35th Annual International Conference of the IEEE EMBS
Osaka, Japan, 3 - 7 July, 2013

Dictionary learning improves subtyping of breast cancer aCGH data

Salvatore Masecchia[1] and Annalisa Barla[1] and Saverio Salzo[1] and Alessandro Verri[1]

Abstract— **The advent of Comparative Genomic Hybridization (CGH) data led to the development of new mathematical models and computational methods to automatically infer chromosomal alterations. In this work we tackle a standard clustering problem exploiting the good representation properties of a novel method based on dictionary learning. The identified dictionary atoms, which show co-occuring shared alterations among samples, can be easily interpreted by domain experts. We compare a state-of-the-art approach with an original method on a breast cancer dataset.**

I. Introduction

Multifactorial pathological conditions, as tumors, are often associated to structural and numerical chromosomal aberrations. The cell loses or varies its function when one or more sections of its DNA has an abnormal number of copies or copy number variations (CNVs). Array-based Comparative Genomic Hybridization (aCGH) is a modern whole-genome measuring technique that evaluates the occurrence of copy variants across the genome of samples (patients) versus references (controls) on the entire genome, extending the original CGH technology [1]. Modern high-resolution aCGH allows for the identification of numerical and structural aberrations or rearrangements.

A signal measured with an aCGH consists of a piecewise linear (and constant) component plus some noise. The typical statistical analysis on such data is the automatic detection of altered recurrent aberrations, that may indicate an oncogene or a tumor suppressor gene. The method should possibly exploit the intrinsic data structure to improve the downstream analysis. Indeed, recent advances in aCGH analysis are based on multi-sample regularization methods for a joint segmentation of many aCGH profiles with the simultaneous detection of shared change-points across samples [2].

The models proposed by [3] and [4] follow this stream, minimizing a functional based on total variation (TV) or fused lasso signal approximation. We developed CGHDL, a dictionary learning based method [5], which is an extension of the model proposed by [3], called FLLat. The main aim of such methods is to obtain a denoised version of the input data as well as a representative dictionary of atoms each containing a meaningful common pattern of genomic alterations. Our model provides a more biologically sound representation of aCGH data thanks to the combination of more complex penalties that explicitly exploit the structured nature of aCGH signals. Despite having a less simple model, we obtain atoms that possibly capture co-occurrences of CNVs across samples leading to results that are more easily interpretable by the domain experts. Moreover, our proposed model is able to deal with signals spanning the entire genome, whereas FLLat in [3] takes into account one chromosome at a time. In dictionary learning, the original signal is approximated by a linear weighted combination of the atoms, *i.e.*, the elements of the dictionary. In our model, we assume that each sample uses just some atoms enforcing sparsity on the coefficient matrix, which is used as the new representation of the input data.

We take into account a clustering problem on a breast cancer dataset comprising three different grades and, first, we compare the clustering properties of CGHDL, FLLat and the raw signals on two separated chromosomes. Then, we show CGHDL clustering properties on the entire genome and we also show how coefficients obtained by CGHDL on the entire genome can be considered a good representation of the data.

In the remainder of the paper we illustrate the model discussing the choice of each penalty, and then we present the experimental setting and the obtained results.

II. CGHDL: a new model for aCGH analysis

We are given a data matrix $Y \in \mathbb{R}^{L\times S}$. The goal is to seek a matrix $B \in \mathbb{R}^{L\times J}$ of J *simple* atoms which possibly give a complete representation of all samples, in the sense that $Y \approxeq \hat{Y} = B\Theta$ for a matrix of coefficients $\Theta \in \mathbb{R}^{J\times S}$.

In [3], the proposed model follows:

$$\min_{\Theta,B} \frac{1}{2}\|Y - B\Theta\|_F^2 + \lambda \sum_{j=1}^{J} \|B(:,j)\|_1 + \mu \sum_{j=1}^{J} TV(B(:,j)) \quad \text{s.t. } \|\Theta(j,:)\|_2^2 \le 1 \quad \forall j = 1,\dots,J. \tag{1}$$

The $\|\cdot\|_1$ penalization term forces each atom $B(:,j)$ to be sparse and the total variation term $TV(B(:,j)) = \sum_{l=1}^{L-1} |B(l+1,j) - B(l,j)|$, induces small variations in the atoms. The hard constraints on the coefficients $\Theta(j,:)$ are imposed for consistency and identifiability of the model. Indeed, multiplying a particular feature $B(:,j)$ by a constant, and dividing the corresponding coefficients by the same constant leaves the fit unchanged, but reduces the penalty.

