Syst. Biol. 57(3):432–448, 2008
Copyright © Society of Systematic Biologists
ISSN: 1063-5157 print / 1076-836X online
DOI: 10.1080/10635150802172192

Sympatry Inference and Network Analysis in Biogeography

DANIEL A. DOS SANTOS, HUGO R. FERNÁNDEZ, MARÍA GABRIELA CUEZZO, AND EDUARDO DOMÍNGUEZ

CONICET, Facultad de Ciencias Naturales, Universidad Nacional de Tucumán, Miguel Lillo 205, 4000 Tucumán, Argentina; E-mail: mayfly@unt.edu.ar (E.D.)

Abstract.— A new approach for biogeography to find patterns of sympatry, based on network analysis, is proposed. Biogeographic analysis focuses basically on sympatry patterns of species. Sympatry is a network (= relational) datum, but it has never been analyzed before using relational tools such as Network Analysis. Our approach to biogeographic analysis consists of two parts: first the sympatry inference and second the network analysis method (NAM). The sympatry inference method was designed to propose sympatry hypothesis, constructing a basal sympatry network based on punctual data, independent of a priori distributional area determination. In this way, two or more species are considered sympatric when there is interpenetration and relative proximity among their records of occurrence. In nature, groups of species presenting within-group sympatry and between-group allopatry constitute natural units (units of co-occurrence). These allopatric units are usually connected by intermediary species. The network analysis method (NAM) that we propose here is based on the identification and removal of intermediary species to segregate units of co-occurrence, using the betweenness measure and the clustering coefficient. The species ranges of the units of co-occurrence obtained are transferred to a map, being considered as candidates to areas of endemism. The new approach was implemented on three different real complex data sets (one of them a classic example previously used in biogeography) resulting in (1) independence of predefined spatial units; (2) definition of co-occurrence patterns from the sympatry network structure, not from species range similarities; (3) higher stability in results despite scale changes; (4) identification of candidates to areas of endemism supported by strictly endemic species; (5) identification of intermediary species with particular biological attributes. [Betweenness; clustering coefficient; dot maps; historical biogeography; intermediary species; units of co-occurrence.]

> Geographic distribution—As such, we will understand, as a simple fact of nature, the records of occurrences at different points of the map of modern world of consanguineous entities forming a taxon.
>
> L. Croizat, 1964:13

The main goal in vicariance biogeography is the determination of area relationships, based on biota distributions (Nelson and Platnick, 1981). Areas of endemism are traditional units for historical biogeography. Several approaches have been proposed to identify and delimit these basic units of analysis in recent years, but all have particular methodological problems (Nelson and Platnick, 1981; Platnick, 1991; Morrone, 1994; Hausdorf, 2002; Szumik et al., 2002; Hausdorf and Hennig, 2003; Mast and Nyfeller, 2003; Szumik and Goloboff, 2004; Hennig and Hausdorf, 2006). Problems common to most of the published works are the criteria to identify areas of endemism and the excessive emphasis given to recognizing their borders (Hausdorf, 2002; Mast and Nyffeler, 2003). Some authors (Nelson and Platnick, 1981; Morrone, 1994; Linder, 2001) established as a criterion for identification that extensive sympatry of at least two species is necessary to identify an area of endemism. Harold and Mooi (1994) and Hausdorf (2002) suggested that extensive co-occurrence of taxa is neither sufficient nor necessary for recognizing an area of endemism. Moreover, Hausdorf (2002) suggested that extensive co-occurrence does not delimit areas of endemism but rather it delimits biotic elements. These previous works emphasized determining how extensive sympatry is between two taxa instead of focusing on the meaning of sympatry. Neither "extensive sympatry" nor "extensive co-occurrence" has been properly defined.

The a priori delimitation of areas of endemism is only possible when the original biotas were separated by a vicariant event, and no subsequent dispersal occurred. Dispersal events obscure the history of the areas, their limits and their significance. Sympatry is inferred by the overlap of ranges of different species.

It is possible to distinguish two modes for deriving species ranges from the literature: raster-like mode and vector-like mode (e.g., Rapoport, 1982; Posadas, 1996; Rathert et al., 1999; Unmack, 2001). The raster-like mode divides the study area into predefined spatial units. The range of a species is the subset of spatial units where there is at least a record. Under this mode, the most common alternative is to plot records on a map divided into quadrants or grids. Here, the species co-occurring in the same cells are considered sympatric. Szumik et al. (2002:807) stated that "The use of grid system seemed unavoidable…" to convert sets of species records into ranges. However, the use of this methodology presents several problems, one of them being scale dependence. As Hausdorf and Hennig (2003:721) stated: "If the grid used is too fine and the distribution data are not interpolated, insufficient sampling may introduce artificial noise in the data set." On the contrary, if the grid size is too coarse "distinct biotic elements may be amalgamated." Also, different grid cell sizes, shapes, and positions have been proposed (Posadas, 1996; Rathert et al., 1999) that alter the results. The grid method is especially unsuitable for mapping river organisms, due to the linear and diverging nature of riverine habitats. For example, a single grid cell could expand ranges to areas without freshwater habitats, that could never have aquatic organisms. On the other hand, headwaters from different basins, with different histories, could end up in the same cell, depending on the scale chosen.

The vector-like mode for deriving species ranges does not divide the study area. Here, species ranges are independently obtained, emerging from the records.

Sympatry is inferred by geometric analysis of range overlap. Under this mode, some techniques of mapping involve contour maps, convex hulls (IUCN, 1994), and Rapoport's mean propinquity method (Rapoport, 1982). The principal problem associated with these procedures is that they emphasize extremes of occurrence and assume homogeneity, convexity and radial nature of distributions.

A more conservative alternative, appealing to direct evidence, is a dot map where records themselves account for the range. Prospecting sympatry via direct evidence is advisable because minimizes distributional assumptions. This proposition requires a new operational definition of sympatry. We consider that two or more species are sympatric when there are *interpenetration* and *relative proximity* among their records. Coincidence of two or more species in the same locality is the maximum expression of these properties.

Biogeographic analysis focuses basically on sympatry patterns of species. This poses a dual challenge. The first is to propose hypotheses of sympatry based on the available data. The second is to analyze these hypotheses to find possible patterns.

The geographic records—the minimal informative unit—correspond to punctual data. As a rule, these punctual data are integrated in different ways to determine the taxon "distributional area." Unfortunately, the "real" distributional areas are beyond our knowledge, and their derivation from the available records is always dependent on artificial assumptions, regardless of the chosen procedure. For this reason, it is imperative to obtain reliable sympatry hypotheses without a priori range delimitation. If we consider the field collection records as *spatial signals* of the real ranges, we can focus on the numerical parameters of these signals to infer the relations among ranges, instead of converting the punctual records into distributional areas before inferring their relationships. Once the hypotheses of sympatry are obtained, they must be analyzed under an adequate theoretical framework.

Network analysis has proven to be a powerful tool for the study of different aspects of biological systems (molecular, ecological, and evolutionary levels; Proulx et al., 2005; Montoya et al., 2006). Sympatry is a network (= relational) datum, not a conventional (attributive) one. In biogeography, conventional data consist of an array of species by locations. In this way, locations of species occurrence are spatial attributes of that species and, conversely, species at a location are biotic attributes of that location. On the other hand, network data consist of an adjacency matrix of measurements, where each entry expresses the relation between pairs of species. The major difference between conventional and network data is that conventional data focus on species and attributes, whereas network data focus on species and relations (Hanneman and Riddle, 2005). Network analysis is designed to detect holistic patterns in the overall sympatry network in which species are embedded.

Species groups satisfying the requirement of within-group sympatry and between-groups allopatry will conform to natural units of co-occurrence (UCs hereafter). Natural UCs are embedded into a more global network when there are other species connecting them. A sympatry network may contain dense groups of species connected through intermediary species. The intermediary species will have higher "connecting" capacity than the other species that are inside the units. The UCs will be evident after the intermediary species are removed. The spatial expression of the resulting UCs will be our candidates for areas of endemism. The species belonging to each UC are strictly endemic (sensu Anderson, 1994). The final status of the candidate areas of endemism will depend on the congruent historical (phylogenetic) relationships of the involved taxa (Humphries and Parenti, 1999; Mast and Nyfeller, 2003). We propose using network analysis to study species sympatry patterns, as a first step in the search for bio-spatial units in biogeography.

Methodological Development

Distributional Data and Sympatry Matrix

Let $S = \{\mathrm{r}_1, \mathrm{r}_2, \ldots, \mathrm{r}_N\}$ the set of N punctual records associated to treated species. Equivalently, $S = \{S_1 \text{ U } S_2 \text{ U } \ldots \text{ U } S_n\}$ is also obtained as the union of record subsets pertaining to n treated species. It is assumed that records are informative fragments of the respective ranges. If the ranges overlap, that is, if the considered species are sympatric, their punctual records must satisfy two properties: (1) relative proximity (*Property I*); and (2) interpenetration (*Property II*).

Relative proximity.—There is relative proximity between two points when the geographic distance between them is smaller or equal than a specified reference value. We need then to define this reference value and establish a procedure to detect dot clouds that comply with this property. This will allow us to extract these clouds from the whole set of records, after evaluating the proximity of the individual points with respect to their neighboring ones.

We opted for the Delaunay triangulation method of the records, which is able to preserve the data proximity information (Bhattacharya and Gavrilova, 2006). Recently, some triangulation-based algorithms have been used to recognize dot patterns (Bhattacharya and Gavrilova, 2006; Papari and Petkov, 2005). The advantage of this method is its capability to detect groupings with different forms, including nonconvex shapes and irregular borders. Briefly, in the Delaunay triangulation, the points are completely interconnected by segments, in such a way that the segments do not cross. For more detail see De Berg et al. (2000). The triangulation segments can be either interspecific or intraspecific, depending on the biological identity of the connected records. To detect coherent dot clouds, it is essential to identify and trim the nonpertinent segments. A segment becomes nonpertinent when its geographic length exceeds the proximity limit admitted by the extreme points; very long segments suggest disjunct ranges.

There is no absolute measure to consider a segment as too long. Its length must be contrasted with the reference values of its end points. Each point has two proximity reference values, one for intraspecific and one for interspecific comparisons. Those values help to decide what distance around a point may be considered as critical for pertinent vicinity.

The geographic distance D that separates the end points of the segment $\overline{rr'}$ can be calculated by the haversine formula:

$$D(r, r') = 2 \times Rd \times \arcsin \sqrt{\begin{matrix} \sin^2\left[\frac{(lat_r - lat_{r'})}{2}\right] + \cos(lat_r) \\ \times \cos(lat_{r'}) \times \sin^2\left[\frac{(lon_r - lon_{r'})}{2}\right] \end{matrix}} \quad (1)$$

$$\forall r, r' \in S \wedge r \neq r'$$

lat_r, $lat_{r'}$: latitude of the points r and r' in radians
lon_r, $lon_{r'}$: longitude of the points r and r' in radians
Rd: radius of the geoid adopted

We adopted the World Geodetic System of 1984 geoid with $Rd = 6{,}378{,}137$ m, because it is the official GPS reference system (El-Rabbany, 2002).

Concerning interspecific comparisons, two punctual records from different species are in *relative proximity* when the distance between them is smaller than the distance separating each of them from its nearest intraspecific neighbor. For each point or record r, its relative proximity for interspecific comparisons (*NRinter*$_r$) is

$$NRinter_r = \min\ D(x, r) \quad (2)$$

$\forall x \in S_i \wedge x \neq r$, S_i being the subset of S associated to species i recorded at r.

