Portugaliae Acta Biol. **19:** 9-19. Lisboa, 2000

‘BIOSYSTEMATICS’ IN THE 90S: INTEGRATING DATA FROM DIFFERENT SOURCES

Gonzalo Nieto Feliner & Inés Álvarez Fernández
Real Jardín Botánico, CSIC
Plaza de Murillo 2. 28014 Madrid (Spain)

Nieto Feliner, G. & Álvarez Fernández, I. (2000). ‘Biosystematics’ in the 90s: integrating data from different sources. *Portugaliae Acta Biol.* **19:** 9-19.

The term biosystematics has always eluded a precise widely accepted definition. Since its coining in the 40s, it has encompassed different approaches and objectives to the study of living organisms. The use of experimental methods, the focus on population level studies, the interest in evolutionary processes, and the emphasis on different sources of evidence to assess relationships have been frequently emphasised. In the 90s, the vast amount of information that new molecular techniques have made available has brought about a new potential problem: the discrepancies that arise between different data sets from the same organisms. Handling those discrepancies to produce an integrated reliable estimate of phylogenetic relationships is perceived by the authors as one the greatest challenges of Systematics in the 90s. To give an idea of how common and relevant those discrepancies are, two real cases recently faced by a research team (ours) are briefly described. Incongruence in the first case (*Doronicum)* appears to be due to high levels of homoplasy in the morphological data, and seems to be better managed by following a total evidence approach. In the second example (*Armeria*), discrepancies are the result of hybridisation and concerted evolution affecting multicopy sequences in the ribosomal DNA.

Key words: Biosystematics, molecular data, multiple data sets, total evidence, reticulation, *Doronicum*, *Armeria*.

Nieto Feliner, G. & Álvarez Fernández, I. (2000). ‘Biosistemática’ en la década de los 90: integración de datos de diferentes fuentes. *Portugaliae Acta Biol*. **19**: 9-19.

El término biosistemática nunca ha tenido una definición precisa y ampliamente aceptada. Desde que fuera acuñado en los años cuarenta, ha sido utilizado para referirse al estudio de organismos vivos bajo enfoques y objetivos muy distintos aunque no excluyentes. Algunos de los enfoques más comunes han sido el uso de métodos experimentales, el énfasis en estudios a nivel poblacional, el interés por procesos evolutivos y el esfuerzo por buscar datos procedentes de distintas fuentes de evidencia para estimar relaciones de parentesco. En los años noventa, la gran cantidad de datos, disponibles gracias a las nuevas técnicas moleculares, han traído un nuevo problema potencial: las discrepancias que pueden aparecer entre distintas matrices de datos de los mismos organismos. Los autores consideran que la gestión de dichas discrepancias para producir una hipótesis integrada y fiable de relaciones filogenéticas es uno de los principales retos de la Sistemática en la actualidad. Para dar una idea de lo comunes que este tipo de problemas puede ser, se describen brevemente dos casos surgidos en un grupo de investigación cualquiera (el nuestro). En el primero de ellos (*Doronicum),* las discrepancias parecen deberse a la cantidad de homoplasia que contienen los caracteres morfológicos y se sugiere que el enfoque de 'evidencia total' es el más adecuado. En el segundo caso (*Armeria*), las discrepancias son el resultado de hibridación y evolución concertada que afecta a las regiones multicopia del ADN ribosomal.

Palabras clave: Biosistemática, datos moleculares, matrices de datos múltiples, 'evidencia total', reticulación, *Doronicum*, *Armeria*.

INTRODUCTION

The topic of this paper occurred to us when asked to give a talk within a session on methods in biosystematics. Whether explicitly named or not, most botanical meetings we have attended included a session on biosystematics. The purpose here was to try to accommodate the discipline in today's systematic practice with the perspective of at least the last two decades. We comment on the wide use of the term biosystematics since its coining and stress the potential problems derived from what systematist always do or recommend when morphological evidence is inconclusive, i.e., search for new characters from other sources, in particular molecular sources. Such potential problems — the integration of heterogeneous data — are nowadays common to practising systematists. As an example of how relevant those kinds of problems are, two cases from our own research where discrepancies between different sources of data have been detected are briefly described.