Our model minimizes the following problem:

$$\min_{\Theta,B} \frac{1}{2}\|Y - B\Theta\|_F^2 + \lambda \sum_{j=1}^{J} \|B(:,j)\|_1^2 + \mu \sum_{j=1}^{J} TV_w(B(:,j)) + \tau \sum_{s=1}^{S} \|\Theta(:,s)\|_1^2$$
$$\text{s.t. } 0 \le \Theta(j,s) \le \theta_{\max}, \quad \forall j = 1,\dots,J \quad \forall s = 1,\dots,S.$$

[1]S. Masecchia, A. Barla, S. Salzo and A. Verri are with DIBRIS, University of Genova, 16146, Genova, Italy {salvatore.masecchia, annalisa.barla, saverio.salzo, alessandro.verri} at unige.it

978-1-4577-0216-7/13/$26.00 ©2013 IEEE

The reconstruction term is penalized with three penalties, based on the biological and structured nature of the data at hand.

The $\|\cdot\|_1^2$ penalization term on the matrix of atoms B forces each atom $B(:,j)$ to be sparse, and gives a structured sparsity along its columns.

The weighted total variation term $TV_w(B(:,j)) = \sum_{l=1}^{L-1} w_l |B(l+1,j) - B(l,j)|$, induces small variations in the atoms with properly chosen weights w_l. The weighting schema is introduced in order to relax at some points the constraint of *small jumps* on the atoms. Actually, we will use weights that are always 1 with some sparse exceptions, where w_l is 0, in correspondence of chromosomes boundaries where the constraint does not have a biological motivations. This allows to treat signals composed by several chromosomes as a whole, but still guaranteeing an independent analysis for each chromosome. This ensures the capability of identifying concomitant alterations occurring on different chromosomes.

The $\|\cdot\|_1^2$ penalization term on the matrix of coefficients Θ induces sparsity along the set of weights associated to each sample separately. This permits to regulate how much different atoms each sample can combine in order to reconstruct the original signal.

The coefficients are constrained to be bounded and positive. This reduces the complexity of the matrix of coefficients Θ and forces the matrix of atoms B to be more informative: *e.g.,* for deletions and amplifications occurring in different samples but on the same locus on the chromosome, different atoms may be selected.

A. Alternating prox minimization algorithm

To solve the minimization, we use a proximal alternating algorithm, as studied in its generality in [6]. We set Y, B and Θ as the matrices of data, atoms and coefficients respectively, and introduce the partial functions:

$$\begin{aligned}\varphi_B(\Theta) &= \frac{1}{2}\|Y - B\Theta\|_F^2 + \delta_{\Delta_{S\times J}}(\Theta) + \tau \sum_{s=1}^{S} \|\Theta(:,s)\|_1^2 \\ \psi_\Theta(B) &= \frac{1}{2}\|Y - B\Theta\|_F^2 + \lambda \sum_{j=1}^{J} \|B(:,j)\|_1^2 + \mu \sum_{j=1}^{J} TV_w(B(:,j)),\end{aligned} \tag{2}$$

where $\delta_{\Delta_{S\times J}}$ is the indicator function of the box set $\Delta_{S\times J} = [0,\theta_{\max}]^{S\times J}$. Then, the *alternating proximal algorithm* is as follows:

$$\left|\begin{array}{l}\Theta_{k+1} = \mathrm{prox}_{\eta_k \varphi_{B_k}}(\Theta_k), \quad \eta_k > 0, \\ B_{k+1} = \mathrm{prox}_{\zeta_k \psi_{\Theta_{k+1}}}(B_k), \quad \zeta_k > 0.\end{array}\right. \tag{3}$$

In (3), $\mathrm{prox}_{\eta\varphi_B}$ and $\mathrm{prox}_{\zeta\psi_\Theta}$ denote the proximity operators with respect to the partial functions (2). They can be computed approximately, by a duality based (inner) algorithm, with a given and controlled precision [7].