In the triangulation context, a pair of points exhibits clear interspecific proximity when the segment that unites them is shorter than their *NRinter* values. However, the retention of an interspecific segment becomes controversial when its length exceeds one of the two end points *NRinter* values. To solve this, we appeal to the weighted average of proximity values, with emphasis on the lower of both *NRinter* values, such as a balance between the more stringent value and the semisum of both extreme values is achieved. Thus, the threshold cut for an interspecific segment (*CTinter*) uniting points r and r' is

$$CTinter(r, r') = 0.75 \times \min(NRinter_r, NRinter_{r'}) + 0.25 \times \max(NRinter_r, NRinter_{r'}) \quad (3)$$

Two considerations are necessary here. First, if two or more species share the same record r, the *NRinter* value of r will be the minimum of all possible values obtained with Equation 2. Second, when the point r corresponds to the single record of a species, *NRinter* will be equal to 0.

Concerning intraspecific comparisons, we need to discern records from disjunct range portions as species ranges are rarely continuous. Without evidence to the contrary, it is reasonable to assume that a record and its nearest intraspecific neighbor belong to a common range portion. The geodetic distance to the nearest intraspecific neighbor varies within each subset S_i. So, for each species i, the proximity reference value for intraspecific comparison (*NRintra*$_i$) will be the mean distance between each record of S_i and its respective intraspecific nearest neighbor:

$$NRintra_i = |S_i|^{-1} \sum_{x=1}^{|S_i|} \min D(x, S_i^x)\ \forall x \in S_i; |S_i| > 1 \quad (4)$$

$|S_i|$: number of records of species i
S_i^x: set of records of species i excluding record x

In the triangulation, an intraspecific segment becomes nonpertinent when its length is beyond the mutual reach of their end points. Given a pair of end points (r, r') belonging to species i, the cut threshold for the intraspecific segment (*CTintra*) will be twice the *NRintra* value of species i records:

$$CTintra(r, r') = 2 * NRintra_i\ \forall r, r' \in\ S_i \quad (5)$$

When both r and r' belong to more than one species, *CTintra(r, r')* will be equal to the highest value obtained with Equation 5, because if cutting proceeds at this value, it also proceeds at the lower ones. Ambiguous segments, that is, segments with lists of species associated to each end point that are partially overlapped (e.g., species A and B in one end point and species A and C in the other), should be treated as intraspecific.

The original Delaunay triangulation, a connected undirected graph, is transformed into a reduced Delaunay triangulation, which may be a disconnected graph, after deletion of nonpertinent segments. The components detected in the reduced Delaunay triangulation are the dot clouds of interest. In reduced Delaunay triangulation, two points are members of the same component if there is a path connecting them. Partial sympatry is hypothesized for species with records in the same component. The final result is then a square matrix with binary entries indicating sympatry (1) or allopatry (0) between species.

Interpenetration.—We consider interpenetration between S_i and S_j if they are spatially intertwined, suggesting that species i and j share a common range portion. In a simple isotropic model, each species i can be found in a radius rad_i around a representative point p_i. Each species has its natural p_i in the record where the geographical distance rad_i to its outermost record is minimized. In this way, p_i will be the center of the minimum circle enclosing all the records of S_i.

Specific p_i and rad_i parameters are submitted to a test to analyze the interpenetration between each pair of species. Thus, for the pair of species (i, j), the input variables will be rad_i, rad_j, p_i, and p_j. The test evaluates if there is at least one record of species i with a distance to point p_j not higher than rad_j (or vice versa for records of species j) and can be formulated as the logical question:

$$\text{Is min } D(S_j, p_i) <= rad_i \text{ OR min } D(S_i, p_j) <= rad_j? \quad (6)$$

min D(S_j, p_i): minimum distance between records of species j and p_i

After completing all pairwise comparisons, a second square matrix is produced, with 1 for positive and 0 for negative test results. There are also special situations that must be considered. Strict acceptance of the *rad* definition implies that species with a single record have a *rad* equal to 0. Singletons will then have an interpenetration of range only with species that were found at the exact same location. On the other hand, although exceptional, a species may exhibit more than one *p*; e.g., species with only two records. In this case, each pair of parameters (*p*, *rad*) detected is submitted separately to the interpenetration test.

Final hypotheses generation.—Given *n* species, the final sympatry matrix **M** is an $n \times n$ adjacency matrix. Each m_{ij} entry denotes presence (1) or absence (0) of a sympatry link between *i* and *j* species. Sympatry matrices are symmetric ($m_{ij} = m_{ji}$, because sympatry is a reciprocal relation) with all the main diagonal terms equal to 1 ($m_{ij} = 1$ when $i = j$, because each species is sympatric with itself). **M** is obtained from the Hadamard product (i.e., element by element product) of the Property I and Property II matrices. In other words, **M** is the strict consensus of unity scores between both strategies. Thus, species will be considered sympatric when their records *interpenetrate* and belong to the same dot cloud of *relative proximity*.

If records are evenly scattered, without noticeable gaps, the first property is prone to overestimate sympatric links by an effect of concatenation. In other words, distant species may remain in a common dot cloud because of inner bridges. For this reason, the complementary use of interpenetration and relative proximity properties was designed to avoid errors of considering sympatry when there is clear allopatry.

Hypothetical example.—Figure 1 illustrates the records associated to 12 species spread over three areas enclosed with a dashed line. These areas represent the subjacent range portions, arbitrarily known here but ever elusive to our understanding of the real world. Peripheral areas are circular. The central area has a more complex shape (derived from a vertical displacement of a cubic polynomial curve).

Species {A, B, C, D} share the upper circular area, {E, F, G} occupy the middle area, and {H, I, J} the lower circular area. On the other hand, species K and L have disjunct ranges. Species K inhabits the upper and middle areas, while L is found in middle and lower areas. For each species, we randomly selected 5 points from its range portion.

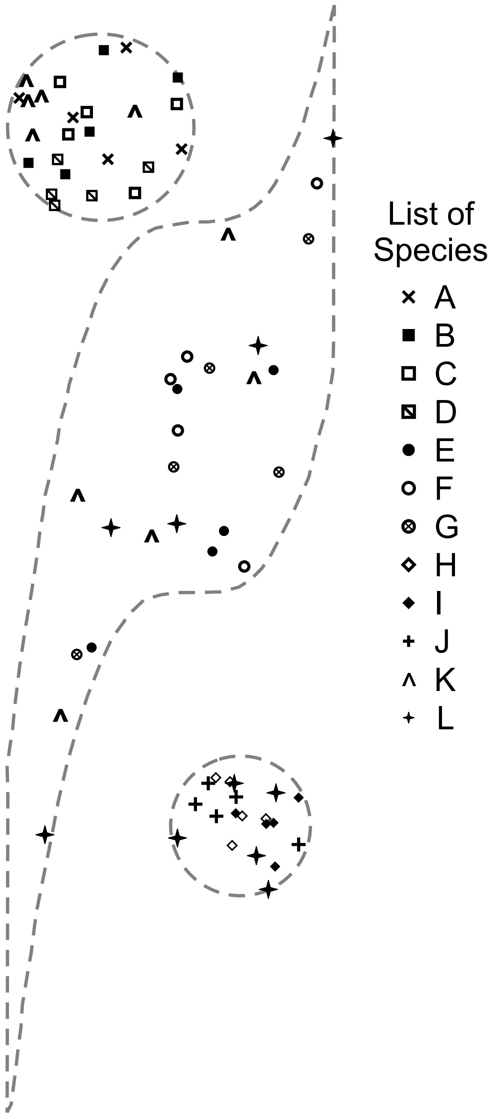
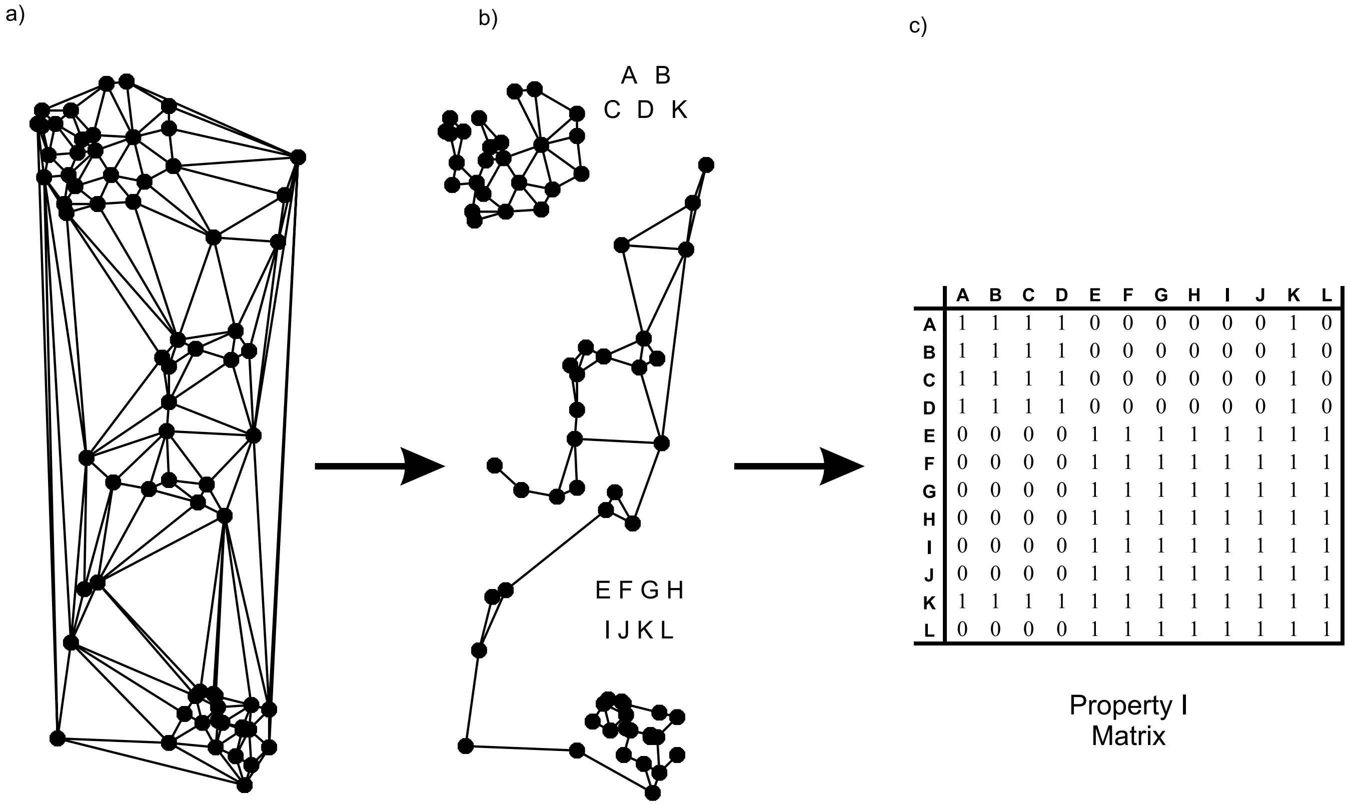


FIGURE 1. Distributional data (randomly generated) of 12 hypothetical species. For more details, see text.

The first phase of sympatry inference starts with the Delaunay triangulation (Fig. 2a) of all records. The Delaunay triangulation evolves to the reduced Delaunay triangulation after removal of nonpertinent segments (Fig. 2b). Finally, a binary matrix indicating inclusion (1) or not (0) of a pair of species into a common dot cloud is generated (Fig. 2c). The second phase of sympatry inference starts with the determination of *rad* and *p* parameters. For each species, *p* and its farthest intraspecific neighbor are located (arrows in Fig. 3a). The distance between origin and head of arrow corresponds to *rad* value. The interpenetration criterion is applied and the respective binary matrix is generated (Fig. 3b). Finally, we obtain our consensus sympatry matrix from the Hadamard product between Property I and Property II matrices. This matrix will be used later during the development of the sympatry network analysis, to obtain the UCs.



