Hierarchical Clustering, Languages and Cancer

Pritha Mahata[1,2], Wagner Costa[1], Carlos Cotta[3], and Pablo Moscato[1,2]

[1] Newcastle Bioinformatics Initiative,
School of Electrical Engineering and Computer Science,
The University of Newcastle, Callaghan, NSW, 2308, Australia

[2] Australian Research Centre in Bioinformatics

[3] Dept. Lenguajes y Ciencias de la Computación, University of Málaga,
ETSI Informática, Campus de Teatinos, 29071 – Málaga, Spain

Contact email: Pablo.Moscato@newcastle.edu.au

Abstract. In this paper, we introduce a novel objective function for the hierarchical clustering of data from distance matrices, a very relevant task in Bioinformatics. To test the robustness of the method, we test it in two areas: (a) the problem of deriving a phylogeny of languages and (b) subtype cancer classification from microarray data. For comparison purposes, we also consider both the use of *ultrametric trees* (generated via a two-phase evolutionary approach that creates a large number of hypothesis trees, and then takes a consensus), and the best-known results from the literature.
We used a dataset of measured 'separation time' among 84 Indo-European languages. The hierarchy we produce agrees very well with existing data about these languages across a wide range of levels, and it helps to clarify and raise new hypothesis about the evolution of these languages.
Our method also generated a classification tree for the different cancers in the NCI60 microarray dataset (comprising gene expression data for 60 cancer cell lines). In this case, the method seems to support the current belief about the heterogeneous nature of the ovarian, breast and non-small-lung cancer, as opposed to the relative homogeneity of other types of cancer. However, our method reveals a close relationship of the melanoma and CNS cell-lines. This is in correspondence with the fact that metastatic melanoma first appears in central nervous system (CNS).

1 Introduction

A large number of articles in bioinformatics use single-objective unsupervised hierarchical clustering algorithms to identify "subtypes". Biologically-oriented journals, in general, have low requirements in terms of algorithmic reproducibility. The validation of the algorithms used in different problem scenarios is either poor or non-existing. Working on a dataset of measured separation times between 84 Indo-European languages we started to notice the deficiencies of a large number of hierarchical clustering schemes. The same problems occur when analyzing datasets derived from mitochondrial DNA distances among species [1].

In this paper, we pose the clustering problem as a graph optimization problem and propose a novel objective function which performs very well in diverse types of datasets. We start with a distance matrix for a set of objects and compute a weighted graph in which vertices represent objects and edges are weighted by the distance between the corresponding vertices. Then our 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. We denote this as *arithmetic-harmonic cut*. The recursive application of such cuts generates a tree-based classification of the data. While our primary concern is the classification of microarray data, we are also interested in testing the robustness of the approach, validating it in other domains. For this purpose, we show results for two different datasets: (a) a dataset for 84 Indo-European languages, and (b) a dataset for 60 cancerous cell-lines (NCI60). Next section will provide more details on the algorithmic methods we have used.

2 Hierarchical Clustering Methods Considered

In addition to the clustering solutions available in the literature for the datasets considered, we have used two unsupervised techniques for computing alternative solutions. The first one is based on arithmetic-harmonic cuts, and the second one relies on the utilization of ultrametric trees. These will be described below.

2.1 Arithmetic-Harmonic Cuts

The method of arithmetic-harmonic cuts approaches the construction of the hierarchy in a top-down fashion. To be precise, it can be described as a recursive process in which we solve a graph optimization problem at each step. Let $G(E, V, W)$ be an undirected, complete weighted graph with no self-loops and such that the weight of any edge is a positive integer number (i.e., $w(e) > 0$) representing distance or some measure of dissimilarity between a pair of objects. We first find a partition of the set V of vertices into $\{S, V \setminus S\}$, which 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 set of edges that link 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 in the same partition). Such a partition is defined by maximizing the following objective function

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

We have implemented an exact backtracking algorithm and also a memetic algorithm (similar to the work of Merz and Freisleben [2] for Graph Bipartitioning) as a meta-heuristic to calculate the best partitioning of the vertices for a given graph. The difference with respect to [2] is that we remove the constraint of equal partitioning of the graph in our memetic algorithm. Thus, the memetic algorithm uses (a) a differential greedy algorithm (similar to that in [3]) for initialization of a set of solutions for the problem, (b) a differential greedy crossover (a modification of the algorithm in [2]) for evolution of the population, and (c) a variable neighborhood local search (see [4]) to improve the newly generated solutions. Whenever the population stagnates, we keep the best solution and re-initialize the rest of solutions in the set. We use this memetic algorithm if the graph contains more than 25 vertices, and a backtracking enumeration algorithm otherwise. Notice that even though backtracking gives us an optimal solution, a memetic algorithm may not. However, in the considered datasets, the memetic algorithm consistently generated the same solution in all runs (thus it is presumably optimal). By applying this method (backtracking or memetic algorithm depending on the number of vertices) recursively, we have at each step a graph as input, and the two subgraphs induced by each of the sets of the vertex partition as output; stopping when we arrive to a graph with just one vertex, we generate a hierarchical clustering in a top-down fashion.