B. Parameter selection

The choice of the parameters (λ,μ,τ) is done according to the Bayesian information criterion (BIC) [8]. The BIC mitigates the problem of overfitting by introducing a penalty term for the complexity of the model. In our case the BIC is written as:

$$(SL)\cdot\log\left(\frac{\|Y - B\Theta\|_F^2}{SL}\right) + k(B)\log(SL)$$

and $k(B)$ is computed as the number of jumps in B and ultimately depends on the parameters (λ,μ,τ). Note that the reconstruction accuracy increases with J, but our aim is not achieving a perfect fit, rather is detecting the relevant alterations. In this context, the value of J may be chosen keeping in mind the compromise between model complexity (smaller J) and reconstruction accuracy (higher J).

III. Experiments

To better understand the underlying properties of the learned dictionary and coefficients by FLLat and CGHDL, we refer to [5] for preliminary results on synthetic generated data.

In this paper, we considered the aCGH dataset from [9], already used by [3] to test FLLat on real data. The dataset consisted of 44 samples of advanced primary breast cancer. Each signal measured the CNV of 6691 human genes. The samples were assigned to 3 classes according to tumor grading: 5 samples were assigned to grade 1, 21 to grade 2, 17 to grade 3 and 1 unassigned.

The aim of our experiments is to prove that CGHDL allows for a more informative representation of the data in terms of main shared patterns of alterations. In order to demonstrate this hypothesis we performed two different experiments: first we would like to demonstrate that CGHDL, even if more complex than FLLat, is able to extract useful information in a chromosome-by-chromosome analysis; then we performed an experiment considering all the chromosomes at the same time and showed how CGHDL can extract all the meaningful genomic alteration providing an overall informative result.

In the first experiment we compared CGHDL and FLLat focusing on chromosomes 8 (241 mapped genes) and 17 (382 mapped genes), identified by [9] as chromosomes with biologically relevant CNVs. Clustering was performed on Y^c, the original raw noisy data matrix restricted to the chromosome $c \in \{8,17\}$, on coefficients matrices Θ^c_{cghdl} and Θ^c_{fllat}, and on the denoised samples matrices $\hat{Y}^c_{cghdl}$ and $\hat{Y}^c_{fllat}$.

As explained in Section II, FLLat cannot analyze an aCGH signal along the entire genome due to the unweighted total variation included into its model, therefore, in the second experiment, we compared the results of CGHDL with a clustering procedure on the raw dataset (6691 probes). Clustering was performed on the original raw noisy data matrix Y, on the coefficients matrix Θ_{cghdl} and the denoised samples matrix $\hat{Y}_{cghdl}$ calculated by CGHDL.

For clustering, we adopted a hierarchical agglomerative algorithm, using the the *city block* or *manhattan* distance between points $d(a,b) = \sum_i |a_i - b_i|$ and the *single linkage* criterion $d(A,B) = min\{d(a,b) : a \in A, b \in B\}$ [10]. The cluster A is linked with the cluster B if the distance $d(A,B)$ is the minimum with respect to all the other clusters B'.

