Correlation

- In probability theory and statistics, correlation indicates the strength and direction of a linear relationship between two random variables. In general statistical usage, correlation or co-relation refers to the departure of two random variables from independence. In this broad sense there are several coefficients, measuring the degree of correlation, adapted to the nature of the data.
Example

Pearson's Correlation Coefficient

**Definition:** Measures the strength of the linear relationship between two variables.

**Characteristics:** Pearson’s Correlation Coefficient is usually signified by $r$ (rho), and can take on the values from -1.0 to 1.0. Where -1.0 is a perfect negative (inverse) correlation, 0.0 is no correlation, and 1.0 is a perfect positive correlation.

$$r = \frac{\sum XY - \frac{\sum X \sum Y}{N}}{\sqrt{\left(\sum X^2 - \frac{(\sum X)^2}{N}\right) \left(\sum Y^2 - \frac{(\sum Y)^2}{N}\right)}}$$

---

**Phylogenetic & Correlation**

- Much of evolutionary biology is inherently historical in nature. Evidence of Natural selection can often be detected by comparing characters across a number of species (Rannala and Huelsenbeck, 2003).

- Comparison among species often examine whether some characters of interest, such as, a developmental mechanism correlates with some other aspect of organisms or their phylogeny, such as life-history strategy, ecological specialization or speciation rates (Maddison, 2000).
Opinion of Character Correlation

- A correlation would suggest an evolutionary process linking to the different features.
- The degree to which ecological or behavioral changes and morphological developmental ones go together can only be investigated when a particular kind of evolutionary change has recurred often within a higher taxon (Arthur and Cubo, 2001).

Testing Character Correlation using Pairwise Comparisons on a Phylogeny

WAYNE P. MADDISON*

*Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, AZ 85721, U.S.A.

Goal:
• To develop the method of pairwise correlation in detail rather than judge the method.
• Focus on pairwise comparison to the terminal taxa.
Some assumptions increase the power of the method, but they may limit its applicability and may lead to incorrect results if assumptions are not met.
Pairwise comparison

<table>
<thead>
<tr>
<th>Advantage</th>
<th>Disadvantage</th>
</tr>
</thead>
<tbody>
<tr>
<td>• Avoid assumptions about ancestral states, branch lengths and elaborate models of evolution (Maddison, 2000).</td>
<td>• It lose information in focusing on only a subset of branches and comparisons (Felsenstein, 1985)</td>
</tr>
<tr>
<td>• Avoid some of the misinterpretations that other methods can yield when other factors cause clustering of changes (Read &amp; Nee, 1995)</td>
<td>• Pairwise comparison are not guaranteed to be statistically independent</td>
</tr>
</tbody>
</table>

• Pairwise test can be valuable as well with complete data and a well-resolved phylogeny, because they avoid some of the assumptions of more parametric methods. With a complete data, a difficult of choice may accompany the test.

• Pairs of taxa must be chosen so that the comparisons are phylogenetically separated from one another, but this constraints does not necessary a unique selection of pairs.
A bat might instead have been compared to squairrel
A study of the correlates of active flight could compare a bird and a non-avian dinosaur such as Velociraptor, a bat and a mouse, a pterosaur and a lizard, and a lepidopteran and a collembolan (Fig. 1). These four pairs are separated on the phylogeny, and represent independent comparisons in the sense that a separate evolutionary change between active flight and its absence must have occurred between the members of each pair. Had nature only supplied us with enough pairs, a general correlation might be found using sign tests or other basic statistical tests. For instance, the flying member of each pair may consistently show a higher value in some variable of interest than the non-flying member.

**Comparison of terminal taxa**

First character vs second character

![Character comparison diagram](image)

Fig. 2. Example with two characters (one with states A, a; other with states B, b) on phylogenetic tree with 20 terminal taxa. Phylogenetically separate comparisons are sought to test whether the states of the two characters are correlated.