The rationale of the use of our objective function can be clear if we rearrange its terms. We can write

$$F = \frac{A_{out}}{H_{in}}(|E| - |E_{out}|)|E_{out}| \tag{2}$$

where A_{out} is the arithmetic mean of the weights that connect vertices of S with $V \setminus S$ (the cut); H_{in} is the harmonic mean of the weights of the edges not in the cut, and $|E_{out}|$ is the cardinality of the cut. Informally, maximizing F is equivalent to try to find a cut that discriminates well the two groups, normalized by the harmonic mean of the intra-cluster dissimilarity, and multiplied by a factor that is maximum when the two groups have a similar number of elements. Normalizing by the harmonic mean allows the denominator being more stable to the presence of outlier samples when associated to either V or $V \setminus S$. For this reason, we denote this partition as *arithmetic-harmonic cut.*

Notice that maximizing the first part of the objective function, i.e., $\sum_{e \in E_{out}} w(e)$ (the total weights of edges across the two sets) is the same as solving the Max-Cut problem for graph G, which is a NP-hard problem. However, it turns out that the hierarchy generated by partitions using Max-Cut does not corroborate the previous knowledge about the datasets. This is probably due to the fact that no importance is given in Max-Cut to the similarity of vertices within the sets. We also considered the objective function

$$F' = \sum_{e \in E_{out}} w(e) - \sum_{e \in E_{in}} w(e) \tag{3}$$

However, the resulting partition by maximizing F' turns out to be no better than the partition obtained from Max-Cut.

2.2 Ultrametric Trees

Ultrametric trees constitute a very amenable approach for fitting distance matrices to trees. In essence, an ultrametric tree T is a weighted tree in which the distance D_{ij} between any two leaves i and j (measured as the sum of the weights of the edges that have to be traversed to reach i from j inside T) verifies that $D_{ij} \leqslant \max\{D_{ik}, D_{jk}\},\ 1 \leqslant i, j, k \leqslant n$, where n is the number of leaves. This equation implies that given any internal node h in T, it holds that $D_{hi} = D_{hj}$ for any leaves i, j having h as ancestor.

The use of ultrametric trees has several advantages in hierarchical classification. First of all, edge weights are very easy to compute: given a distance matrix M containing dissimilarity values for a collection of objects, and a candidate tree T, the minimum weights such that $D_{ij} \geqslant M_{ij}$ and T is ultrametric can be computed in $O(n^2)$ [5]. Secondly, they adapt very well to dynamical processes evolving at a more or less constant rate. Finally, even if the latter is not the case, they provide a very good approximation to more relaxed criteria such as mere additivity, that would be much more computationally expensive to calculate. Notice also that finding the optimal topology T for a given distance matrix M under the ultrametric assumption is NP-hard [5].

Ultrametric trees have been computed using an evolutionary two-phase procedure: firstly, a collection of high quality tentative trees are generated; subsequently, a consensus method is used to summarize this collection into a single tree. Beginning with the former, the generation of high quality (i.e., minimum weight) ultrametric trees has been approached using an evolutionary algorithm based on the scatter search template. Starting from the solution provided by the complete-link agglomerative algorithm, an initial population of trees is produced by perturbation (internal exchanges of branches). Then, an evolutionary cycle is performed using tree-based path relinking for recombination [6], and internal rotations for local search (no mutation is used). Whenever the system stagnates, the population is restarted by keeping the best solution and generating new trees by exchanging branches among existing trees.

Once a collection of high quality trees has been found, the consensus method is used to amalgamate them. This is done using the TreeRank measure [7] as similarity metric among trees. This measure is based on counting the number of times we have to traverse an edge upwards or downwards in order to go from a certain leaf to another one. By computing how different these figures are for two trees, we obtain a dissimilarity value. The TreeRank measure is currently being used in TreeBASE[4] –one of the most widely used phylogenetic databases– for the purposes of handling queries for similar trees.

