Open problems in a recent algebraic method in phylogenetics*

Óscar Rivero Salgado and Pol Torrent i Soler
Projectes d'Enginyeria Física 2
Universitat Politècnica de Catalunya
(Dated: Spring of 2015)

In this project we aim to improve the Erik+2 method for obtaining the right distributions at the leaves of a phylogenetic tree by addressing the problems that are due to the lack of enough experimental data when dealing with a high number of species. We introduce a new procedure based on successive applications of the Erik+2 method to take into account the most filled rows and columns of the observed data matrix and on balancing the scores obtained from both rows and columns. We also propose normalizations to compare the scores based on the dimensions of the data matrix.

Keywords: Phylogenetics, Algebraic Geometry, Flattening matrix, Erik+2 Method

I. THEORETICAL INTRODUCTION

The evolution of species is usually modelled in a phylogenetic tree $\mathcal{T}$. The leaves of the tree represent current species and the root the common ancestor. The aim of phylogenetics is to determine the phylogenetic tree of a set of species from the DNA sequences of current species. Due to its structure, we can deal with DNA sequences as if they were a sequence of nucleotids (A, C, G, T). For this reason, we need a statistic model for the substitutions of nucleotids to face our problem. We will work under the following assumptions:

- The trees are binary (which means that two branches come out of the root, if it exists, and that they are divided into another two branches in each node).
- The processes in each branch do only depend on the common father node.
- Mutations of the DNA chain occur randomly.
- Each position of the DNA evolves independently and under the same mutation probabilities. This means it is enough to model one position of the chain.

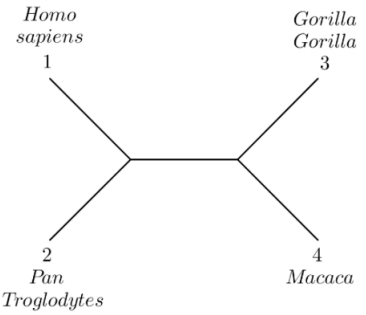

FIG. 1. A example of an unrooted 4-leaf phylogenetic tree

Following these assumpions we can think the nucleotid mutation process as a Markov process by assigning to each edge e a transition matrix

$$S_e = \begin{pmatrix} \mathrm{P}(\mathtt{A}\,|\,\mathtt{A},e) & \mathrm{P}(\mathtt{C}\,|\,\mathtt{A},e) & \mathrm{P}(\mathtt{G}\,|\,\mathtt{A},e) & \mathrm{P}(\mathtt{T}\,|\,\mathtt{A},e) \\ \mathrm{P}(\mathtt{A}\,|\,\mathtt{C},e) & \mathrm{P}(\mathtt{C}\,|\,\mathtt{C},e) & \mathrm{P}(\mathtt{G}\,|\,\mathtt{C},e) & \mathrm{P}(\mathtt{T}\,|\,\mathtt{C},e) \\ \mathrm{P}(\mathtt{A}\,|\,\mathtt{G},e) & \mathrm{P}(\mathtt{C}\,|\,\mathtt{G},e) & \mathrm{P}(\mathtt{G}\,|\,\mathtt{G},e) & \mathrm{P}(\mathtt{T}\,|\,\mathtt{G},e) \\ \mathrm{P}(\mathtt{A}\,|\,\mathtt{T},e) & \mathrm{P}(\mathtt{C}\,|\,\mathtt{T},e) & \mathrm{P}(\mathtt{G}\,|\,\mathtt{T},e) & \mathrm{P}(\mathtt{T}\,|\,\mathtt{T},e) \end{pmatrix}$$

where $\mathrm{P}(\mathtt{I}\,|\,\mathtt{J},e)$ is the probability of the nucleotid in the father node J becoming I after the edge e. These entries are unknown and along with the distribution in the root $\pi = (\pi_{\mathtt{A}}, \pi_{\mathtt{C}}, \pi_{\mathtt{G}}, \pi_{\mathtt{T}})$ are the parameters of our model. By imposing conditions on the matrix S_e one obtains different models.