- With a complete data, a difficult of choice may accompany the test.
- If the result of a test dependent on selection, and a particular set of taxon pairs is chosen by the investigator, then an arbitrary or subjective element could enter the interpretation of character association.
- For this reason, automated means to select pairing of taxa would be valuable, so that the arbitrary choice could be eliminated.
Pairwise test can be valuable as well with complete data and a well-resolved phylogeny, because they avoid some of the assumptions of more parametric methods. With a complete data, a difficult of choice may accompany the test.

zxiao, 3/29/2009
Criteria for Choosing Pairs

• One obvious criterion is to prefer pairing with as many comparison as possible (Purvis & Bomham, 1997). The more pairs, the greater the sample size in the pairwise comparison.
• It should represent a comparison relevant for the question of interest.
• If the dependent character is continuous variable, then its value is likely to differ between any two taxa, and thus any pair differing in the independent character could be useful.

Figs. 3–6: A tree of seven terminal taxa with a single character with states A, a, to illustrate alternative pairings that could be chosen. Bold lines indicate pairs of terminal taxa selected for comparison.

---


Research article

Patterns of correlated character evolution in flightless birds: a phylogenetic approach

JORGE CUBO1* and WALLACE ARTHUR2

1Equipe ‘Formations squelettiques’, UMR CNRS 8570, Université Paris 6, 2, Place Jussieu – Case 7077, F-75251 Paris Cedex 05, France; 2Ecology Centre, School of Sciences, University of Sunderland, Sunderland, SR1 3SD, United Kingdom
(*author for correspondence; tel.: +33-1-44-27-31-24; fax: +33-1-44-27-36-53; e-mail: cabo@ccr.jussieu.fr)

Goal: Map the evolution of different characters onto phylogenies.
• The cladistic techniques and the addition of molecular to morphological data have resulted in robust phylogenies for at least a handful of higher taxa.
• Study the degree to which ecological changes and morphological ones go together can only be investigated when a particular kind of evolutionary change has recurred often within a higher taxon.
Two Necessary Questions

- Is a particular morphological change associated with a particular ecological change to a sufficient extent that we can be confident the association is not simply a result of chance?
- Considering other morphological characters that might expect to be correlated with the “prime” character under consideration, is evidence that these do indeed co-vary as expect (Arthur and Cubo, 2001)?

Phylogenetic Hypothesis of Birds

<table>
<thead>
<tr>
<th>Major Group</th>
<th>Material</th>
<th>Cons and Pros</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cracraft 1988</td>
<td>Morphological data</td>
<td>Not consider Psittaciformes</td>
</tr>
<tr>
<td>Sibley 1990</td>
<td>DNA/DNA hybridization</td>
<td>Good for conducting large scale analysis, but controversial</td>
</tr>
<tr>
<td>Mindel 1997</td>
<td>Mitochondrial DNA</td>
<td>Not consider Psittaciformes</td>
</tr>
</tbody>
</table>
Concentrated Changes Test

This is a method for testing whether changes in one character are associated phylogenetically with the state of another character. Supposes that character evolution on the phylogeny is first reconstructed, that is, the states of hypothetical ancestors are reconstructed.

This reconstruction would most likely be done for binary characters using *parsimony algorithms* presented by Fitch (1971) and Hartigan (1973).
Figure 1 shows a phylogenetic tree with the evolution of two characters mapped onto it. Evolution in the first character is indicated by stated name (0 or 1) beside the tick marks on the branches; evolution of the second character is indicated by the shading of branches (black for one state, white for the other). Different conventions are used to display evolution of two characters to facilitate discussion: in subsequent discussion, any reference to state 0 or 1 at node refers to the first character; any reference to black or white refers to the second character.

In this example, there are five gains (0 to 1 transition) of state 1 and one loss (1 to 0 transition) to state 0 in the first character.

The incidence of flightlessness has been traced on to the phylogenetic tree (Black branch). Afterwards, the occurrence of peramorphic features of the plevic apparatus has also been traced onto the tree (Horizontal bars). This concentrated changes test gives us the probability, under the null hypothesis that gains and losses are randomly distributed over the branches, that seven gains of peramorphic pelvic apparatus would occur in the whole clade.
Moreover, the independent variable was traced in the tree. The presence of peramorphic features in the skull has also been traced onto the tree (Horizontal bars).