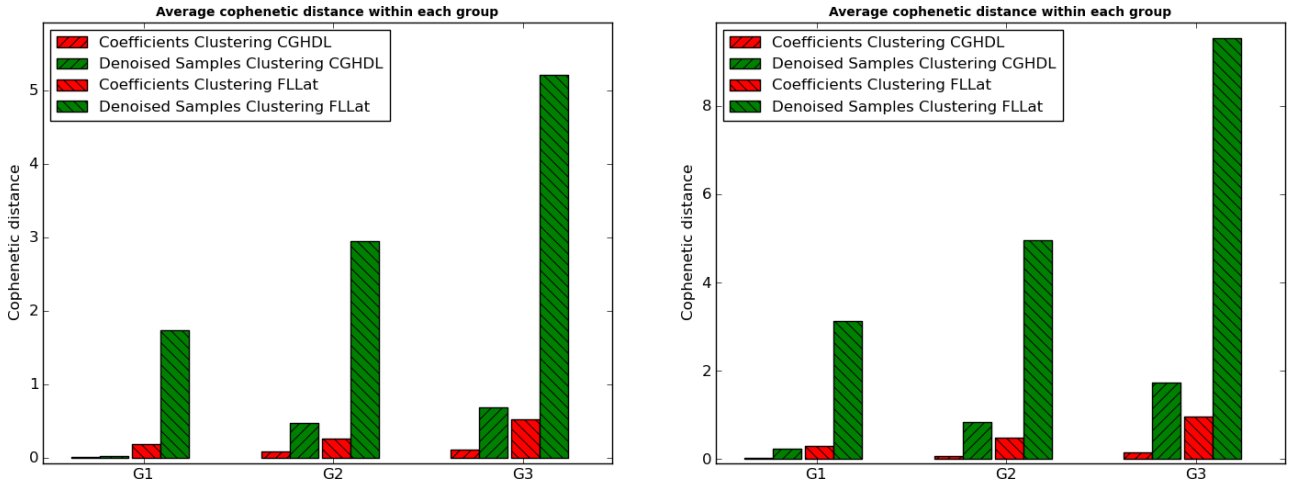

Fig. 1. Average cophenetic distances for the groups G1, G2 and G3 on chromosome 17 (left) and chromosome 8 (right). CGHDL always has better clustering results (see also Tables I and II). Moreover, is also interesting to note that clustering the denoised samples by CGHDL and FLLat, the former has better results, suggesting also an higher quality of the dictionary atoms used to reconstruct the samples.

The *manhattan* distance allows us to calculate a point-wise difference both for the coefficients vectors and the raw/denoised aCGH signals.

Moreover, to evaluate the coherence of the obtained dendrogram with respect to the groups G1, G2 and G3, we measured the *cophenetic distance* among the samples within each group [11]. For each pair of observations (a,b), the cophenetic distance is the distance between the two clusters that were merged to assign the two points in a single new cluster. The average of the cophenetic distances within each clinical group provides an objective measure of how the resulting dendrogram "describes" the differences between observations, using the clinical grades as ground truth.

Note that, by design, the values contained into the coefficients matrix produced by FLLat and CGHDL could have different range of values (in CGHDL the values are positive and bounded). In order to calculate comparable distance metrics, before clustering and cophenetic distances evaluation, each estimated coefficients matrix Θ was normalized by is maximum absolute value. The same preprocessing was also applied on the original aCGH signals and on the estimated ones $\hat{Y} = B\Theta$.

IV. Results

Both FLLat and CGHDL choose the optimal parameters over a grid using a BIC-based searching algorithm. In particular, for FLLat the grid was defined by some heuristics implemented into the given R package. The parameter θ_{max} in CGHDL was set to 1.0. This choice forces the algorithm to find atoms with signal amplitude comparable with the original data.

A. Analysis restricted to chromosomes 17 and 8

In Figure 1 (left) we show the means of the cophenetic distances calculated for each group of samples (the unannotated one was not considered) restricted to the chromosome 17. In this experiment, following [3], we fixed $J = 5$ and initialized B with the first 5 principal components of the matrix Y.

	G1	G2	G3
Θ^{17}_{cghdl}	$\mathbf{0.008 \pm 0.004}$	$\mathbf{0.079 \pm 0.112}$	$\mathbf{0.111 \pm 0.124}$
$\hat{Y}^{17}_{cghdl}$	0.022 ± 0.019	0.476 ± 0.720	0.687 ± 0.795
Θ^{17}_{fllat}	0.178 ± 0.044	0.265 ± 0.173	0.517 ± 0.446
$\hat{Y}^{17}_{fllat}$	1.737 ± 0.484	2.945 ± 2.074	5.212 ± 3.851
Y^{17}	19.284 ± 2.374	19.589 ± 3.961	23.941 ± 5.870

TABLE I
AVERAGE COPHENETIC DISTANCES AFTER CLUSTERING FOR THE ANALYSIS RESTRICTED TO CHROMOSOME 17

We searched, for CGHDL, the best triple of parameters in $\mu \in \{0.01, 0.1, 1.0, 10, 100\}$, $\lambda \in \{0.01, 0.1, 1.0, 10, 100\}$ and $\tau \in \{0.1, 1.0, 10\}$. It is clear that the clustering on the coefficients matrix produced by CGHDL places the samples belonging to homogeneous clinical groups (G1, G2 and G3) closer in the dendrogram. Moreover, also the denoised data matrix $\hat{Y}^{17}_{cghdl}$ shows better discriminative performances with respect to $\hat{Y}^{17}_{fllat}$. This may be due to the capability of our model to better detect the main altered patterns in the signals, despite a possibly higher reconstruction error [5]. Such property ultimately induces a more effective clustering. In Table I we report a summary of the averaged cophenetic distances, also including the clustering on raw signals.