The consensus algorithm we have used is an evolutionary metaheuristic that evolves tentative trees following [8]. Given the collection of trees we want to summarize, the sum of dissimilarities to the tentative tree is used as the fitness function (to be minimized). Evolution is performed using the prune-delete-graft operator [9, 10] for recombination, no mutation, binary tournament selection, and elitist replacement. In our experiments, we have considered all different trees generated by the scatter search method in one hundred runs, and then running the consensus algorithm one hundred times on this collection. The best solution out of these latter 100 runs is kept as the final consensus tree.

[4] `http://www.treebase.org`

3 Classifying Indo-European Languages

Two major hypotheses exist about the origin of Indo-European languages: the *'Kurgan expansion'* and the *'Anatolian farming'* hypotheses. Based on archaeological evidences, the Kurgan theory [11] says that Kurgan horsemen (near current Russia) went to Europe and the Near East around 6000 years B.C. On the other hand, the Anatolian theory [12] claims that Indo-European languages expanded with the spread of agriculture from Anatolia (present Turkey) around 8000-9500 years B.C. Scientists have used genetic [13–15] and numerical [16, 17] methods to complete the linguistic phylogeny of the Indo-European languages. However, there are still some doubts about its exact topology.

The dataset we analyse is the distance-matrix for 84 Indo-European languages generated by Dyen *et al.* [18]. They used a list of basic vocabulary and estimated historical relationships (similarity) between two languages by computing the ratio of number of words (cognates) shared by them and the total number of words. Furthermore, one also considers the replacement rates of each word in the vocabulary. By considering the above ratio and replacement rates, they generated the so-called "separation time" between pairs of languages, which they provided as a distance-matrix of 84 languages. This dataset is the same that was used in [17] where the *neighbor-net analysis* method provided some hints on possible lateral information transfer while still provide an overall hierarchical relationships among the languages.

Gray and Atkinson used a Bayesian Markov chain Monte Carlo method to generate and analyze a large number of (10000) tentative trees [16]. Although their method is in remarkable agreement with the timing of the Anatolian hypothesis, the method also shows how much uncertainty still exists to validate some subtrees. The posterior probability of some subtrees can be as low as 0.36 in some cases, e.g., in dividing a sub-tree into Indo-Iranian and Albanian languages, and 0.4 in dividing the Greek and Armenian groups from the most of the languages (with the exception of the Tocharian and the Hittite, extinct languages introduced by authors to the original dataset by Dyer *et al.*) [18].

We have applied our method of arithmetic-harmonic cut to Dyen *et al.*'s dataset. The language tree we have obtained using this cut is shown in Fig. 1(a). As it can be seen, this method separates the relatively older languages (a Greco-Armenian-Albanian-Indo-Iranian group) from the relatively newer languages (a Celtic-Slavic-Germanic-Romanic group). Unlike our tree, the tree in [16] (see Fig. 1(b)) first separated the extinct Tocharians, Hittite, and the Greco-Armenian group languages from the rest. Their tree had the Albanian branch together with the Indo-Iranian group. We note that the resulting subtree of Indo-Iranian-Albanian languages are divided with only 0.36 bayesian posterior probability. This raises a concern, as it may be the case that Indo-Iranian-Albanian languages are older than what was claimed in [16] and they are more suited to be temporally closer to the Greco-Armenian languages. In fact, the work in [17] with neighbor-net analysis has produced a network with Albanians very close to the Greco-Armenian languages. Thus, our topology for the main divisions seem reasonable and is in closer relationship with the spread of farming from