	A	B	C	D	E	F	G	H	I	J	K	L
A	1	1	1	1	0	0	0	0	0	0	1	0
B	1	1	1	1	0	0	0	0	0	0	1	0
C	1	1	1	1	0	0	0	0	0	0	1	0
D	1	1	1	1	0	0	0	0	0	0	1	0
E	0	0	0	0	1	1	1	1	1	1	1	1
F	0	0	0	0	1	1	1	1	1	1	1	1
G	0	0	0	0	1	1	1	1	1	1	1	1
H	0	0	0	0	1	1	1	1	1	1	1	1
I	0	0	0	0	1	1	1	1	1	1	1	1
J	0	0	0	0	1	1	1	1	1	1	1	1
K	1	1	1	1	1	1	1	1	1	1	1	1
L	0	0	0	0	1	1	1	1	1	1	1	1

FIGURE 2. First phase of sympatry inference. (a) Delaunay triangulation. (b) Reduced Delaunay triangulation. Dot clouds with associated species (capital letters). (c) Property I matrix.

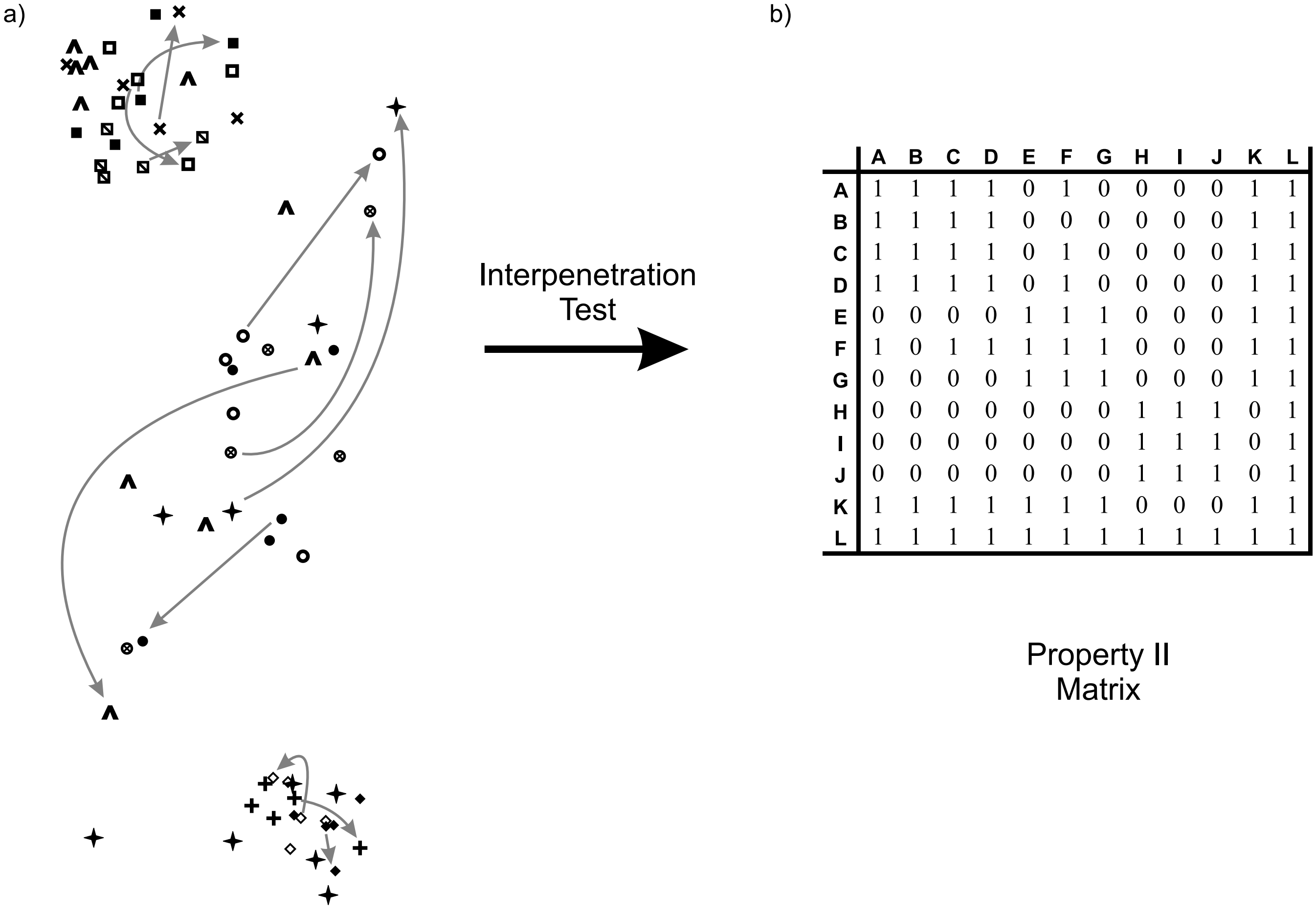
Alternative strategies for inferring sympatry hypothesis.—Distributional data are frequently available in the form of presence/absence matrices of species in operative geographical units (grids, political divisions, basins, etc.). The amount of data published in this way, force us to consider other alternatives for inferring sympatry. In these cases, ranges consist of the sum of locations where species occur and, consequently, range overlap (sympatry) proceeds when co-occurrence of species in at least one location has been observed. In case of ranges obtained by a vector-like mode (different from the dot maps already considered), geometric analysis of range overlap encompasses the inference of sympatry. Therefore, sympatry is assumed when the intersection of species ranges is not null.

Sympatry Network Analysis

The network analysis method (NAM) is oriented to identify groups of species that satisfy the requirement of within-group sympatry and between-group allopatry. These groups of species correspond to UCs. In sympatry networks, UCs will be hardly perceived as entities initially. On the contrary, the different UCs are usually embedded into a more global network due to intermediary species connecting allopatric groups. The removal of intermediary species will segregate the UCs. To illustrate the reasoning, let us consider the network associated with the hypothetical sympatry matrix (Fig. 4a).

Intermediary species and betweenness measure.—A sympatry network is described as a graph $G = (V, E)$, where the set V of nodes represents species, and the set E of edges represents sympatric relations. For simplicity, we will consider the graph as *unweighted*, ignoring the strength or intensity of the relations. Let w be a weight *function* on edges, defining $w(e) = 1, e \in E,$ for an unweighted graph. A sympatry network is thus an undirected graph where an edge exists between two species for which there is an inferred sympatry relation in the sympatry matrix, and no edge otherwise. This graph can be either connected or disconnected from the onset, depending on the sympatry relations present in the sympatry matrix.

A *path* from $s \in V$ to $t \in V$ is an alternating sequence of nodes and edges, beginning with s and ending with t, such that each edge connects its preceding node with its succeeding node. The path *length* is the sum of the weights of its edges (Brandes, 2001). A fundamental concept in graph theory is the "geodesic" or shortest path linking two given nodes (Newman, 2001). The minimum length of any path connecting a pair of nodes is called *geodesic distance*. There may be more than one geodesic path between two nodes.



	A	B	C	D	E	F	G	H	I	J	K	L
A	1	1	1	1	0	1	0	0	0	0	1	1
B	1	1	1	1	0	0	0	0	0	0	1	1
C	1	1	1	1	0	1	0	0	0	0	1	1
D	1	1	1	1	0	1	0	0	0	0	1	1
E	0	0	0	0	1	1	1	0	0	0	1	1
F	1	0	1	1	1	1	1	0	0	0	1	1
G	0	0	0	0	1	1	1	0	0	0	1	1
H	0	0	0	0	0	0	0	1	1	1	0	1
I	0	0	0	0	0	0	0	1	1	1	0	1
J	0	0	0	0	0	0	0	1	1	1	0	1
K	1	1	1	1	1	1	1	0	0	0	1	1
L	1	1	1	1	1	1	1	1	1	1	1	1

FIGURE 3. Second phase of sympatry inference. (a) For each species, an arrow connects the representative point (p) to its farthest intraspecific neighbor. (b) Property II matrix.

Freeman (1977) introduced a centrality measure called *betweenness* that can be used to identify those intermediary species. Betweenness (B) is a measure of the frequency that a node occurs in the geodesic path connecting two other nodes:

$$B_{(v)} = \sum_{s \neq v \neq t \in V} \frac{\sigma_{st}(v)}{\sigma_{st}} \tag{7}$$

Let $\sigma_{st} = \sigma_{ts}$ denote the number of geodesic paths from $s \in V$ to $t \in V$, where $\sigma_{ss} = 1$ by convention. Let $\sigma_{st(v)}$ denote the number of geodesic paths from s to t that some $v \in V$ belong to. Thus, the betweenness measure is the number of times a given node occurs on a geodesic path (Borgatti et al., 2002). The best performance algorithms to calculate B are in Brandes (2001) and Newman (2001).

According to Hanneman and Riddle (2005), the betweenness measure of a node indicates its degree of control position in the network. In a *connected* network of sympatry, the shortest paths from the species of one UC to another unit must pass through intermediary species. So, intermediary species will have a high betweenness due to their connecting capacities.

Units of co-occurrence and clustering coefficient.—Two concepts are necessary to introduce here: *density* and *degree of a node* in a binary network. Density is the fraction of possible edges actually present in the network (Equation 8). The degree of node i (k_i) is its number of incident edges (neighbors; Equation 9).

$$Density = \frac{2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} m_{ij}}{n(n-1)} \tag{8}$$

$$k_{(i)} = \sum_{j=1}^{n} m_{ij} - 1 \tag{9}$$

In Equations 8 and 9, m_{ij} represents the element of a sympatry matrix where row i and column j cross. In Equation 8, n is the number of nodes in the network under consideration.

The group of species that integrate a UC must be cohesive. The maximal cohesiveness is reached when each

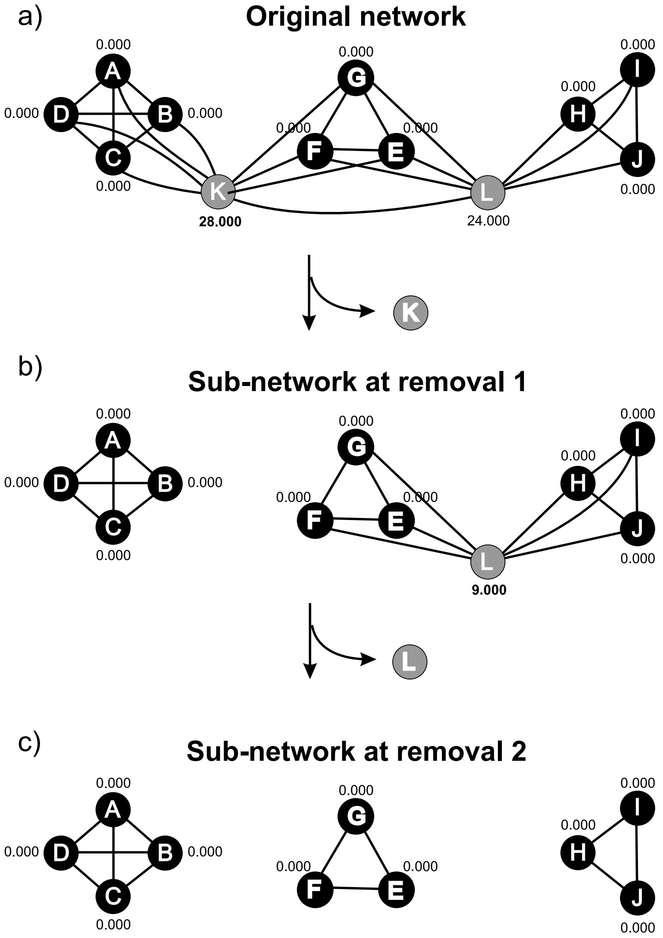


FIGURE 4. Removal process based on betweenness scores. (a) Original network and successive subnetworks with betweenness scores near the respective nodes. Nodes represent species, while edges connect pairs of sympatric species. In the original network, node K obtains the highest betweenness (bold type). (b) The removal of node K with all its incident edges produces the first subnetwork composed of two components. Betweenness scores are recalculated, and node L acquires the maximum betweenness (bold type). (c) After node L removal, the last subnetwork is generated, with three components, and betweenness scores are recalculated again. The removal process stops at this level since all remnant nodes obtain zero betweenness scores.

species is sympatric with all others, a *clique* in graph terms. Watts and Strogatz (1998) introduced the clustering coefficient to account for the tendency in many real-world networks to be structured in dense groups of nodes. That is, the density in local neighborhoods tends to be higher than expected for a random graph of the same size. Clustering coefficient of a node $i(C_i)$ is defined in Equation 10.

$$C_i = \frac{2E_i}{k_i(k_i - 1)}\,(C_i = 0\,if\; k_i < 2) \qquad (10)$$

k_i: degree of node i
E_i: number of links between k_i neighbors of node i

C_i can be interpreted as the probability that any two neighbors of i are connected.