MEANING OF BIOSYSTEMATICS

The term 'Biosystematics was introduced by CAMP & GILLY (1943): "Seeks (1) to determine the natural biotic units and (2) to apply to these units a system of nomenclature adequate to the task of conveying precise information regarding their defined limits, relationships, variability and dynamic structure". Such definition has been effective enough so as to launch a tremendously successful movement within the Systematic community over the second half of this century. Its success, which can be measured by the number of symposia, books, papers, and even organisations involved, is due to several causes including the ample demand for enlarging the objectives and improving the results of an already old scientific discipline.

Probably it could be agreed that the objectives were sufficiently stated. However, the term biosystematics has always eluded a precise, widely accepted definition. Probably one of the reasons is that to accomplish the pursued aims there were several possible ways. What researchers from outside and inside of the taxonomic community demanded in the 40s and after was that scientific advancements related to the classification of organisms be introduced into the taxonomic practice. Those new scientific advancements came from several disciplines, such as genetics, ecology, and physiology, as well as from sub-disciplines within them. As a consequence, many researchers working on rather different things, although all oriented towards a more or less precise goal claimed they were all doing biosystematics. Therefore, it is not surprising that there have been difficulties to define precisely the term. As an example, while introducing one symposium on biosystematics in Montreal David D. Keck said "I am not sure that it can be taken for granted that we all define biosystematics the same way" (*Regnum Veg*. 27: 7. 1963).

Along these five decades, there have been attempts to restrict the use of the term. For instance STACE (1980) criticised misapplications and denounced that "the term biosystematics has been widened by some taxonomists to cover virtually any taxonomic activity not pursued in the herbarium and almost any acquired technique". As a whole however a diffuse usage of the term has been rather the rule than the exception. In fact, a common practice has been to consider that "the concept is employed for taxonomic work on living material" (HEDBERG 1997).

In sum, we can safely say that the term has been applied to virtually any study related to systematics departing from what we describe nowadays as alpha-taxonomy. In this sense, divergence from traditional taxonomy has been a main objective. And if we looked at the ways studies have implemented such departure, we will have some of the approaches to biosystematics cultivated in the last four or five decades. Four of the most common approaches have been the use of experimental methods, the focus on population level studies, the interest in evolutionary processes, and the emphasis on different sources of evidence to assess relationships, kariology occupying a significant role among those sources. Those four approaches are in no way exclusive. Many studies have more or less extensively focused on all of them. Here, we want to stress on the consequences of placing the last one in the current scene, where a wide variety of molecular techniques providing large amounts of data are available.

As you know, in systematics we try to identify homologies that can group organisms. For most systematists, these homologies reflect a single phylogenetic history that we aim to recover. It has been a recurrent idea in the history of Systematics that the larger the data sampled the better the estimate of relationships. The numerical taxonomists and its precursor Adanson defended this idea as applied to morphological data (HULL 1988). But systematists using molecular data are also increasingly adopting it. Molecular data are apt to fulfil our, sometimes desperate, search for characters to produce more natural classifications and more sound estimates of phylogenetic relationships. The reasons are well known. There are a potentially enormous number of characters and most of the variation contained in this kind of characters is inherited. But this source of data is not devoid of problems for systematists. Large amounts of information may be difficult to handle even if they contain true phylogenetic signal. The fact is that information from molecular sources is not always consistent with morphological information, and indeed there are good reasons for this to happen. We thus need to handle the discrepancies that arise commonly between different data sets from the same organisms. This problem has been addressed already by several authors but it is by no means solved (EERNISSE & KLUGE 1993; PATTERSON *et al.* 1993; BULL *et al.* 1993; WIENS 1998; WENDEL & DOYLE 1998; JOHSON & SOLTIS 1998).

We consider that one of the greatest challenges in Systematics is to integrate heterogeneous data into a single reliable estimate of phylogenetic relationships within a group of organisms, and that such is also the challenge of the ‘biosystematics’ of the 90s. But, how difficult is to integrate heterogeneous data or how frequent are problems derived from this attempt? To address such question, as a example, we present two cases recently faced by our research group. Both of them involve discrepancies between morphological and molecular data for the same organisms and both have arisen during the last three years in systematic or evolutionary oriented studies.