In conclusion,
1. the concentrated changes test shows that the association between this ecological change and this trait is significant.
2. Cases of peramorphosis of plevic apparatus and adaptations associated with a compensatory shift in appendicular specialization of hind limbs to running or to foot-propelled driving, which have been repeatedly selected in many clades of flightless birds.
Phylogenetic Tree for Genomic Research

Calculating the similarity matrix

<table>
<thead>
<tr>
<th>Gene 1</th>
<th>Gene 2</th>
<th>Gene 3</th>
<th>Gene 4</th>
<th>Gene 5</th>
<th>Gene 6</th>
<th>Gene 7</th>
<th>Gene 8</th>
<th>Gene 9</th>
<th>Gene 10</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>1.3</td>
<td>0.2</td>
<td>0.8</td>
<td>-1.2</td>
<td>0.3</td>
<td>0.7</td>
<td>-1.5</td>
<td>0.4</td>
<td>-0.5</td>
</tr>
<tr>
<td>1.3</td>
<td>1.0</td>
<td>1.0</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.4</td>
<td>0.4</td>
<td>0.5</td>
</tr>
<tr>
<td>0.2</td>
<td>1.0</td>
<td>1.0</td>
<td>0.8</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.4</td>
<td>0.4</td>
<td>0.5</td>
</tr>
<tr>
<td>0.8</td>
<td>0.2</td>
<td>0.8</td>
<td>1.0</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.4</td>
<td>0.4</td>
<td>0.5</td>
</tr>
<tr>
<td>-1.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>1.0</td>
<td>0.2</td>
<td>0.2</td>
<td>0.4</td>
<td>0.4</td>
<td>0.5</td>
</tr>
<tr>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>1.0</td>
<td>0.2</td>
<td>0.4</td>
<td>0.4</td>
<td>0.5</td>
</tr>
<tr>
<td>0.3</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>1.0</td>
<td>0.2</td>
<td>0.4</td>
<td>0.5</td>
</tr>
<tr>
<td>0.7</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>0.2</td>
<td>1.0</td>
<td>0.2</td>
<td>0.5</td>
</tr>
<tr>
<td>-1.5</td>
<td>-0.4</td>
<td>-0.4</td>
<td>-0.4</td>
<td>-0.4</td>
<td>-0.4</td>
<td>-0.4</td>
<td>-0.4</td>
<td>1.0</td>
<td>0.5</td>
</tr>
<tr>
<td>0.4</td>
<td>0.4</td>
<td>0.4</td>
<td>0.4</td>
<td>0.4</td>
<td>0.4</td>
<td>0.4</td>
<td>0.4</td>
<td>0.5</td>
<td>1.0</td>
</tr>
<tr>
<td>-0.5</td>
<td>-0.5</td>
<td>-0.5</td>
<td>-0.5</td>
<td>-0.5</td>
<td>-0.5</td>
<td>-0.5</td>
<td>-0.5</td>
<td>-0.5</td>
<td>1.0</td>
</tr>
</tbody>
</table>

End up with a symmetrical table of Pearson correlations
Different methods of tree building

1. Single-linkage clustering
2. Complete-linkage clustering
3. Average-linkage clustering
4. Centroid linkage clustering

Visualization: Data are often converted to a colorimetric scale

Each box: a transcript measurement
Each row of boxes: transcript measurements for a given gene
Each column of boxes: transcript measurements from a single array

Red: higher transcript abundance in one sample
Green: higher transcript abundance in the other sample
weighted pearson correlation

\[ S_{x,y} = \frac{1}{\sum_{i=1}^{N} w_i} \sum_{i=1}^{N} (X_i)(Y_i) \]

Where \( w_i = \frac{1}{L_i} \)

\[ L(i) = \sum_{\text{all } i \neq j \text{ where } d(i,j) < k} \left( \frac{k - d(i,j)}{k} \right)^n \]

- \( k = \text{array corr. cutoff} \)
- \( d = \text{Pearson distance } (= 1 - \text{Pearson corr}) \)
- \( n = \text{exponent (usually 1)} \)

Gene X: \( X_1, X_2, X_3, X_4, X_5 \)
Gene Y: \( Y_1, Y_2, Y_3, Y_4, Y_5 \)

For example: if these arrays are identical, the data are over-represented 3X

--- can weight experiments \( i = 3,4,5 \) by \( w = 0.33 \)
Can also cluster *array experiments* based on global similarity in expression?