Extracting the units of co-occurrence..—To extract the UCs from the original network (Fig. 4a), it is necessary to identify and remove the intermediary species (Fig. 4b, c). The criterion for removing a species is B (betweenness score). Starting from the original network, the species with the highest B is removed, giving a subnetwork. If there are two or more species with the same highest B, they are removed simultaneously. For each remnant node, B must be recalculated after each removal because the pattern of connections is modified. As nodes are removed from the sympatry network, the graph becomes unconnected and components emerge. The iterative removal stops when all remnant nodes have B equal to zero. One of the subnetworks obtained in this way, the one holding the components (i.e., connected subgraphs) corresponding to the UCs, must be selected.

In the hypothetical example, it is clear that the last subnetwork (Fig. 4c) must be selected. The three components (A-B-C-D; E-F-G; H-I-J) are maximally cohesive by virtue of their clique status. However, UCs do not always correspond to cliques, and successive removals may generate "atomized" subnetworks with many diads (components with two nodes) and isolated nodes. This realistic situation prevents the selection of the last subnetwork by default. On the contrary, all instances of the removal process must be explored to select the one that maximizes a measure sensible to segregation of UCs.

The species of a UC embedded in a major component will have two kinds of neighbors: either other species of the unit, or intermediary species. Therefore, the removal of intermediary species is a selective pruning in the neighborhood of UC species. Thus, to evaluate the segregation of UCs we must measure changes operating at their species neighborhood, using the clustering performance in the neighborhood (*CP*). Given a sympatry network (or subnetwork), for each species i connected to others, CP_i is defined as the minimum clustering value observed among neighbors of i (Equation 11). If species i is an isolated node (without neighbors), a zero is assigned to CP_i.

$$CP_i = \min C_{i'} \;\; \forall i' \in N \wedge i' \neq i \wedge m_{ii'} = 1 \qquad (11)$$

By definition, a clustering value of intermediary species is always suboptimal. In fact, the open neighborhood of intermediary species exhibits species without sympatric links. So, until intermediary species are removed, some nodes of UCs will present suboptimal values of *CP*. As removal progresses and UCs become segregated entities, the neighborhood of their nodes will restrict to themselves. In a cohesive UC, specific *CP* values should rise after successful removal.

A collective increase of *CP* values suggests the arrangement of species into UCs. This increase allows us to propose a global descriptor of successive removals for selecting the subnetwork that holds the UCs. The proposed descriptor is the overall change of clustering performance (*OCP*), which measures the magnitude of change in *CP* values with respect to the original network.

For each subnetwork x generated at level x of the removal process, its OCP_x is obtained as the sum of differences between *CP* value of species i at subnetwork $x(CP_i^x)$ and the *CP* value of the same species at original network (CP_i^0):

$$OCP_x = \sum_{i=1}^{n_x} (CP_i^x - CP_i^0) \qquad (12)$$

n_x: size of subnetwork at level x of removal process

We select the subnetwork that maximizes the *OCP* value. Of the components forming this subnetwork, those composed of more than two nodes will correspond to the UCs. Diads do not constitute UCs because the cohesiveness of their nodes is null ($C_i = 0$). In case of *OCP* ties, the smallest subnetwork is preferred. Table 1 provides the values of *C* and *CP* for every node from both the original and the successive subnetworks of the hypothetical example. The last row of the table shows the gradual increase of *OCP*, with a maximum at the last subnetwork, from which UCs should be extracted.

Testing adequacy of network.—The adequacy of a sympatry network for segregation into UCs must be tested before its analysis. The partition index (*PI*; Equation 13) depicts the tendency of species towards a clustered arrangement. *PI* varies between 0 (sparse nodes) and 1 (compact UCs with few or no intermediary species). *PI* selects the clustering parameter (*C* or *CP*) with the highest value of each species, sum them, and divides the total by the theoretical optimum (= species number).

$$PI = n^{-1} \sum_{i=1}^{n} \max(C_i^0, CP_i^0) \qquad (13)$$

High *C* or *CP* indicates that a species is inside a unit of co-occurrence or in an intermediary condition, respectively. The significance of observed *PI* is tested against a series of *PI*s associated to random networks. If the observed *PI* is significantly higher than those randomly obtained ($P < \alpha$), an arrangement of nodes into different UCs is suggested. In our example, 5000 networks were randomly generated, following a Bernoulli model (each species pair is observed independently and their link is selected with a probability equal to the density of the original network). The adequacy test resulted in a very high value ($PI = 0.905$, $P < 0.001$).

The construction of random networks can be more or less realistic. An alternative to conferring more realism to the network is to randomize the distributional data, constrained by known biogeographical processes, such as autocorrelation and carrying capacity (e.g., Hennig and Hausdorf, 2004). Then, the sympatry network inference can be based on these new random data. However, this procedure has underlying arbitrary assumptions with regards to vagility, the available area for dispersion, etc. Moreover, "... unduly restrictive null models are available that may mask the detection of nonrandom pattern because too much of that pattern may be built into the null model itself" (Moore and Swihart, 2007:764). We prefer to focus on the network structure, avoiding the stochastic generation of ranges.

TABLE 1. Clustering coefficient (*C*), clustering performance (*CP*), and overall clustering performance (*OCP*) along removal process. The intermediary species (K and L) have smaller values of *C* because some pairs of their neighbors lack links. *OCP* values increase gradually to a maximum at sub-network 2 where three ideal components, each one a clique, remain.

	Original network		Subnetwork 1 (after removal of node K)		Subnetwork 2 (after removal of nodes K and L)	
ID node	*C*	*CP*	*C*	*CP*	*C*	*CP*
A	1	0.429	1	1	1	1
B	1	0.429	1	1	1	1
C	1	0.429	1	1	1	1
D	1	0.429	1	1	1	1
E	1	0.429	1	0.4	1	1
F	1	0.429	1	0.4	1	1
G	1	0.429	1	0.4	1	1
H	1	0.429	1	0.4	1	1
I	1	0.429	1	0.4	1	1
J	1	0.429	1	0.4	1	1
K	0.429	0.429	—	—	—	—
L	0.429	0.429	0.4	1	—	—
OCP	0.000		2.681		5.710	

Mapping Units of Co-occurrence

Units of co-occurrence are in a geographical context from which sympatry among species emerges. The geographical correspondence of a UC is its *spatial expression*. The exact procedure to derive spatial expressions will depend on the information source. When distributional matrices are used, the spatial expression will be the union of locations inhabited by the species in question. On the other hand, when individual maps are used, the spatial expression will be the juxtaposition of respective species maps. The spatial expressions of the hypothetical example are represented in Figure 5.

EMPIRICAL EXAMPLES

The performance of our method was evaluated studying the patterns of distribution in three real cases. The formalization of the algorithm is included in Appendix 1.

Case Study 1: Epiphragmophora Doering, 1874, *a Land Gastropod*

This genus is a component of the land snail family Xanthonychidae (Gastropoda: Stylommatophora) and is found exclusively in South America. Distributional information from a data set, consisting of the coordi-

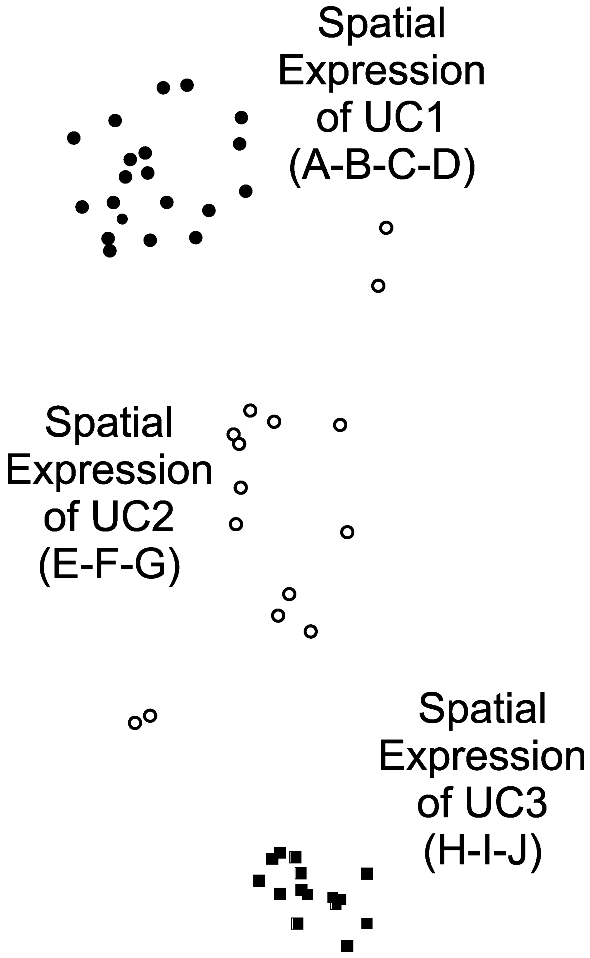

Spatial Expression of UC1 (A-B-C-D)
Spatial Expression of UC2 (E-F-G)
Spatial Expression of UC3 (H-I-J)

FIGURE 5. Spatial expressions of the three units of co-occurrence (UC) detected in the hypothetical example. Supporting species for each spatial expression between brackets.

nates of 145 localities of occurrence of 21 species of *Epiphragmophora* distributed in Argentina, was used and analyzed with the proposed new methodology (Appendix 2; Appendices 2 to 4 are available at the *Systematic Biology* Web site at http://www.systematicbiology.org). After triangulation and interpenetration test were applied, a sympatry matrix was created and the respective basal sympatric network was obtained (Fig. 6a). The partition index was significantly high ($PI = 0.\ 56$, P $\ll$ 0.01). NAM recognized four UCs after the removal of two intermediary species (Fig. 6a, b): UC1 composed by *E. walshi*, *E. cryptomphala*, *E. jujuyensis*, and *E. trigrammephora* is located in Salta and Jujuy Provinces in the Yungas and Chacoan areas. UC2 is composed by *E. parodizi*, *E. tucumanensis*, and *E. argentina*, inhabiting Tucumán province mainly in Yungas. UC3 composed by *E. variegata*, *E. hemiclausa*, and *E. hieronymi*, which inhabit Yungas and Pre-Puna areas from Catamarca to western Tucumán Provinces. UC4 composed by *E. trifasciata*, *E. puntana*, *E. guevarai*, and *E. trenquelleonis*, inhabiting Chacoan areas of northwestern Cordoba, southern Santiago del Estero, and northern San Luis Provinces (Fig. 6c). The two species removed due to their highest betweenness scores were *E. rhathymos* and *E. tomsici*. Five other species, *E. escoipensis*, *E. puella*, *E. quirogai*, *E. saltana*, and *E. villavilensis*, are isolated nodes, mainly because they consist of single localities records in the data set that are also geographically distant from each other. Each of the units identified are nonoverlapping candidates to areas of endemism supported by strictly endemic species. These units represent patterns obtained from the sympatry network structure, not from species range similarities.