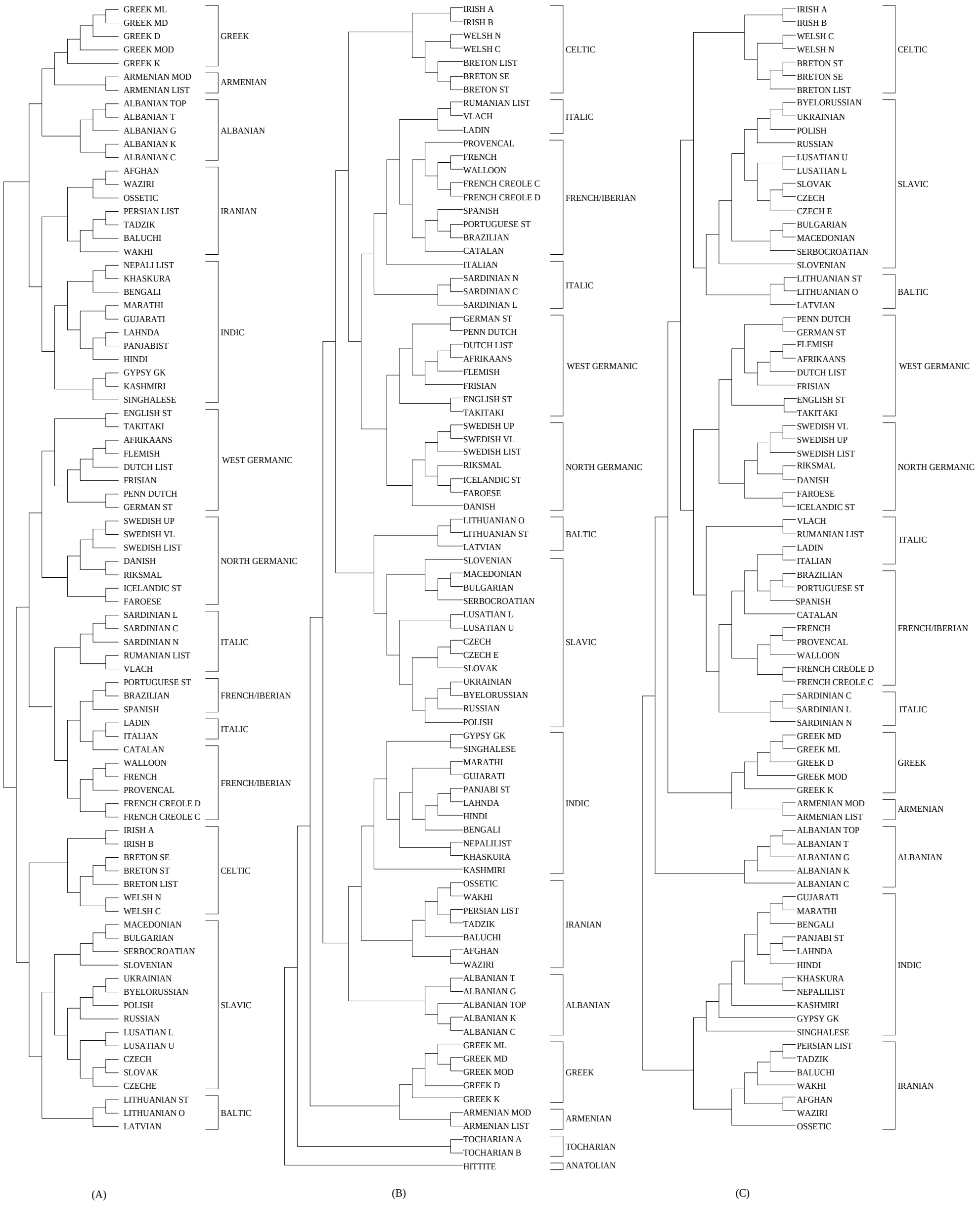

Fig. 1. Three proposed language-Trees: (a) tree using arithmetic-harmonic cuts, (b) Gray-Atkinson's tree [16], (c) consensus ultrametric tree.

the Middle East between 10,000 and 6,000 years ago, which also correlates well with the first principal component of the genetic variation of 95 polymorphisms [19], which solely accounts for 28 % of the total variation.

In addition to this fact, there are certain strange relations between languages at the leaves of Gray's tree. After checking the distance matrix, we find several cases in which our tree seems to produce a more reasonable branching than Gray and Atkinson's. First of all, the closest neighbor of Czech and Slovak languages are Lusatian languages. It is probably natural to have Czech, CzechE and Slovak placed in a subtree closer to Lusatian languages. In the trees generated from both arithmetic-harmonic cut (Fig. 1(a)) and the ultrametric trees (Fig. 1(c)), we see that these languages are placed next to each other. However in Fig. 1(b) generated by [16], Czech and Slovak are placed closer to Ukrainian, Byelorussian and Russian. Secondly, Catalan is a language evolved from Latin, with strong influences from French and Spanish. As a consequence of its Latin origin, Italian is the closest language of Catalan in the dataset. The position of Catalan with Italian and Ladin in our tree seems very natural, as hybridizations with French and Spanish occurred later (note that the bayesian posterior probability is 0.83 for its link with the Spanish-Portuguese group). See [16] for the details of probabilities in the Figure 1(b). Although Italian's closest language is Ladin, the latter was placed closer to RomanianList and Vlach with the posterior probability of 0.88. Also, notice the position of the Italian with 0.59 posterior probability. Finally, there are also small differences in the topology of small subtrees between our hierarchy and Gray's, namely, those regarding Dutchlist-Afrikaans-Flemish, Greek languages, Albanian languages and the position of Bengali in the Aryan languages among others. The differences seem to occur mainly where the posterior probability of one or several branchings is low.