DORONICUM (ASTERACEAE, SENECIONEAE)

The first example is based on the PhD dissertation of one of us (I.A.) *Doronicum* comprises around 30 species distributed in Europe, Asia and Northern Africa. There is only a comprehensive taxonomic review for this genus, by CAVILLIER (1907, 1911). We have gathered information potentially useful for phylogeny reconstruction, and thus for classification purposes, mainly from three sources of evidence. These are morphology [12 parsimony informative characters], sequences from the internal transcribed spacers of the nuclear ribosomal DNA (ITS) [117 parsimony informative characters], and sequences from another spacer region (*trn*L – *trn*F) in the chloroplastic DNA [19 parsimony informative characters]. Details of the sampling, methods and results will be given elsewhere. Here, we will only refer to aspects related with congruence between data sets.

Comparisons of the topologies of the cladograms resulting from the independent analyses of the three data sets do not provide a clear agreement between them. We thus need to follow one of the available approaches to handling more that one data set for the same organisms. The two opposite views are the *total evidence* (character congruence or simultaneous analysis - EERNISSE & KLUGE 1993) and the *consensus approach* (taxonomic congruence or partitioned analysis – MIYAMOTO & FITCH 1995). The former advocates combining all data from the same organisms into a single matrix. The latter recommends separate analyses of different data sets from the same organisms and a comparison of the resulting trees for common clades.

However, because the rightness of the decision depends critically on the specific data, JOHSON & SOLTIS (1998), formalising BULL *et al.* (1993) approach, recommend a third approach, the conditional combination. This consists in merging the data in a single matrix if and only if there is evidence that the various matrices are not heterogeneous, i.e., that they do not represent different branching histories or that they have not been affected by different evolutionary mechanisms, for instance reticulation. This is achieved by going through a series of tests that assess if the two of more data sets for the same organisms contain serious incongruence.

The first stage is to determine if the topologies of the parsimony trees resulting from the different data sets are similar enough so as to consider that their differences are due to sampling error. We have computed two of these *topological congruence indices* for *Doronicum*: the Partition Metric (PM) and the Greatest Agreement Subtree Metric (D_1). The first simply measures the rearrangements needed to transform one of the two trees we compare into the other. The second measures the number of taxa we have to prune in two trees to arrive at a minimum topology in which the two trees agree.

However, because what these indices compare is the topologies of the most parsimonious trees (that is, the best summaries of the data) but not the data sets themselves, it is necessary to go through a second stage. This consists of assessing how much conflicting phylogenetic information exists between the two data sets, by using *character congruence indices*. For *Doronicum*, we have computed two of them: the incongruence metric of Mickevich and Farris (I_{MF}) and the incongruence metric of Miyamoto (I_M). Both of them try to estimate the strength of support within each data set for relationships other than those implied by the most parsimonious solution in terms of extra homoplasy needed.

These indices provide a quantitative measure of incongruence but we do not know how large they need to be to indicate a serious conflict between data sets. For this reason, a significance test for heterogeneity that assesses the null hypothesis of homogeneity between data sets can be computed. We have computed the Partition Homogeneity Test (HT_F) of FARRIS *et al.* (1995), which uses the I_{MF} as a distance measure. In Doronicum, this statistical test confirms what the former indices suggested, i.e., that ITS and *trn*L-F are homogeneous. On the contrary, it rejects the null hypothesis in the other two comparisons. In other words, morphological data is discordant with the two molecular data sets. This implies that the two molecular data sets can be combined into a single matrix, and the results (topologies), compared with those from the independent analysis of the morphological data.

When this conditional combination approach is followed, and we compared the strict consensus of the molecular trees with the morphological trees, we arrived at the striking conclusion that there are no common clades. Normally, we tend to think that when molecular trees and morphological trees disagree, it is more likely that morphological ones are a better representation of the species phylogeny. This is due to mechanisms potentially affecting genes such as lineage sorting, gene duplication and introgression (DOYLE 1992). In the *Doronicum* data, however, it seems that such is not the case. First, because despite the fact that ITS is inherited biparentally and *trn*L-F is inherited maternally, both data sets are homogeneous. That the branching histories of the two data sets agree suggests that they are a good reflection of the species branching pattern. Besides, the parameters resulting from the parsimony analysis of the morphological data set are clearly worse than those from the combined molecular data set. The CI is similar (0.62) even though the number of informative characters is almost twelve times greater in the molecular data, and the bootstrap support for most groups is lower in the morphological trees as is the resolution. Therefore, although morphological characters are few, they contain a considerable amount of incongruence among them.