Interestingly, NAM found that two geographically proximal species, *E. hieronymi* and *E. parodizi*, both coincident in western portion of Tucumán province but living at different altitudes, belong to different UCs. The same situation occurs with *E. variegata* and *E. tucumanensis*, both species inhabiting geographically close cloud forest areas (Yungas). UC4 and UC1 are geographically well separated from each other and also with UC2 and UC3 due to the existence of clear geographical gaps. However, UC2 and UC3 are contiguous, located between Tucumán and Catamarca provinces (Fig. 6c). *E. tucumanensis* + *E. argentina* form a monophyletic clade (Cuezzo, 2006), both are part of UC2 and according to the cladistic analysis these are sister species of *E. hemiclausa* and *E . variegata*, both components of UC3. This situation may indicate that both units of co-occurrence might have been originated by vicariance. However, this hypothesis needs further testing.

NAM detected two intermediary species: *E. rhathymos* and *E. tomsici*. *E. rhathymos* shows a remarkably wide area of distribution, ranging from Salta to Catamarca and Cordoba provinces. This species is highly variable in shell morphology and size among the different populations. Conversely, *E. tomsici* is only distributed in Tucumán and Catamarca provinces, but again morphological characters are highly variable between populations of Yungas and Chacoan areas. Molecular studies are needed in both cases to test if these taxa are really single species or group of species not detected by morphological characters.

Roig Juñent and Flores (2001) proposed two areas of endemism in the Chacoan biogeographical subregion in Argentina based on the analysis of beetles and vegetational formations: the "Llanos Chaqueños," an area encompassing Santiago del Estero, western Cordoba, eastern Catamarca, northern San Luis, La Rioja and San Juan Provinces, and the "Occidental Chaco" ranging from southern Bolivia and western Paraguay to northern Cordoba. Our analysis shows some coincidences with these previous subdivisions. However, the limits of these previously mentioned areas of endemism correspond with vegetational ecorregions. Areas of endemism or candidates of areas of endemism should not be delimited or proposed based on types of environments and vegetational formations. Our results are independent of predefined spatial units.

Case Study 2: The Curimatidae, a Characiform Family of Freshwater Fishes

Curimatidae occur in a broad range of freshwater Neotropical ecosystems. These habitats range from

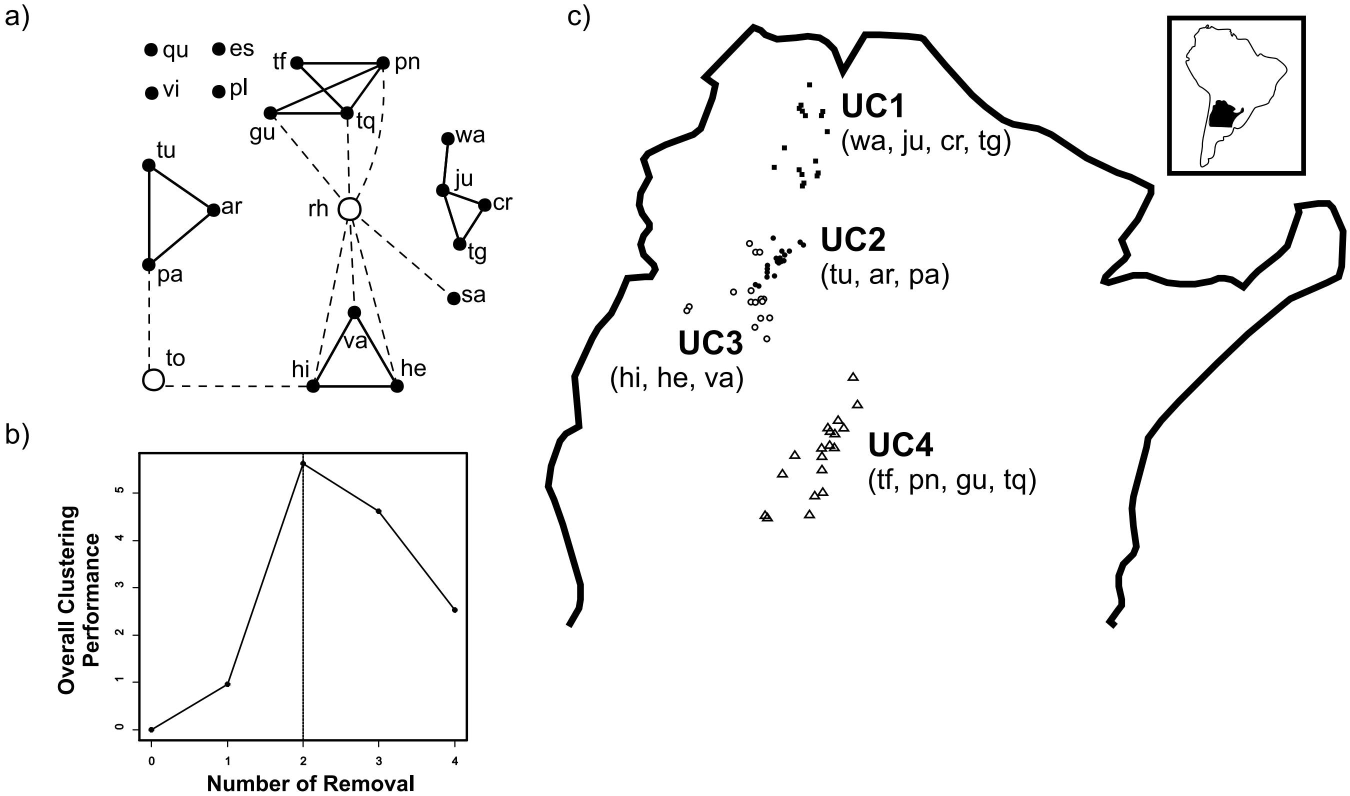


FIGURE 6. Inference and analysis of sympatry network applied on *Epiphragmophora* data. (a) Basal sympatry network. Intermediary species (empty circles) connected to network by dashed line. Elements of units of co-occurrence (UC) connected by full lines. (b) Overall clustering perfomance (*OCP*) along the revomal process. Vertical line indicates instance of removal where *OCP* is maximized (selected subnetwork). (c) Spatial expressions of the four UCs detected. Species codes: ar = *E. argentina*; cr = *E. cryptomphala*; es = *E. escoipensis*; gu = *E. guevarai*; he = *E. hemiclausa*; hi = *E. hieronymi*; ju = *E. jujuyensis*; pa = *E. parodizi*; pl = *E. puella*; pn = *E. puntana*; qu = *E. quirogai*; rh = *E. rhathymos*; sa = *E. saltana*; tf = *E. trifasciata*; tg = *E. trigrammephora*; to = *E. tomsici*; tq = *E. trenquelleonis*; tu = *E. tucumanensis*; va = *E. variegata*; vi = *E. villavilensis*; wa = *E. walshi*.

streams and meandering rivers typical of lowland floodplains to tributaries and rapids of the Andean piedmont and the upland of the Guyana and Brazilian Shields (Vari, 1988). The Curimatidae had been chosen to be analyzed using NAM because they are typical lowland components of the Neotropical fish fauna on both sides of the Andes. Members of this family inhabit the trans-Andean Pacific drainages of Central and South America from Costa Rica to Peru. They also inhabit the Caribbean drainages in northern South America, but the greatest species diversity occurs in the Atlantic drainages from the Orinoco basin through the Amazon basin and numerous rivers of the Guianas, Brazil, Uruguay, and Argentina.

Vari (1988) proposed nine areas of endemism based on the distribution of the Curimatidae: Western, Orinoco, Guianas, Northeast, São Francisco, Coastal, Upper Parana, Paraguay, and Amazon (Fig. 7a). However, he also considered the subdivision of Western area of endemism into six smaller nonoverlapping areas: Maracaibo, Río Magdalena, Atrato, Patia, Guayas, and Chira with few or unique endemic species in each of the river basins. The Curimatidae in the trans-Andean Pacific and Caribbean drainages are distinct from those of the east of the Cordilleras of the Andes, which acts as a barrier. The largest and most specious area of endemism recognized by Vari (1988) was the Río Amazonas. According to Vari (1988), as a consequence of the overlapping distribution patterns among those species, it is not possible to recognize areas of endemism within this vast geographic extension. The Amazon shares species with the three different areas: the Río Orinoco, the Guiana, and the Río Paraguay system. Vari stated that the largest number of species is shared with the Río Orinoco due to the broad Río Casiquiare connecting the two basins. Endemism among curimatid species is more pronounced in the two regions south of the Amazon basin, the upper Río Parana and the area consisting of Río Paraguay, lower Parana, Río Uruguay, and the river south of San Pablo.

A set of 1639 records corresponding to 98 species of Curimatidae was used. The records were compiled from two Internet databases, NEODAT Project (http://www.neodat.org/) and Fishbase (Froese and Pauly, 2007). Additional geographical data were taken from published taxonomic revisions (Vari 1984, 1988, 1989a, 1989b, 1991, 1992). The definitive list of records with their coordinates in decimal format is given in Appendix 3.