An important difference is that in our classification the Celtic languages are considered closer to Baltic-Slavic languages. This goes against the current belief of Celtic's closeness to Romanic and Germanic languages. Note that in Gray and Atkinson's classification, the branchings of (Germanic,Romance) and (Celtic,(Germanic,Romance)) have low posterior probabilities (0.46 and 0.67, respectively). The minimum-weight ultrametric tree (see Fig. 1(c)) for this dataset also considers Celtic and Slavic languages to be the closest ones as groups. However, this tree disagrees with our tree in the primary branches. For instance, it first takes out Indo-Afghan languages as outliers, then considers Albanian and Greco-Armenian languages as outliers successively. In the tree obtained by the arithmetic-harmonic cut, all these outliers are grouped together. Notice that even at the successive branchings, the consensus ultrametric tree often produces a large number of outliers (see e.g., Indic and Iranian branches of Figure 1(c)).

4 A Molecular Classification of 60 Tumors

Validation of our methodology on the languages dataset has allowed us to have confidence in applying it in our primary problem domain, classification of cancer samples. In this section, we show how our partitioning algorithm finds subtypes of human cancers. We study a dataset from 60 tumor cell-lines used in National Cancer Institute's (NCI) screen for anti-cancer drugs. We use the gene expression of these cell lines given as a cDNA microarray with 6,830 genes for each cell-line.

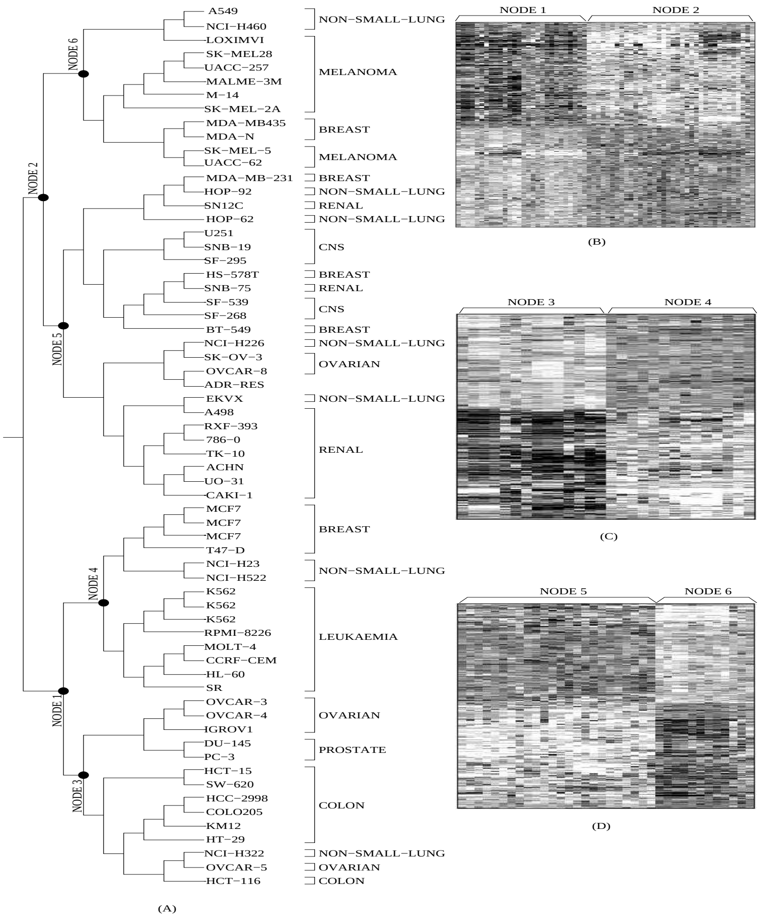

Fig. 2. (a) Classification of NCI60 dataset using arithmetic-harmonic cuts. Genetic signatures with (b) 1101 genes for the first partition into Node 1 and Node 2, (c) 695 genes for the second partition into Node 3 and Node 4, and (d) 696 genes for the third partition into Node 5 and Node 6.

The analysis of this dataset was done by Ross *et al.* in 2000 [20], where a result of a hierarchical clustering for this dataset was first discussed. Their result shows that the cell lines from same origin were grouped together in case of leukaemia, melanoma, colon, renal and ovarian cancers, with a few exceptions. However, cell-lines derived from non-small lung carcinoma and breast tumors were distributed in multiple places suggesting a heterogeneous nature.

Fig. 2(a) shows the result of applying arithmetic-harmonic cut on this dataset. In Fig. 2(b),(c) and (d), we show the genetic signatures (most differentially expressed genes in the two sides of the partition) of the first three partitions using the cut with 1,101, 696 and 695 genes respectively. In the genetic signatures (computed with the method described in [21] and [22]), each row corresponds to a gene and each column corresponds to a tumor sample.

