skin, upregualted in Psoriasis, and may have an important role in local mechanisms of mucosal and cutaneous immunity [39]. Our combinatorial methods allow a systematic investigation of what can be “master genes” as being key players in a variety of pathways implicated in the disease and allow for high-throughput bioinformatic analysis.