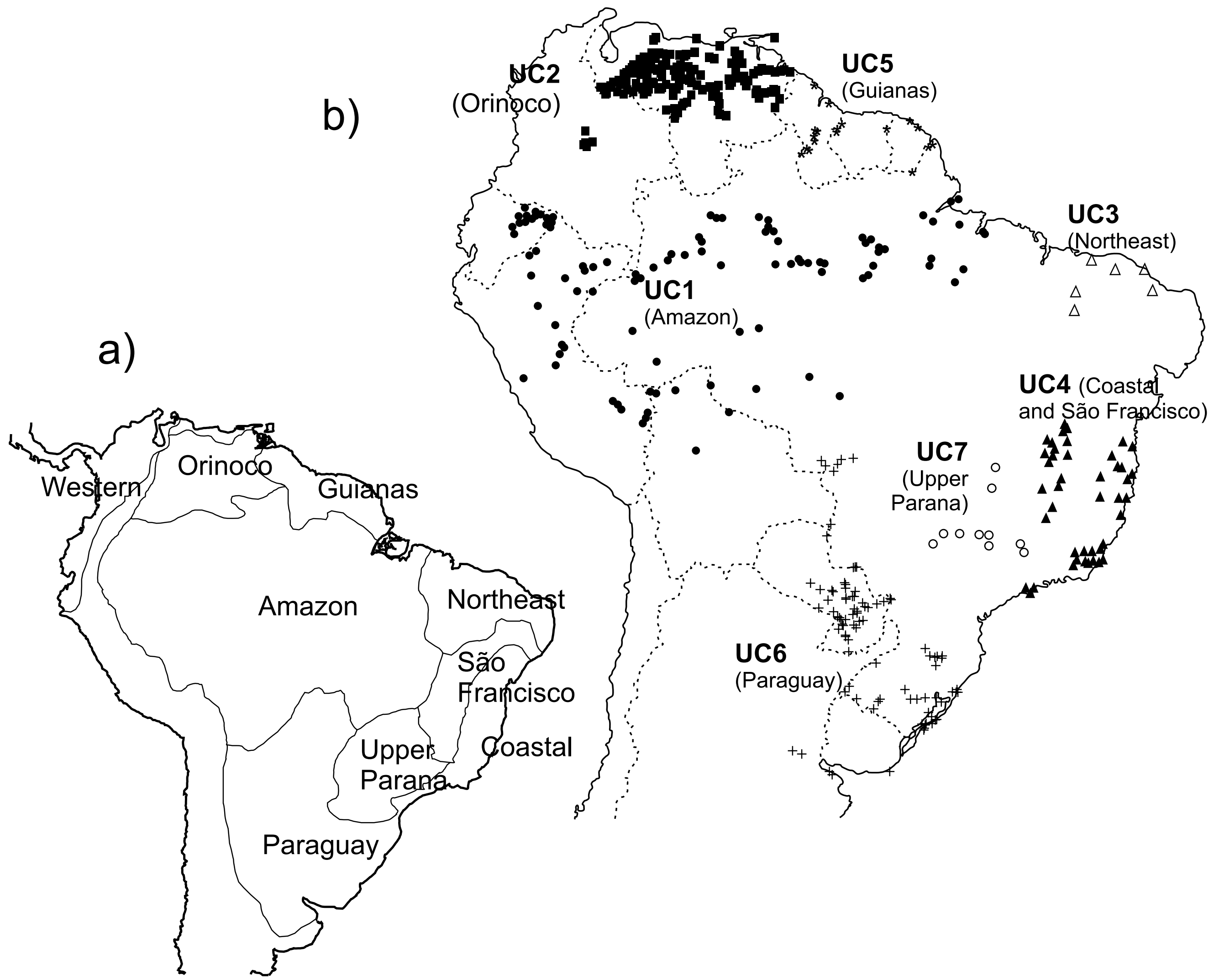


FIGURE 7. Inference and analysis of sympatry network applied on Curimatidae data. (a) Areas of endemism proposed by Vari (1988). (b) Spatial expressions of the seven units of co-occurrence (UC) detected (species listed in Table 2). For discussion see text.

NAM was applied to the sympatry matrix derived from punctual records. The partition index was significantly high ($PI = 0.74$, $P << 0.01$). Seven UCs were obtained (Table 2). Our results (Fig. 7b) show high coincidence with the areas delimited by Vari (1988). The following difference with Vari's results was found: São Francisco and coastal areas comprise a single UC instead of two areas of endemism, with three species endemic to this unit: *Steindachnerina elegans*, *Cyphocharax gilbert*, and *Curimatella lepidura*.

Pseudocurimata is the only generic-level clade of the Curimatidae endemic to the rivers of western slopes of the Andes. Vari (1988, 1989) included the species of this genus to justify the creation of the Western area of endemism. However, the species ranges within *Pseudocurimata* are restricted to single rivers. Such generic and specific levels of endemism of *Pseudocurimata* are noteworthy both in comparison to the more extensive ranges typical of most components of the South American ichthyofauna and relative to the much broader geographic distribution of all other curimatid genera (Vari, 1989a).

NAM did not associate the western species into a common UC, because the species ranges are restricted and not interpenetrated. Single endemic species that characterize some of Vari's subdivisions, as, for example, *Steindachnerina atratoensis*, endemic to the Atrato area of endemism, and *Pseudocurimata patiae*, endemic to the Patia, are isolated nodes in the network. Sympatry is not hypothesized in these cases due to the restricted ranges of single endemic species in single rivers that do not allow deriving co-occurrence patterns from these data. Other elements such as *Cyphocharax aspilos* and *Potamorhina laticeps* from Maracaibo constitute diads, entities not qualifying for UCs.

The largest UC found with NAM is Amazonas in concordance with the Amazon area of endemism established by Vari (1988, 1992), with 31 endemic species.

TABLE 2. Curimatidae units of co-occurrence (UC) detected by sympatry network analysis method (NAM).

Units of Co-Occurrence

UC1 (Amazon)
Curimata aspera
Curimata cisandina
Curimata inornata
Curimata knerii
Curimata ocellata
Curimata roseni
Curimata vittata
Curimatella meyeri
Curimatopsis microlepis
Cyphocharax gangamon
Cyphocharax gouldingi
Cyphocharax laticlavius
Cyphocharax leucostictus
Cyphocharax mestomyllon
Cyphocharax nigripinnis

Cyphocharax notatus
Cyphocharax pantostictos
Cyphocharax plumbeus
Cyphocharax spiluropsis
Cyphocharax stilbolepis
Cyphocharax vexillapinnus
Potamorhina latior
Potamorhina pristigaster
Psectrogaster amazonica
Psectrogaster falcata
Psectrogaster rutiloides
Steindachnerina fasciata
Steindachnerina hypostoma
Steindachnerina leucisca
Steindachnerina quasimodoi

UC2 (Orinoco)
Curimata cerasina
Curimata incompta
Cyphocharax meniscaprorus
Cyphocharax oenas
Steindachnerina argentea
Steindachnerina pupula
Steindachnerina dobula

UC3 (Northeast)
Curimata macrops
Psectrogaster rhomboides
Psectrogaster saguiru
Steindachnerina notonota

UC4 (Coastal and São Francisco)

Curimatella lepidura
Cyphocharax gilbert
Steindachnerina elegans

UC5 (Guianas)
Curimatopsis crypticus
Cyphocharax microcephalus
Steindachnerina runa

UC6 (Paraguay)
Curimatopsis myersi
Cyphocharax gillii
Cyphocharax platanus
Cyphocharax saladensis
Cyphocharax spilotus
Cyphocharax voga
Potamorhina squamoralevis
Steindachnerina biornata
Steindachnerina brevipinna
Steindachnerina conspersa

UC7 (Upper Parana)
Cyphocharax modestus
Cyphocharax nagelii
Cyphocharax vanderi

Steindachnerina insculpta

Paraguay area of endemism is also recognized by NAM with 10 endemic species characterizing this unit. The rest of the UCs found have 6 or less endemic species: Orinoco (6), Upper Parana (4), Northeast (4), Guiana (3), Coastal and São Francisco (3).

Nineteen species were intermediary and were eliminated from the analysis. From them, *Curimatella dorsalis* has the highest betweenness at first instance of removal. It exhibits a high morphological variability and a wide area of distribution ranging from Río Orinoco and Río Amazonas to the lower Parana system (Vari, 1992). Most of the intermediary species connecting Amazonas with Orinoco were mainly found in the Río Casiquiare, coinciding with Vari's results. Other species indicated by Vari as widely distributed and with the greatest variation in the body form (*Curimata cyprinoides* and *Psectrogaster essequibensis)* have been identified with NAM as intermediary species.

In comparison to Vari's (1988, 1992) results, the present study represents a testable methodology to analyze Curimatidae distribution. The spatial expressions of the UCs found using the available information are supported by exclusive endemic species. There was no need to define line limits of areas because the units arose from the sympatry analysis. Intermediary species that obscure the recognition of the units were identified to define the biogeographical pattern. We hypothesize then that the UCs found are candidates for areas of endemism until more information concerning other taxa is gathered to test this hypothesis.

Case Study 3: Sciobius *Schönherr, a Southern African Weevil*

The method was also implemented on a classic example, namely the distributional data of 47 *Sciobius* species published by Schoeman (1983) as dot maps. We have selected this example because it has been previously employed by many different authors (Morrone, 1994; Szumik et al., 2002; Hausdorf and Hennig, 2003; Mast and Nyfeller, 2003) in order to identify either areas of endemism or biotic elements. For comparison with their results, we used the same grid system. The sympatry matrix was obtained under the assumption that species co-occurring in at least one cell are sympatric. We also considered the original data (dot maps) for inferring sympatry. Then, we analyzed the sympatry matrices derived from different data inputs ($2^\circ \times 2^\circ$, $1^\circ \times 1^\circ$, and punctual data sets) with NAM. Table 3 summarizes the results of the three analyses.

We used the $2^\circ \times 2^\circ$ grid data matrix obtained by Morrone (1994) and corrected later by Mast and Nyfeller (2003). Nomenclatural conventions introduced by these authors are retained. The partition index of data was significantly high ($PI = 0.892$, $P \ll 0.01$). Four UCs were detected after removal of nine intermediary species. The respective spatial expressions of UCs correspond to a set of four nonoverlapping areas (Fig. 8a). We are able to compare and discuss our results with the other studies.

Parsimony analysis of endemicity (PAE; Morrone, 1994; Mast and Nyfeller, 2003): Original matrix (Morrone,

TABLE 3. Comparison of results from three different data inputs for *Sciobius*. 1–4 = units of co-occurrence; int = intermediary species; iso = isolated node. Boxes enclose partial matches. Note the high coincidence of species assignment among the three different partitions.

Sciobius species (abbreviations)	Data input		
	1°× 1°	2°× 2°	Dot maps
S. angustus (ag)	1	1	1
S. peringueyi (pe)	1	1	1
S. viduus (vd)	1	1	1
S. anriae (ai)	2	2	2
S. arrowi (ar)	2	2	2
S. barkeri (ba)	2	2	2
S. brevicollis (br)	2	2	2
S. cognatus (co)	2	2	2
S. cultratus (cu)	2	2	2
S. dealbatus (de)	2	2	2
S. holmi (hl)	2	2	2
S. marginatus (mg)	2	2	2
S. panzanus (pa)	2	2	2
S. pollinosus (pi)	2	2	2
S. prasinus (pr)	2	2	2
S. tenuicornis (te)	2	2	2
S. wahlbergi (wa)	2	2	2
S. asper (as)	3	3	3
S. capeneri (ca)	3	3	3
S. cinereus (ci)	3	3	3
S. griseus (gi)	3	3	3
S. minusculus (mi)	3	3	3
S. nanus (na)	3	3	3
S. oneili (on)	3	3	3
S. scapularis (sa)	3	3	3
S. schoenlandi (se)	3	3	3
S. tottus (to)	3	3	3
S. lateralis (la)	4	4	4
S. planipennis (pl)	4	4	4
S. pondo (pn)	4	4	4
S. scholtzi (sl)	4	4	4
S. aciculatifrons (ac)	int	int	int
S. granosus (go)	int	int	int
S. marshalli (ms)	int	int	int
S. obesus (ob)	int	int	int
S. pullus (pu)	int	int	int
S. endroedyi (en)	**4**	**4**	iso
S. transkeiensis (tr)	**4**	**4**	iso
S. granipennis (gp)	**4**	**4**	3
S. horni (ho)	**int**	**int**	2
S. kirsteni (ki)	**iso**	**iso**	2
S. vittatus (vt)	**1**	**1**	iso
S. viridis (vr)	1	**int**	**int**
S. impressicollis (im)	4	**int**	**int**
S. thompsoni (th)	4	**int**	**int**
S. spatulatus (sp)	int	**2**	**2**
S. bistrigicollis (bi)	int	**2**	**2**

1994) gave rise to three areas of endemism (N-O-R-S-T; P; I-J-L-M). After the analysis of corrected data, Mast and Nyfeller (2003) reduced the N-O-R-S-T candidate to a single cell (N). NAM also detected a fourth UC in B-C-E-F, which was not previously detected with PAE.