The analysis on chromosome 8 gives similar results. Following [3], we fixed $J = 6$, initialized B with the first 6 principal components of the matrix Y and searched, for CGHDL, the best parameters in $\mu \in \{0.01, 0.1, 1.0, 10, 100\}$, $\lambda \in \{0.01, 0.1, 1.0, 10, 100\}$ and $\tau \in \{0.1, 1.0, 10\}$. Figure 1 (right) shows the means of the cophenetic distances calculated for each group of samples, and Table II shows the corresponding averaged cophenetic distances.

B. Whole genome analysis

We ran the experiments with three different $J \in \{10, 18, 24\}$ which correspond to the number of principal components of Y able to explain respectively the 50%, 70% and 80% of the variance. Than we searched the best param-

	G1	G2	G3
Θ^8_{cghdl}	$\mathbf{0.016 \pm 0.007}$	$\mathbf{0.054 \pm 0.024}$	$\mathbf{0.147 \pm 0.142}$
$\hat{Y}^8_{cghdl}$	0.222 ± 0.095	0.842 ± 0.410	1.720 ± 1.135
Θ^8_{fllat}	0.301 ± 0.095	0.469 ± 0.236	0.951 ± 0.657
$\hat{Y}^8_{fllat}$	3.135 ± 1.090	4.962 ± 2.605	9.547 ± 6.638
Y^8	12.363 ± 1.165	15.484 ± 4.124	20.150 ± 6.200

TABLE II

AVERAGE COPHENETIC DISTANCES AFTER CLUSTERING FOR THE ANALYSIS RESTRICTED TO CHROMOSOME 8

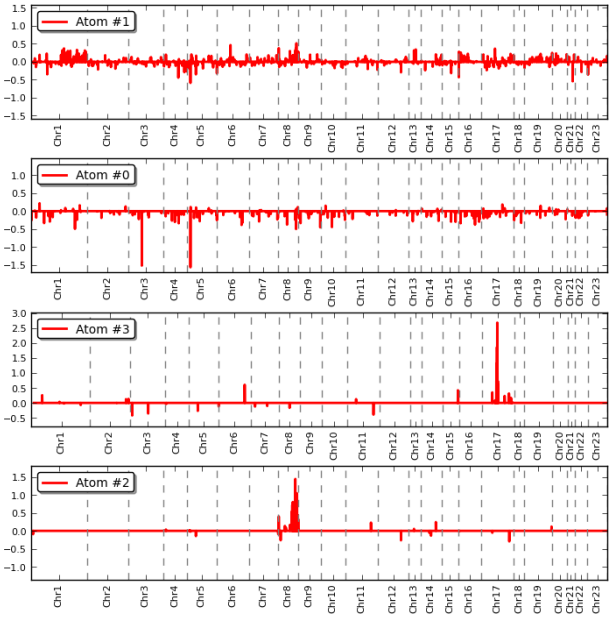

Fig. 2. Profiles of the first 4 more used atoms for sample reconstruction (sum of the row of Θ) extracted by CGHDL on all chromosomes. The atom #1 maps a general pattern of alterations, and it is responsible of a high proportion of signal reconstruction. CGHDL found the alterations on chromosomes 8 and 17, and also detected co-occurring alterations on chromosomes 3 and 5. Vertical lines indicate chromosomes boundaries.

eters $\mu \in \{0.01, 0.1\}$, $\lambda \in \{0.01, 0.1\}$ and $\tau \in \{0.01, 0.1\}$. Here, we present the results obtained with $J = 10$: the resulting atoms (see Figure 2) describe co-occurrent alterations along different chromosomes but are still fairly simple for a visual interpretation by the domain experts. For different Js we did not note relevant differences in terms of fit and clustering.

It is important to note that the four more used atoms of the dictionary extracted by CGHDL detect the main genomic alterations on chromosomes 8 and 17 as well as a co-occurrence of deletions on chromosome 3 and 5. In [9] all these alterations were already indicated as very common but the relation between chromosomes 3 and 5 was not indicated as co-occurrence and needs further biological validation.