We will now compare the hierarchy generated by arithmetic-harmonic cuts to Ross *et al.*'s classification tree and the consensus ultrametric tree, shown in Fig. 3. All clustering methods agree in the fact that some of the tumors, (e.g., leukaemia, renal, melanoma, central nervous system) are relatively homogeneous, i.e., most samples of these tumors are grouped together with a few exceptions. However, in Ross *et al.*'s solution and in the ultrametric tree, all melanoma samples except LOXIMVI are grouped with colon and leukaemia tumors (see the lower branches of both trees in Figure 3), whereas arithmetic-harmonic cut shows a marked difference in the first division. It groups *all* melanoma tumor samples (including LOXIMVI) with CNS (Central Nervous System), renal, ovarian and some of the breast and lung tumor samples (see Node 2 in Fig. 2(a)). Since CNS is the closest *group* to the melanoma samples in this figure, we may infer that this clustering supports the hypothesis that central nervous system (CNS) and melanoma may share important genetic pathways with similar expression. We note that CNS is a favorite site of metastasis in patients with advanced melanoma and that it is the first site of relapse in $15 - 20\%$ of melanoma cases [23, 24]. Also notice that CNS metastases from melanoma and renal tumors are also often misdiagnosed as primary brain tumors [25]. However, our literature survey did not reveal a close relationship of melanoma with colon or leukaemia, as compared to its relation with CNS.

The three methods, however, behave slightly differently in clustering non-small-lung, breast and ovarian tumors. First of all, all clustering methods applied on NCI60 group together ovarian tumor samples OVCAR-3, OVCAR04 and IGROV1. However, the positions of OVCAR-5, SK-OV-3 and OVCAR-8 differ in the outcomes of the different methods suggesting a possible heterogeneity of ovarian tumors. Also, it was suggested by Perou *et al.* [26] that there are 6 types of breast-cancers. All clustering methods more or less agree with the fact that the breast tumors (HS-578T, BT-549, MDA-MB-231, MCF7, T-47D, MDA-MB435, MDAN) are scattered in 4-5 places. In all methods, breast tumors HS-578T, BT-549 and MDA-MB231 are together with CNS/renal tumor samples; MCF7 and T-47D tumors are clustered with colon tumors; MDA-N and MDA-MB435 tumors are grouped with melanoma tumor samples. This is a definite indication that the breast cancer is a heterogeneous disease. Similarly, small-lung-cancer samples are distributed in all the three methods.

In the above comparison, we need to remember that ultrametric trees are largely used in cladistics, and assume that all species evolve at a constant rate (cf. Sect. 2). Such an assumption suits rather well the universe of languages (they evolve according to mutual contact between sub-populations, assimilating the words or phonemes from each other). However, unlike biological beings like bacteria or more complex life forms, the tumor cell-lines do not have a common ancestor. Tumors are defects in DNA, which cause malignant cell proliferation. Thus, the ultrametric approach may be susceptible to errors in the classification of cancer samples. Therefore, the position of melanoma samples in the tree