Biotic elements analysis (Hausdorf and Hennig, 2003): Four elements were produced, with geographical cores approximating our areas. However, unlike our areas, the biotic elements overlap geographically among each other and include in their list of supporting species clearly widespread taxa such as *S. pullus* (it belongs to element 4, occupying 48% of the cells under study and extending from Natal to Southern Cape Province) and *S. marshalli* (it belongs to element 2, traversing Natal and Transvaal regions from cell A to cell M). The inclusion of these widespread taxa is a consequence of the distance measure (Kulkczinsky distance) adopted by the authors. Kulkczinsky distance is justified because it relates the range overlap between species in a balanced way (Hausdorf and Hennig, 2003). However, in the same way, Kulkczinsky distance is misleading because it favors the inclusion of widespread species into biotic elements, clustering them with species of smaller ranges. Our method, due to the identification of intermediary species, recognized and removed these dispersal elements that obscure the patterns.

Stability in cluster analysis is strongly dependent on the data set (Hennig, 2007). Particularly, when the results of *Sciobius* example obtained with Prabclus package (Hausdorf and Hennig, 2003) were tested, we found that the same data matrix yield different results depending on row (species) ordering. Row permutation from a random order to an alphabetical order produced considerably different results. For this reason, caution should be taken when the biotic element methodology is applied.

The grid-based method (Szumik et al., 2002): It was difficult to compare to our results because optimal and suboptimal results are mixed together. However, there are sets of cells roughly equivalent to our results (Transvaal, Natal, Southern, and Eastern Cape regions). Our areas are supported by strictly endemic species, precluding records outside the areas. Furthermore, our areas are supported by unique species that are not supporting elements elsewhere, whereas different combinations of cells in Szumik et al. (2002) share many of their supporting species (e.g., sets 1, 4, 5, 6, 7, 8, 9). If different areas of endemism are proposed, each one must be supported by a unique or distinctive pool of supporting species as our method effectively does.

When using NAM, the shift from a 2°× 2° to a 1°× 1° grid did not critically affect the species composition of the UCs already found. Moreover, a better spatial resolution was achieved as a consequence of cell size reduction (Fig. 8b). Seventy pairs of sympatric species at the coarser scale became allopatric at the finer scale. Nevertheless, the general structure of the sympatry network was not greatly modified. Seventy-seven percent of deleted edges was incident to intermediary species previously detected, without major erosion of links inside UCs. The stability of our results contrasts with those derived from PAE and biotic element analysis, where 1°× 1° grid cell data resulted in lower resolution or increased noise, respectively. The analysis at this scale was not performed with the grid-based method, so we could not compare it with our results.

Finally, we inferred sympatry from punctual evidence (Fig. 8c1), to demonstrate that our approach has independence of the predefined areas. Schoeman (1983) used a quarter-degree dot for plotting records on the distribution maps. We assigned the geographical coordinates to the center of dots. A total of 311 coordinates were

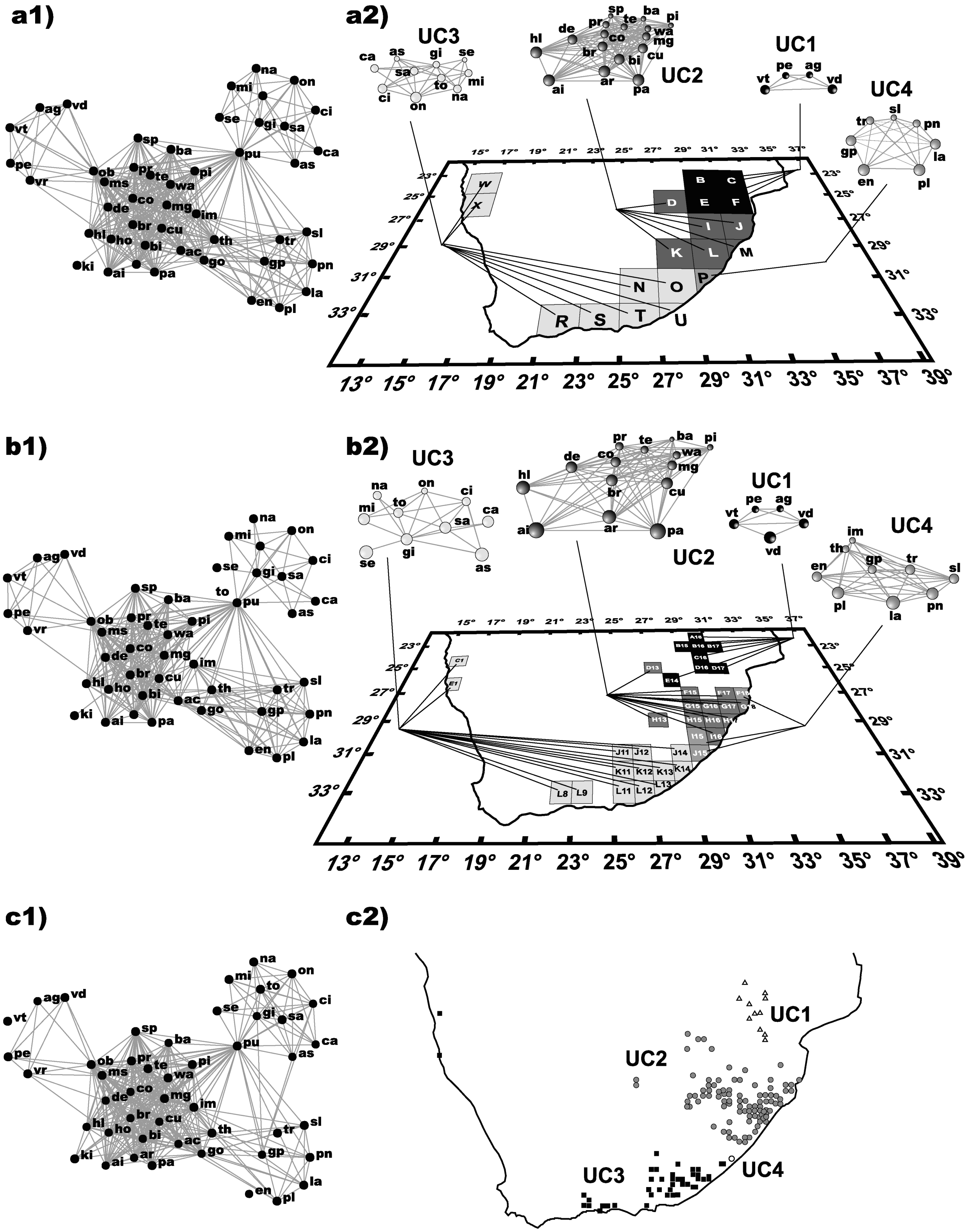

a1)
a2)
UC3
UC2
UC1
UC4
ca as gi se sa mi to na ci on
hl de pr sp te ba pi wa co mg br bi cu ai ar pa
vt pe ag vd
tr sl pn gp la en pl
B C D E F I J K L M N O P R S T U W X
15° 17° 19° 21° 23° 25° 27° 29° 31° 33° 35° 37°
23° 25° 27° 29° 31° 33°
13° 15° 17° 19° 21° 23° 25° 27° 29° 31° 33° 35° 37° 39°
b1)
b2)
A16 B15 B16 B17 C16 D13 D16 D17 E1 C1 E14 F15 F17 G15 G16 G17 H13 H15 H16 I15 I16 J11 J12 J14 J15 K11 K12 K13 K14 L8 L9 L11 L12 L13
c1)
c2)
na mi on to se gi ci sa ca pu as ag vd vt pe vr sp ba ob pr te pi ms wa co de mg im hl br cu ho bi th tr sl ki ai ar pa ac go gp pn la en pl

FIGURE 8. Inference and analysis of sympatry network applied on *Sciobius* (Coleoptera: Curculionidae) distribution data. (a) 2° grid data set. (b) 1° grid data set. (c) Punctual data set. (a1, b1, c1) Basal sympatry networks. Nodes (black dots) represent species, whereas edges (connecting lines) represent sympatry links between species. Network was constructed as an undirected graph using Netdraw 2.29 (Borgatti, 2002). (a2, b2, c2) Spatial expressions of units of co-occurrence (UC). (a2, b2) Each UC with corresponding components of selected subnetwork in same gray tone. (c2) Each UC with same symbol.

TABLE 4. Comparison of characteristics between NAM and the other discussed methods.

	Analysis performed			
Characteristics	PAE: Maximum parsimony of grid cells using presence/absence of taxa as characters with an outgroup with taxa absent in every cell (Rosen, 1988)	Grid-based method: Scoring of sets of cells depending on the adjustment of taxa to them (Szumik et al., 2002)	Biotic elements analysis: Partition of species into clusters according to their range similarity (Hausdorf, 2002)	NAM: Analysis of a sympatry matrix directed to extract species groups sympatrically cohesive by removal of intermediary species
Dependence on predefined areas	Yes	Yes	Yes	No
Emphasis on range similarities	No	No	Yes	No
Geographical overlap of supporting elements	No	Yes	Yes	No
Uniqueness of supporting elements	No	No	Yes	Yes
Curvature of earth into consideration	No	No	No	Yes
Evaluation of randomness in data structure	No	No	Yes	Yes
Relative stability of results to scale change	No	No	Medium	High

extracted and analyzed to obtain the sympatry matrix. The list of records with their coordinates in decimal format is given in Appendix 4. The partition index was also highly significant ($PI = 0.83$, $P \ll 0.01$). NAM obtained four UCs and one isolated node after removal of eight intermediary species. The spatial expressions of UCs (Fig. 8c2) are similar to those of our analyses based on 2° × 2° and 1° × 1° grid data matrices.

S. marshalli and *S. pullus* showed high connecting capacity in all the analyses, meaning high betweenness scores caused their early removal. This quantitative consideration correlates with the introduced and pest nature of these species, already remarked by Schoeman (1983). Table 4 lists the most conspicuous differences among the various approaches discussed.

DISCUSSION AND CONCLUSIONS

Traditionally, the theoretical concept of sympatry among species has been synthetically stated as "range overlap," leading first to the derivation of the species ranges, and later to the identification of sympatry. In the process of obtaining the ranges, either with grids or vectorial alternatives, the information provided by direct evidence (punctual records) is masked. On the contrary, we rely here on the direct evidences to analyze these spatial signals, their interactions of proximity and interpenetration, to infer sympatry.

There are some similarities between sympatry network analysis and the algorithm proposed by Girvan and Newman (2002) to detect communities in social networks. The main differences are as follows: (1) the Girvan and Newman algorithm is based on "edge betweenness," whereas ours is based on "node betweenness"; (2) the communities in the Girvan and Newman approach can be connected, whereas the UCs are mutually exclusive; (3) the Girvan and Newman method assigns all individuals to communities, whereas ours assigns some species to groups but others not (i.e., intermediary species, isolated nodes and diads).

In biogeography, Page (1987) applied graph concepts to formalize Croizat's ideas. His method, although using network analysis, operated on locality adjacency matrices derived from minimum spanning trees, whereas our method operates on species adjacency matrices. In this way, the resulting networks are different. In Page's method, the pattern of interest is the topological recurrence of locality graphs projected on a map, whereas in ours the goal is to detect groups of species cohesively sympatric.

The identification of UCs and the derivation of their spatial expressions liberate us from the need of the traditional delimitation of areas of endemism. These spatial expressions are the candidates to areas of endemism, until biogeographical analysis is performed to test their historical significance. In this way, taxa from cladograms should be replaced with the defined UCs to obtain area cladograms. The congruence between area cladograms from different lineages will serve to test the nonrandom co-occurrence of the taxa inhabiting the units and their possible common biogeographical history.