We define now the random variables X_i as the state of the leaf i for $i \in \{1, \dots, n\}$ so that X_i takes value in $\{\mathtt{A}, \mathtt{C}, \mathtt{G}, \mathtt{T}\} = \mathcal{K}$, where n is the number of leaves of the tree. Now let $p_{x_1 x_2 \cdots x_n} = \mathrm{P}(X_1 = x_1, \dots, X_n = x_n)$ be the joint distribution at the leaves of the tree. Those probabilities can be calculated using only the entries of the transition matrices.

We are now ready to state the main definition and the main theorem we will need to understand Erik+2 method. Let $A|B$ be a partition of the leaves (that is, if $L(\mathcal{T})$ is the set of leaves of the rooted tree $\mathcal{T}$ then $L(\mathcal{T}) = A \cup B$ and $A \cap B = \varnothing$). Then we define the *flattening matrix* $\mathrm{flat}_{A|B}$ associated to the partition $A|B$ as the $4^{|A|} \times 4^{|B|}$ matrix

$$\mathrm{flat}_{A|B} = \begin{pmatrix} p_{\mathtt{AA}\cdots\mathtt{AA}} & p_{\mathtt{AA}\cdots\mathtt{AC}} & p_{\mathtt{AA}\cdots\mathtt{AG}} & \cdots & p_{\mathtt{AA}\cdots\mathtt{TT}} \\ p_{\mathtt{AC}\cdots\mathtt{AA}} & p_{\mathtt{AC}\cdots\mathtt{AC}} & p_{\mathtt{AC}\cdots\mathtt{AG}} & \cdots & p_{\mathtt{AC}\cdots\mathtt{TT}} \\ p_{\mathtt{AG}\cdots\mathtt{AA}} & p_{\mathtt{AG}\cdots\mathtt{AC}} & p_{\mathtt{AG}\cdots\mathtt{AG}} & \cdots & p_{\mathtt{AG}\cdots\mathtt{TT}} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{\mathtt{TT}\cdots\mathtt{AA}} & p_{\mathtt{TT}\cdots\mathtt{AC}} & p_{\mathtt{TT}\cdots\mathtt{AG}} & \cdots & p_{\mathtt{TT}\cdots\mathtt{TT}} \end{pmatrix}$$

That is, each column of the flattening matrix corresponds to a state of the leaves in B and each row to state of the leaves in A. We will call such a partition

* Advisor: Marta Casanellas Rius. The authors would like to thank her for providing us all the material, support and guide we needed to carry out the project.

proper if we can remove an edge such that all the leaves in A are in the same connected component and all the nodes leaves in B are in the other one. For instance, in the previous example $12|34$ is a proper partition, while $13|24$ is not. Now we are ready to state the following

THEOREM. Let $A|B$ be a partition of the set of leaves of the tree $\mathcal{T}$. If that partition is proper, then

$$\text{rank flat}_{A|B} \leq 4$$

whereas if it is not a proper partition then

$$\text{rank flat}_{A|B} > 4.$$

For the case with $n = 4$ species at the leaves if the parameters are "general enough" one can show that the rank of the flattening matrix for partitions which are not proper is maximum (i.e. 16), but since we will be dealing with cases with $n = 12$ we can not assume this as true.

II. THE ERIK+2 METHOD

A. Fundamentals and the algorithm

We start off with a set of nucleotid sequences (one for each leaf in our tree, as they are the observed DNA chains of current species) which we will assume that have nonempty entries and have all the same length. We will call an *alignment* such a set, for instance, consider the following 3-species alignment:

Specie α `AAAGAGTTCCA`
Specie β `AGACAGATGCA`
Specie γ `AGGGGGAAAGA`

From this experimental data we can calculate the relative frequencies $\tilde{p}_{x_1x_2\cdots x_n}$, which we will use as estimators for the true probabilities $p_{x_1x_2\cdots x_n}$ (in fact it can be shown that those are the maximum likelihood estimators for the true probabilities). Given a partition of the leaves $A|B$, we can build the estimated flattening matrix $\widetilde{\text{flatt}}_{A|B}$ just like we did in the previous chapter but this time using the relative frequencies instead of the true probabilities. We aim to determine the right topology of the tree (i.e. to determine which specie is at each leaf) by studying which partitions of the leaves are proper according to the experimental data and which not.