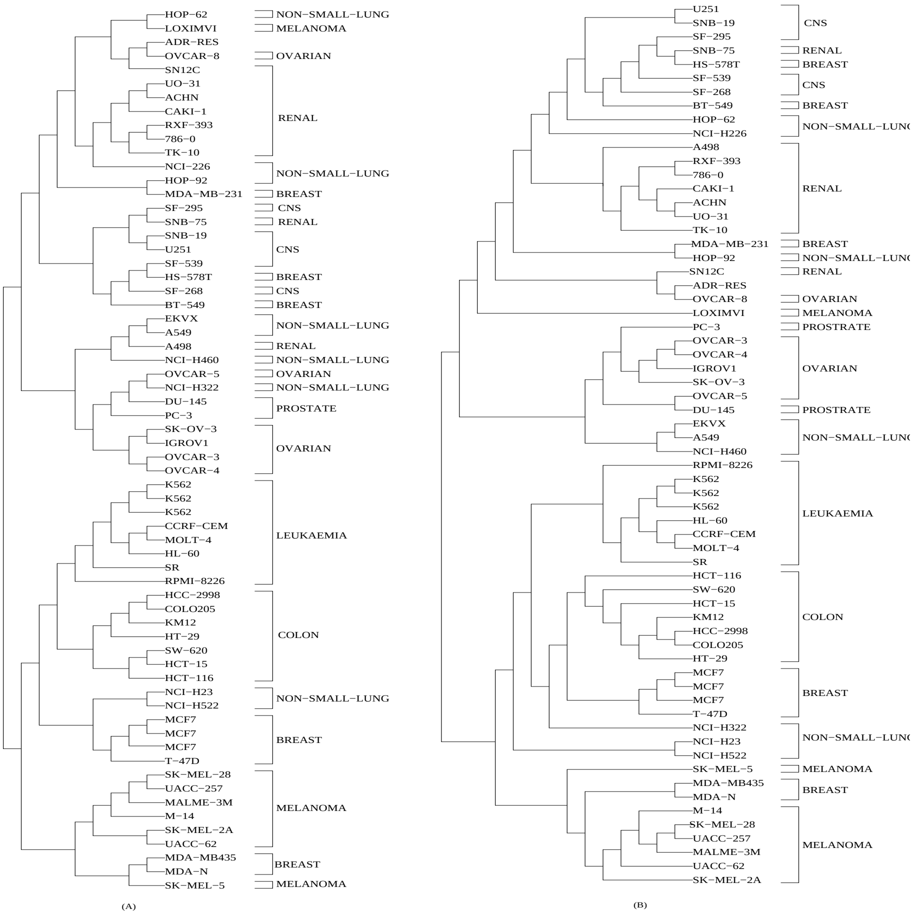

Fig. 3. Classification of NCI60 Dataset from (a) Ross *et al.* and (b) ultrametric tree.

produced by the arithmetic-harmonic cut should not be considered incorrect. Actually, the position of melanoma with CNS suggests an interesting quest, that of finding common activated genetic pathways, as it is known that the brain is a primary point of metastasis to an individual with melanoma condition [23, 24].

5 Conclusions

We proposed two new approaches for hierarchical clustering and showed the result of applying these methods on two very diverse datasets. The hierarchies we produce for both languages and cancer samples in this method agree very well with existing data about these datasets. It also raises some interesting questions. The arithmetic-harmonic cut seems to correlate well with the results of the first component of the genetic variation provided by Cavalli-Sforza and his coauthors [19]. It indicates a branching in two major groups, with an earlier group "moving" towards Europe (actually, the advanced farming hypothesis at work), later followed by another group moving in the same direction (evolving in Greek and Albanian and Armenian languages) while another group "moved" south-east and later differentiated in Iranian and Indic languages. It also suggest a commonality of Celtic, Baltic and Slavic (a hypothesis also raised in the past, and also supported by the consensus of the ultrametric trees). These differences, as well as small others, with the solution provided by Gray and Atkinson's seem to be in branchings where the bayesian posterior probability is low, and our methods agree where the posterior probability is high. The consensus of the ultrametric trees seem to suggest a single wave towards Europe, but a first branching in an Albanian group, followed by a second branching with the Greek and Armenian in one subgroup seems less plausible to us.

Overall, our results seem to indicate that it is important to use several hierarchical clustering algorithms and to analyze common subgroupings. In the case of tumor samples, it is indeed the case that this is the most relevant outcome as we do not have any guarantee that the samples "share" a common "ancestor". The question: *"Which tree is the best one ?"* might actually be highly irrelevant to the the real problem at hand, as it seems to be the consensus of these trees the most important outcome. Results on a number of other clustering algorithms on these datasets (which we were unable to show here for reasons of space), indicates that more research in robust algorithm methods needs to be done for molecular subtype classification in cancer and that validation of the methodology with different problem settings is highly beneficial to develop it.

References

1. Cotta, C., Moscato, P.: A memetic-aided approach to hierarchical clustering from distance matrices: Application to phylogeny and gene expression clustering. Biosystems **71** (2003) 75–97
2. Merz, P., Freisleben, B.: Fitness landscapes, memetic algorithms, and greedy operators for graph bipartitioning. Evolutionary Computation **8** (2000) 61–91
3. Battiti, R., Bertossi, A.: Differential greedy for the 0-1 equicut problem. In: Proc. of DIMACS Workshop on Network Design: Connectivity and Facilities. (1997)
4. Festa, P., Pardalos, P., Resende, M.G.C., Ribeiro, C.C.: Randomized heuristics for the MAX-CUT problem. Optimization Methods and Software **7** (2002) 1033–1058
5. Wu, B., Chao, K.M., Tang, C.: Approximation and exact algorithms for constructing minimum ultrametric trees from distance matrices. Journal of Combinatorial Optimization **3** (1999) 199–211
6. Cotta, C.: Scatter search with path relinking for phylogenetic inference. European Journal of Operational Research **169** (2006) 520–532
7. Wang, J., Shan, H., Shasha, D., Piel, W.: Treerank: A similarity measure for nearest neighbor searching in phylogenetic databases. In: Proceedings of the 15th International Conference on Scientific and Statistical Database Management, Cambridge MA, IEEE Press (2003) 171–180
8. Cotta, C.: On the application of evolutionary algorithms to the consensus tree problem. In Gottlieb, J., Raidl, G., eds.: Evolutionary Computation in Combina-

torial Optimization. Volume 3248 of Lecture Notes in Computer Science., Berlin, Springer-Verlag (2005) 58–67