Hierarchical Clustering

This method builds the hierarchy from the individual elements by progressively merging clusters. In our example, we have six elements \{a\} \{b\} \{c\} \{d\} \{e\} and \{f\}. The first step is to determine which elements to merge in a cluster. Usually, we want to take the two closest elements, according to the chosen distance.

Optionally, one can also construct a distance matrix at this stage, where the number in the \(i\)-th row \(j\)-th column is the distance between the \(i\)-th and \(j\)-th elements. Then, as clustering progresses, rows and columns are merged as the clusters are merged and the distances updated. This is a common way to implement this type of clustering, and has the benefit of caching distances between clusters. A simple agglomerative clustering algorithm is described in the single-linkage clustering page; it can easily be adapted to different types of linkage.
Genes that are involved in common biological processes and/or physically interact in protein-protein complexes display very frequently similar expression patterns

So, if two genes display similar expression patterns under a very high number of conditions they are likely related.

Systematic studies have shown that the correlation is quite good; however it is also clear that if two genes are co-expressed in one species, it does not mean necessarily that they are functionally related.

If one should use this criterion to predict a link between two genes, a very high number of false positives must be expected.

zhongyi, 4/6/2009
Hierarchical trees of gene expression data are analogous to phylogenetic trees

Distance between genes is proportionate to the total branch length between genes (not the distance on the y-axis)

Orientation of the nodes is irrelevant … although some clustering programs try to organize nodes in some way.

Yap1p targets are coregulated in response to oxidative stress but not other conditions
Hierarchical clustering
Advantages v.s. Disadvantages

Advantages:
1) Straightforward
2) Captures biological information relatively well

Disadvantages:
1) Doesn’t give discrete clusters, and need to define clusters with cutoffs
2) Hierarchical arrangement does not always represent data appropriately sometimes a hierarchy is not appropriate: genes can belong to one cluster.
3) Get different clustering for different experiment sets

THERE IS NO ONE PERFECT CLUSTERING METHOD

K-means clustering

Approximate k-means algorithms have been designed that make use of coresets: small subsets of the original data.
The quality of the final solution depends largely on the initial set of clusters, and may, in practice, be much poorer than the global optimum
 Partitioning (or top-down) clustering method

- Randomly split the data into $k$ groups of equal number of genes
- Calculate the centroid of each group
- Reassign genes to the centroid to which it is most similar
- Calculate a new centroid for each group, reassign genes, etc ... iterate until stable

$k$-means clustering

What are the disadvantages of $k$-means clustering?

- Need to know how many clusters to ask for
  (can define this empirically)

- Genes are not organized within each cluster
  (can hierarchically cluster genes afterwards or use SOM analysis)

- Random process makes this an indeterminate method
What kinds of information can we extract from whole-genome expression data?

1. Hypothetical functions for uncharacterized genes
   -- genes encoding subunits of multi-subunit protein complexes
   are often highly coregulated
   example: ribosomal protein genes, proteasome genes in yeast
   -- genes involved in the same cellular processes are often coregulated

2. New roles for characterized genes

2. Better understanding of the experimental conditions
   -- based on expression patterns of characterized genes

3. Implications of gene regulation
   -- WT vs. mutants can identify transcription factor targets
   -- promoter analysis of coregulated genes = upstream elements
   -- gene coregulation with known pathway targets can implicate
     pathway activity

5. Understanding developmental pathways

6. Defining experimental samples based on expression profiles
   example: comparing tumor samples from patients

Software for clustering and visualization:

Cluster (Mike Eisen): http://rana.lbl.gov (for PC only)

Cluster (de Hoon): http://bonsai.ims.u-tokyo.ac.jp/~mdehoon/software/cluster/

Java Treeview (Alok Saldana): http://jtreeview.sourceforge.net/