V. CONCLUSIONS

In this paper we presented a novel method for aCGH data analysis and compared our result with a state-of-the-art method. We demonstrate the good properties of CGHDL for representing aCGH signals (coefficients) and extracting relevant information (atoms). The clustering results were

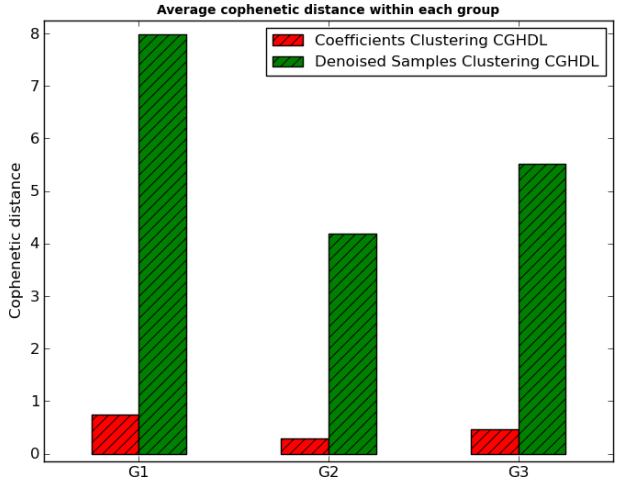

Fig. 3. Average cophenetic distances for the groups G1, G2 and G3 on all chromosomes and $J = 10$. See also Table III

	G1	G2	G3
Θ_{cghdl}	$\mathbf{0.738 \pm 0.541}$	$\mathbf{0.290 \pm 0.213}$	$\mathbf{0.463 \pm 0.406}$
$\hat{Y}_{cghdl}$	7.988 ± 4.663	4.191 ± 2.795	5.512 ± 3.632
Y	305.76 ± 39.85	290.26 ± 38.04	302.86 ± 34.76

TABLE III

AVERAGE COPHENETIC DISTANCES AFTER CLUSTERING FOR THE ANALYSIS EXTENDED TO ALL CHROMOSOMES WITH $J = 10$

validated using clinical grading as ground truth. We expect to apply the method on higher resolution aCGH data and possibly validate the ability to extract meaningful CNVs to give the domain experts an effective method to understand underlying biological processes.

REFERENCES

[1] A. Kallioniemi, O.-P. Kallioniemi, D. Sudar, D. Rutovitz, J. W. Gray, F. Waldman, and D. Pinkel, "Comparative genomic hybridization for molecular cytogenetic analysis of solid tumors," *Science (New York, N.Y.)*, vol. 258, no. 5083, pp. 818–21, Oct. 1992.

[2] H. Wang and J. Hu, "Identification of Differential Aberrations in Multiple-Sample Array CGH Studies," *Biometrics*, vol. 67, no. 2, pp. 353–62, Jul. 2010.

[3] G. Nowak, T. Hastie, J. R. Pollack, and R. Tibshirani, "A fused lasso latent feature model for analyzing multi-sample aCGH data," *Biostatistics*, Jun. 2011.

[4] J.-P. Vert and K. Bleakley, "Fast detection of multiple change-points shared by many signals using group LARS," *Advances in Neural Information Processing Systems 23*, vol. 1, pp. 1–9, 2010.

[5] S. Masecchia, S. Salzo, A. Barla, and A. Verri, "A dictionary learning based method for acgh segmentation," in *Proc. of ESANN*, 2013.

[6] H. Attouch, J. Bolte, P. Redont, and A. Soubeyran, "Proximal alternating minimization and projection methods for nonconvex problems: an approach based on the kurdyka-lojasiewicz inequality," *Math. Oper. Res.*, vol. 35, no. 2, pp. 438–457, 2010.

[7] S. Villa, S. Salzo, L. Baldassarre, and A. Verri, "Accelerated and inexact forward-backward algorithms," *Optimization Online*, 2012.

[8] G. Schwartz, "Estimating the dimension of a model," *Annals of Statistics*, 1978.

[9] J. Pollack, T. Sørlie, C. Perou, C. Rees, S. Jeffrey, P. Lonning, R. Tibshirani, D. Botstein, A. Børresen-Dale, and P. Brown, "Microarray analysis reveals a major direct role of dna copy number alteration in the transcriptional program of human breast tumors," *Proceedings of the National Academy of Sciences*, vol. 99, no. 20, 2002.

[10] R. Sibson, "Slink: an optimally efficient algorithm for the single-link cluster method," *The Computer Journal*, vol. 16, no. 1, 1973.

[11] R. Sokal and F. Rohlf, "The comparison of dendrograms by objective methods," *Taxon*, pp. 33–40, 1962.