The notion of intermediary species offered in this paper is new, and it is only captured with a holistic approach like network analysis. Several circumstances can lead to intermediacy. The most common is widespread distributions. However, intermediary species do not always possess this attribute. For example, in the *Sciobius* analysis with the 2° × 2° data set, some intermediary species, such as *S. granosus* and *S. thompsoni*, have a narrow range of two cells. From this case, it becomes clear that species are intermediary due to their position in the network, not necessarily because of the size of their range. Additional causes of intermediary condition may include the lack of either taxonomic resolution or speciation after a vicariant event.

We have considered sympatry as a binary relation; but when there is partial range overlap, it is possible to use weights to quantify the degree of superposition. Newman (2004) proposed a method for analyzing weighted networks, mapping them onto multigraphs. Although promising, this development transcends the scope of the present paper. Certainly, our approach is susceptible to improvements, but we consider it as a viable option to other methods overly dependent on a priori delimited spatial units.

The new methods presented here represent a novel platform of analysis in biogeography and can be differentiated from other methods by the following characteristics: (1) sympatry inference generates sympatry hypotheses from direct evidence, as a previous necessary step in the search of biogeographical patterns; (2) sympatry hypotheses are proposed from punctual data, allowing the consideration of nonconvex ranges; (3) NAM emphasizes the treatment of sympatry (a relational or network datum) with the appropriate analytical tool (network analysis); (4) NAM may explore ranges obtained through different methods, being versatile and applicable to available data; (5) NAM deals with distributions that obscure subjacent patterns, identifying intermediary species and facilitating heuristic tasks with them; (6) NAM explores randomness in the data to be analyzed without appeal to stochastic generation of ranges; (7) NAM yields results strictly adjusted to a notion of endemism in the sense of species restricted to an area; (8) NAM is independent of comparative criteria among ranges as congruence or similarity; and (9) NAM shows higher stability of results despite scale change.

Acknowledgments

We thank J. M. Carpenter, A. García Muñiz, R. M. Hughes, J. L. Molina, C. Molineri, and C. Nieto for helpful comments; H. Barber-James and A. S. Schoeman for bibliographical support; A. Mast for providing his original data matrix; C. Hennig and B. Hausdorf for clarifying some problems detected in Prabclus package; and CONICET (PIP 2563/01, PIP 6048) and ANPCyT (PICT 01-12529) from Argentina for financial support. Editors J. Sullivan and A. Paterson and three anonymous reviewers made very helpful critique and stimulated us to formalize our idea of sympatry inference based on punctual data.

References

Anderson, S. 1994. Area and endemism. Quart. Rev. Biol. 69:451–471.

Bhattacharya, P., and M. L. Gavrilova. 2006. CRYSTAL—A new density-based fast and efficient clustering algorithm. Pages 102–111 *in* 3rd International Symposium on Voronoi Diagrams and Engineering (ISVD'06), July 2–5, 2006, Calgary, Canada.

Borgatti, S. P. 2002. NetDraw: Graph visualization software. Harvard, MA: Analytic Technologies.

Borgatti, S. P., M. G. Everett, and L. C. Freeman. 2002. Ucinet 6 for Windows. Harvard, MA: Analytic Technologies.

Brandes, U. 2001. A faster algorithm for betweenness centrality. J. Math. Sociol. 25:163–177.

Croizat, L. 1964. Space, time, form: The biological synthesis. Caracas, Venezuela. Published by the author.

Cuezzo, M. G. 2006. Systematic revision and cladistic analysis of *Epiphragmophora* Doering from Argentina and Southern Bolivia (Gastropoda: Stylommatophora: Xanthonychidae). Malacologia 49:121–188.

de Berg, M., van Kreveld, M., Overmars, M., and O. Schwarzkopf (eds.). 2000. Computational Geometry, 2nd edition. Springer, Berlin.

El-Rabbany, A. 2002. Introduction to GPS: The Global Positioning System. Artech House mobile communications series. Artech House, Boston, Massachusetts.

Freeman, L. C. 1977. A set of measures of centrality based on betweenness. Sociometry 40:35–41.

Froese, R., and D. Pauly (eds.). 2007. FishBase. Available at http://www.fishbase.org/ (accessed September 2007).

Girvan, M., and M. E. J. Newman. 2002. Community structure in social and biological networks. Proc. Natl. Acad. Sci. USA 99:8271–8276.

Hanneman, R. A., and M. Riddle. 2005. Introduction to social network methods. University of California Press, Riverside, California.

Harold, A. S., and R. D. Mooi. 1994. Areas of endemism: Definition and recognition criteria. Syst. Biol. 43:261–266.

Hausdorf, B. 2002. Units in biogeography. Syst. Biol. 51:648–652.

Hausdorf, B., and C. Hennig. 2003. Biotic element analysis in biogeography. Syst. Biol. 52:717–723.

Hennig, C. 2007. Cluster-wise assessment of cluster stability. Comp. Stat. Data Anal. 52:258–271.

Hennig, C., and B. Hausdorf. 2004. Distance-based parametric bootstrap tests for clustering of species ranges. Comput. Stat. Data Anal. 45:875–896.

Hennig, C., and B. Hausdorf. 2006. A robust distance coefficient between distribution areas incorporating geographic distances. Syst. Biol. 55:170–175.

Humphries, C. J., and L. Parenti. 1999. Cladistic biogeography: Interpreting patterns of plant and animal distributions. Oxford University Press, New York.

IUCN. 1994. IUCN Red List Categories. International Union for the Conservation of Nature, Species Survival Commission, Gland, Switzerland.

Linder, H. P. 2001. On areas of endemism, with an example from the African Restionaceae. Syst. Biol. 50:892–912.

Mast, A. R., and R. Nyffeler. 2003. Using a null model to recognize significant co-occurrence prior to identifying candidate areas of endemism. Syst. Biol. 52:271–280.

Montoya, J. M., S. L. Pimm, and R. V. Sole. 2006. Ecological networks and their fragility. Nature 442:259–264.

Moore, J. E., and R. K. Swihart. 2007. Toward ecologically explicit null models of nestedness. Oecologia 152:763–777.

Morrone, J. J. 1994. On the identification of areas of endemism. Syst. Biol. 43:438–441.

Nelson, G., and N. I. Platnick. 1981. Systematics and biogeography: Cladistics and vicariance. Columbia University Press, New York.

NEODAT PROJECT (Inter-Institutional Database of Fish Biodiversity in the Neotropics). Available at http://www.neodat.org/ (accessed September 2007).

Newman, M. E. J. 2001. Scientific collaboration networks. II. Shortest paths, weighted networks and centrality. Phys. Rev. E64:1–7.

Newman, M. E. J. 2004. Analysis of weighted networks. Phys. Rev. E70:1–9.

Page, R. D. M. 1987. Graphs and generalized tracks: Quantifying Croizat's panbiogeography. Syst. Zool. 36:1–17.

Papari, G., and N. Petkov. 2005. Algorithm that mimics Human perceptual grouping of dot patterns. Lect. Notes Comput. Sci. 3704:497–506.

Platnick, N. I. 1991. On areas of endemism. Aust. Syst. Bot. 4:xi–xii.

Posadas, P. 1996. Distributional patterns of vascular plants in Tierra del Fuego: A study applying parsimony analysis of endemism (PAE). Biogeographica 72:161–177.

Proulx, S. R., D. E. L. Promislow, and P. C. Phillips. 2005. Network thinking in ecology and evolution. Trends Ecol. Evol. 20:345–353.

Rapoport, E. H. 1982. Areography: Geographical strategies of species. Pergamon Press, Oxford, UK.

Rathert, D., D. White, J. C. Sifneos, and R.M. Hughes. 1999. Environmental correlates of species richness for native freshwater fish in Oregon, U.S.A. J. Biogeogr. 26:257–273.

Roig-Juñent, S., and G. Flores. 2001. Historia Biogeográfica de las areas áridas de América del Sur austral. Pages 257–266 *in* Introducción a la biogeografía en Latinoamerica: Teorías, conceptos, métodos y aplicaciones (J. J. Morrone, and J. Llorente Busquets, eds.). Las Prensas de Ciencia, Facultad de Ciencias, UNAM, México D.F.

Rosen, B. R. 1988. From fossils to earth history: Applied historical biogeography. Pages 437–481 *in* Analytical biogeography: An integrated approach to the study of animal and plant distribution (A. Myers and P. Giller, eds.). Chapman and Hall, London.

Schoeman, A. S. 1983. Revision of the genus *Sciobius* Schönherr (Coleoptera: Curculionidae). Entomol. Mem. Dep. Agric. Repub. S. Afr. 59:1–50.

Szumik, C. A., F. Cuezzo, P. A. Goloboff, and A. E. Chalup. 2002. An optimality criterion to determine areas of endemism. Syst. Biol. 51:806–816.

Szumik, C., and P. A. Goloboff. 2004. Areas of endemism: An improved optimality criterion. Syst. Biol. 53:968–977.

Unmack, P. J. 2001. Biogeography of Australian freshwater fishes. J. Biogeogr. 28:1053–1089.

Vari, R. P. 1984. Systematics of the Neotropical Characiform Genus *Potamorhina* (Pisces: Characiformes). Smith. Cont. Zool. 400:1–36.

Vari, R. P. 1988. The Curimatidae, a lowland Neotropical fish family (Pisces: Characiformes): Distribution, endemism, and phylogenetic biogeography. Pages 343–377 *in* Proceedings on the Workshop on Neotropical Distribution Patterns (W. R. Heyer and P. E. Vanzolini, eds.). Acad. Bras. Cienc., Río de Janeiro.

Vari, R. P. 1989a. Systematics of the neotropical characiform genus *Pseudocurimata* Fernández-Yepes (Pisces: Ostariophysi). Smith. Cont. Zool. 490:1–28.

Vari, R. P. 1989b. Systematics of the Neotropical characiform genus *Curimata* Bosc (Pisces: Characiformes). Smith. Cont. Zool. 474:1–61.

Vari, R. P. 1991. Systematics of the neotropical characiform genus *Steindachnerina* Fowler (Pisces: Ostariophysi). Smith. Cont. Zool. 507:1–116.

Vari, R. P. 1992. Systematics of the neotropical characiform genus *Curimatella* Eigenmann and Eigenmann (Pisces: Ostariophysi) with summary comments on the Curimatidae. Smith. Cont. Zool. 533:1–48.

Watts, D. J., and S. H. Strogatz. 1998. Collective dynamics of "small-world" networks. Nature 393:440–442.

First submitted 13 November 2007; reviews returned 8 February 2008; final acceptance 7 April 2008
Associate Editor: Adrian Paterson

Appendix 1: Formalization of Algorithm

Sympatry inference and NAM are structured around the following algorithm, implemented in the program package SyNet, which is an add-on package for the statistical software R (available at http://www.cran.r-project.org).

1. Is distributional evidence based on punctual data? If yes, go to 2; else go to 3.
2. Obtain the sympatry matrix applying the analysis of proximity and interpenetration of points. Go to 6.
3. If data are available as raster-like ranges, go to 4. If data are available as vector-like ranges, go to 5.
4. Obtain the sympatry matrix applying co-occurrence of species. Go to 6.
5. Obtain the sympatry matrix applying geometric analysis of range overlap. Go to 6.
6. Test the significance of *PI*. If test yields a significant index, then go to step 7; else interrupt analysis.
7. Initialize *OCP* to zero. Calculate betweenness value for each species of original network.
8. Remove species with highest betweenness value.
9. For resulting subnetwork, calculate *OCP*. Recalculate betweenness value for each remnant node.
10. Are betweenness scores of all remnant nodes equal zero? If no, go to 8; else next.
11. From original network to last subnetwork, select the instance of removal where *OCP* is maximized. In case of *OCP* ties, select the latter instance.
12. Extract from 11 the UCs (components with three or more nodes) and map them.