By the theorem we stated above, if those matrix was exactly the flattening matrix we should be able to distinguish between proper partitions and the other ones because proper ones would have exactly rank 4 and the other ones would not. This could be done easily by checking whether all 5×5 minors vanish or not, but since we only have the estimated matrices we have to develop a method to decide which one is "closer" to rank 4 matrices. Since we want to know how close is a matrix to the set of rank 4 matrices $\mathcal{V}$ we should define a distance. In this case we will work with the distance induced by the Frobenius norm (which is the square root of the quadratic sum of all elements in the matrix). It can be shown that the distance in the Frobenius norm of a matrix to $\mathcal{V}$ can be computed as the quadratic sum of the singular values of the matrix other than the first four. This is, if M is an $m \times n$ matrix with values in $\mathbb{C}$, its singular value decomposition is a factorization of the form

$$M = U\Sigma V^*$$

where U is a $m \times m$ unitary matrix, Σ is a $m \times n$ diagonal matrix with entries $\sigma_i \geq 0$ and V^* is the Hermitian transpose of a $n \times n$ unitary matrix. Then we stated that the distance of an arbitrary matrix to rank 4 matrices in the Frobenius norm is simply

$$d(M, \mathcal{V}) = \sum_{i=5}^{\min\{m,n\}} \sigma_i^2.$$

The Erik method uses this fact to give a score to each flattening matrix. Indeed, it works as follows: given an alignment and a partition $A|B$, it computes the estimated flattening matrix and then it obtains the singular value decomposition of the matrix and computes the distance $d(\widetilde{\text{flatt}}_{A|B}, \mathcal{V})$ which is the score assigned to the partition. Hence the partition which is estimated to be proper is the one that has the lower score.

The Erik+2 method slightly modifies the previous procedure by noticing that since there are mutations that are more probable than other ones (e.g. similar species have similar nucleotid sentences) then there are rows and columns which are more filled than other and this fact can lead to wrong predictions. The solution given by the Erik+2 method is to normalize first rows and then columns so as each one sums up to 1. Scores obtained after normalizing by both rows and columns are taken into account to compute the final score.

B. Some issues of the method

One has to take into account that if we are dealing with a case with $n = 4$ then the flattening matrices for 2×2 partitions will have dimension 16×16. But in our case we used the algorithm to treat cases with 12 species, which leads to flattening matrices with dimensions $4^2 \times 4^{10}$ for 2×10 (actually the dimensions of the matrix we were dealing with computationally were about 16×60000 since we were only taking into account nonempty rows and columns) and $4^5 \times 4^7$ for 5×7 partitions. This explains why alignments with size 100000 work fine with 4 species but often are not enough to fill bigger flattening matrices so as to give a closer approach to the theoretical situation. It is

important to notice that this important problem is only due to the lack of enough data, since if we were working with theoretical flattening matrices and not with the estimated ones then the Erik method would assign a perfect 0 score to the proper partition and a strictly greater one to not proper ones.

Another issue we tried to address is to make scores which come from partitions with different size comparable. With the existing method it was not possible since the diagonal matrices Σ dimensions depend on the size of the partition and hence when dealing with partitions where both sets have similar cardinal we have more singular values to sum than when one set has significantly more elements than the other.

III. OUR PROPOSED MODIFICATIONS

In this section we describe some of the most successful modifications out of the ones we tried. We start off with the observation that for the 2×10 sized partitions the flattening matrices have lots of columns which contain a single element due to the lack of data and that this fact can easily alter the rank of the matrix. Since the theoretical model stated that we should be dealing with matrices of rank approximately 4 we conjectured that there should be an important amount of data in a few rows and columns.