9. Moilanen, A.: Searching for the most parsimonious trees with simulated evolution. Cladistics **15** (1999) 39–50
10. Cotta, C., Moscato, P.: Inferring phylogenetic trees using evolutionary algorithms. In Merelo, J., et al., eds.: Parallel Problem Solving From Nature VII. Volume 2439 of Lecture Notes in Computer Science. Springer-Verlag, Berlin (2002) 720–729
11. Mallory, J.P.: Search of the Indo-European languages. Archaelogy and Myth (1989)
12. Renfrew, C.: Time-depth in historical linguistics. The McDonald Institute for Archaeological Research (2000) 413–439
13. Richards, M.: Tracing european founder lineage in the near easter mtDNA pool. Am. K. Hum. Genet **67** (2000) 1251–1276
14. Semoni: The genetic legacy of Paleolithic Homo Sapiens in extant europeans: a Y chromosome perspective. Science **290** (2000) 1155–1159
15. Chikhi, L., Nichols, R., Barbujani, G., Beaumont, M.: Y genetic data support the Neolithic demic diffusion model. Prod. Natl. Acad., Sci. **67** (2002) 11008–11013
16. Gray, R.D., Atkinson, Q.D.: Language-tree divergence times support the anatolian theory of indo-european origin. Nature **426** (2003) 435–439
17. Bryant, D., Filimon, F., Gray, R.: Untangling our past: Languages, trees, splits and networks. In Mace, R., Holden, C., Shennan, S., eds.: The Evolution of Cultural Diversity: Phylogenetic Approaches. UCL Press (2005) 69–85
18. Dyen, I., Kruskal, J.B., Black, P.: An Indo-European classification: A lexicostatistical experiment. Transactions of the American Philosophical Society, New Ser. **82** (1992) 1–132
19. Cavalli-Sforza, L.: Genes, peoples, and languages. Proceedings of the National Academy of Sciences of the United States of America **94** (1997) 7719–7724
20. Ross, D.T., Scherf, U., Eisen, M., Perou, C., Rees, C., Spellman, P., Iyer, V., Jeffrey, S., Rijn, M., Waltham, M., Pergamenschikov, A., Lee, J.C., Lashkari, D., Shalon, D., Myers, T., Weinstein, J.N., Botstein, D., Brown, P.: Systematic variation in gene expression patterns in human cancer cell lines. Nature Genetics **24** (2000) 227–235
21. Cotta, C., Langston, M., Moscato, P.: Combinatorial and algorithmic issues for microarray data analysis. In: Handbook of Approximation Algorithms and Metaheuristics. Chapman and Hall (2005)
22. Hourani, M., Mendes, A., Berretta, R., Moscato, P.: A genetic signature for parkinsons disease using rodent brain gene expression. In Keith, J., ed.: Bioinformatics. Humana Press (2006)
23. Ferraresi, V., Ciccarese, M., Zeuli, M., Cognetti, F.: Central system as exclusive site disease in patients with melanoma: treatment and clinical outcome of two cases. Melanoma Res. 2005 **15** (2005) 467–469
24. Marchetti, D., Denkins, Y., Reiland, J., Greiter-Wilke, A., Galjour, J., Murry, B., Blust, J., Roy, M.: Brain-metastatic melanoma: a neurotrophic perspective. Pathology Oncology Research **9** (2003) 147–158
25. Buell, J., Gross, T., Alloway, R., Trofe, J., Woodle, E.: Central nervous system tumors in donors: Misdiagnosis carries a high morbidity and mortality. Transplantation Proceedings **37** (2005) 583–584
26. Perou, C.M., Jeffrey, S.S., Rijn, M., Rees, C.A., Eisen, M.B., Ross, D.T., Pergamenschikov, A., Williams, C.F., Zhu, S.X., Lee, J.C.F., Lashkari, D., Shalon, D., Brown, P.O., Botstein, D.: Distinctive gene expression patterns in human mammary epithelial cells and breast cancers. Genetics **96** (1999) 9212–9217