First of all we looked at how data should be distributed if the matrix was completely random to compare it to the actual flattening matrices. We obtained (assuming alignments of size 10^5 as the ones we had) for instance that for the 2×10 partition there would be on average 95380 nonempty columns where 90869 of them have only one entry. The actual matrices have about 60000 columns, 40000 of them having a single entry, hence dispersion is lower than in the random model but not much lower. For 5×7 partitions, we observed that random matrices have entries in almost all columns (we computed an average of 16347 nonempty columns out out $4^7 = 16384$ possible and we expected that just 98 columns had one entry). In this case we observed that on average we had 9000 nonempty columns so dispersion was also lower than in the random case.

We also looked with detail to some cases and found out the following patterns for flattening matrices coming from a proper partition. They usually:

- Have a lower amount of nonempty rows and columns.
- Have less rows and columns with only 1 entry.
- Have more entries in the most populated rows.

This led us to think that a good idea was to reorder rows and columns according to their number number of entries in order to have the most populated (and hence most significant) rows and columns in the first place. Then we consider the sub-matrices obtained by taking the m rows and the first k columns. We apply the Erik+2 method to those sub-matrices and then we increase the value the k a certain amount (we usually increased by 1000 because we saw heuristically it gave a nice balance between time of computation and data considered), compute the score again and so on and finally we add up all the scores. In order to compare the scores between partitions of different size, it is a good idea to divide the score for the number of total SVDs done, but when dealing with partitions of the same size this does not help since usually the wrong matrices have a higher number of columns and so in those cases we do a higher amount of SVDs, helping to increase the score for not proper partitions.

We also considered to do an analogous procedure for the rows, i.e. considering sub-matrices of size $k_1 \times k_2$ and then increase both k_1 and k_2, but since we are usually dealing with matrices which have $m \ll n$ we did not see a significant improvement of the results. Due to this fact we also need to multiply by m the score obtained by normalizing the columns and by n the score obtained by normalizing the rows in order to have the same order of magnitude for them both.

Since we are adding up scores of matrices with different dimensions the next step is to give estimates for the value of those scores so we can normalize. We will try a simple model that gives us a lower bound for the Frobenius norm of the matrix. Assume that each row has e/m entries which are equal to 1, where e is the total number of entries of the matrix. After normalizing by rows, those 1 entries becomes m/e. Hence the Frobenius norm of the matrix is

$$\|M\|_2 = \sqrt{\left(\frac{m}{e}\right)^2 \cdot \frac{e}{m} \cdot m} = \frac{m}{\sqrt{e}}$$

and since we are multiplying this score by the number of columns n then then our bound becomes $mn/\sqrt{e}$. It is easy to see that an analogous argument gives the same result when we normalize by columns and then multiply by m. An upper bound for the Frobenius norm is obtained easily since the maximum is attained when all the data of the row is in a single position, so that $\|M\|_2 = m$ and after multiplying by n we obtain mn. Again the argument is symmetric for both rows and columns.

Since dispersion is high we assume that our data will be closer to the lower bound model and hence the score we assign to a $m \times n$ sub-matrix (the overall score is obtained after adding up all the scores given to sub-matrices).

$$\text{score}(M) = \frac{n \cdot \text{rowscore} + m \cdot \text{colscore}}{mn/\sqrt{e}}$$

where rowscore and colscore are the scores assigned by the Erik+2 method after normalizing rows and columns respectively. After computing the overall score we can divide by either the number of SVDs done (so as to compare our score to scores coming from partitions with different size) or by the expected number of SVDs for that size of the partition, in order to keep a penalty to flattening matrices which require a higher number of SVDs because they have a higher number of columns.

Another option (which takes more into account the algebraic nature of the problem) is to notice that $\mathcal{V}$ is an algebraic variety (since it can be described in terms of a system of polynomial equations because the condition to be rank 4 is to have all 5×5 minors equal to zero), which dimension varies when we change the dimensions of the matrix, but were not able to reach a good conclusion from this idea.

IV. PERFORMANCE OF THE METHOD

To test the performance of our method and to compare it to the original Erik+2 method, we were given a set of 100 data files corresponding to trees with 12 leaves with the same topology but with random branch lengths. For every data set, we obtained the scores for 9 partitions, 3 of size 2×10, 3 of size 3×9 and 3 of size 5×7, where one partition of each size was proper and the rest were not.

The following table contains the information of the performance for the following methods: sc_1 is the number of rows of the flattening matrix, sc_2 the number of columns, sc_3 the score given by the original Erik+2 method, sc_4 the Erik+2 using the $mn/\sqrt{e}$ normalization, sc_5 our score without dividing for the number of SVDs, sc_6 our score taking the arithmetic mean of the scores for each submatrix, and sc_7 our score taking a pondered mean of the sub-scores.

TABLE I. Percentage of success for the different methods (we count draws as a success)

Partition	sc_1	sc_2	sc_3	sc_4	sc_5	sc_6	sc_7
2 vs 10	100	64	33	40	**70**	65	67
3 vs 9	100	50	39	31	**42**	35	36
5 vs 7	*97*	*76*	**58**	21	47	24	26

TABLE II. Average of the score given to proper partitions

Partition	sc_1	sc_2	sc_3	sc_4	sc_5	sc_6	sc_7
2 vs 10	16	57418	3209	181	9972	176	177
3 vs 9	62	38453	13926	296	10929	291	291
5 vs 7	890	8745	75489	589	4530	620	677

TABLE III. Average of the score given to wrong partitions

Partition	sc_1	sc_2	sc_3	sc_4	sc_5	sc_6	sc_7
2 vs 10	16	59347	3206	181	10502	179	183
3 vs 9	64	39422	14396	293	11134	288	288
5 vs 7	954	9816	87601	560	4814	584	638

V. CONCLUSIONS

We can see that our method (score 5) works significantly better than the original Erik+2 method since it recognizes the proper partition out of the three 70 out of 100 times, since the original method worked fine only 33% of the time. This could be explained by the fact that the Erik+2 method does a single SVD where only 16 singular values are obtained (notice that the Erik+2 method is more accurate as the partition is more balanced), and that the dispersion present in those type of matrices fits nicely with our model.

For the 3 vs 9 case our method turns out to be slightly better but not significantly, but neither the original method nor ours provided a satisfactory result, so we thing new ideas should be introduced to deal with this problem. In the 5 vs 7 case the most effective score turns to be the original Erik+2 method, but we should notice that the percentage of success of taking the score as simply the number of columns is really high and the averaged difference of columns between proper and wrong partitions is percentage-wise the most significant. The number of rows, which was not reliable for the other partitions, could be considered as the number of rows increases and for this case it gives a huge percentage of success.

We can also see that while we have reduced the relative difference between scores when averaging (although by doing this we are decreasing the percentage of success) those scores are not yet comparable. A noticeable fact is that for the method that works better (without averaging) scores obtained for the first two sizes are really close, but for the 5 vs 7 it reduces to less than one half (this is due to the fact that we make much less SVDs, as one can see looking at the averaged score), while for the original Erik+2 the score shows a steady increasing trend when the partition gets more balanced. We have to keep in mind that our estimations were for the 2-norm of the matrix, while we were actually dealing with the distance to a variety, so we are not taking into account the first 4 singular values and we do not know whether this is going to make the score change always in the same way, and also that score differences between proper and wrong partitions are not big enough to ensure that if we were able to reduce them to a common scale we would be able to distinguish between proper and wrong partitions which have different size.

[1] M. Casanellas, *Algebraic tools for evolutionary biology*, EMS Newsletter, December 2012, p.12-18.

[2] M. Garrote, *Phylogenetics and rank of matrices*. Bachelor's degree thesis at FME. Advisor: M. Casanellas.