In sum, we have to make a decision as to what is the best estimate of phylogenetic relationships with the available data. If we followed the approach recommended by JOHNSON & SOLTIS (1998), since we have no common clades recovered from the molecular and morphological data sets and in view of the apparently low information in the morphological data, we would probably have to rely exclusively on the molecular trees. But this implies entirely discarding a set of empirical data, a criticism that was made in the last decade to advocators of the character compatibility analysis (FARRIS 1983). Alternatively, we may at this point choose the total evidence approach, based on the argument that the most stringent test for a set of characters is to analyse it together with other characters to see which patterns are reinforced and which are questioned by congruence. The results of such test are read on the resulting cladograms generated from the combined matrix, following the total evidence approach. What we find is a confirmation that half of the morphological characters are highly homoplasious. In particular, the fact homocarpy requires five independent gains (7 steps) indicates that it fails to pass a test of homology by congruence with the rest of the characters.

This first example in *Doronicum* illustrates a situation where discrepancies between data sets are not strongly supported by the respective data sets and where a total evidence approach is preferable.

ARMERIA (PLUMBAGINACEAE)

The second example differs from the first in various ways. We do not have an explicit phylogeny based on morphological characters. The reason is that morphological data in the genus are not suitable for a cladistic analysis. A high percentage of the characters that distinguish species are continuous, and their ranges overlap largely so that even if some type of gap-coding was applied, the matrix would be very inappropriate. Besides, high levels of intraspecific variability require that a large number of characters be coded as polymorphic or missing. Such a pattern of morphological variation is most likely due to hybridisation and introgression. Therefore, in this example we have a likely cause for discrepancies between different sources of data.

Over the last ten years, we have been conducting research on *Armeria* and learned a number of things about the biology of this genus, part of them discovered thanks to an experimental crossing program (NIETO FELINER 1990, 1997; NIETO FELINER *et al.* 1996). 1) Internal reproductive barriers are weak; 2) Most populations of *Armeria* are diploid obligate outcrossers due to a self-incompatibility system (BAKER 1966); 3) sympatric situations among congeners are frequent. 4) Morphological patterns of variation, detectable doing herbarium work, suggest the occurrence of extensive hybridisation. 5) Some morphometric characters are reliable markers for introgression because they have clear additive effects and thus are able to reflect intermediacy.

In the last three years we have been trying to find molecular evidence to document cases of hybridisation that we had previously hypothesised based on other data, mainly morphological. One of them was the presumed hybridity of *Armeria villosa* subsp. *carratracensis* (NIETO FELINER *et al.* 1996). After some trials using RAPDs and restriction site data from amplified chloroplastic regions, the molecular marker selected was the ITS regions. Because of the strong possibility of gene flow, the sampling covered not only different populations of the presumed parents (*A. villosa* subsp. *longiaristata* and *A. colorata*), but also other taxa occurring in geographically proximal locations. In a first study, we sampled 55 specimens from 33 taxa, and obtained the ITS sequences directly from PCR products (FUERTES *et al.* 1999b). We analysed them with parsimony and rooted the trees using *Psylliostachys suworowii* as outgroup following a phylogeny of Plumbaginaceae (LLEDÓ *et al.* 1998).

In the strict consensus from 15 fundamental trees based on 18 informative characters, we can see that the discrepancy consists on the splitting of a single species, and even a single subspecies, in different clades (Fig. 1). Therefore, this gene tree conflicts with the taxonomic arrangement fundamentally based on morphology, ecology and distribution. Specifically, different sequences (12) from the same subspecies (*A. villosa* subsp. *longiaristata)* appear in three of the five major clades, and four of the five major clades contain accessions of one or more of the six subspecies recognised within *A. villosa*.

The geographic structure of the data is the first clue to throw some light on the discrepancy (Fig. 1). Each outlined area coincides with one of the major clades in the consensus. So, the composition of the ITS clades depends on the geographical origin of the samples rather than on their systematic placement. Such geographical structure supports an interpretation of this pattern in terms of gene flow between different species, which we know is possible, based on previous evidence. This implies that shared nucleotide positions supporting the major clades are due, in some of the samples, to extensive gene flow rather than to common ancestry. Other alternative but rather unlikely interpretations have been discussed elsewhere (FUERTES *et al.* 1999b).

There is some further supporting data for the interpretation of the discrepancy in terms of gene flow. Lack of internal resolution of major clades in the strict consensus (Fig. 1) is not due to different alternative topologies in the 15 fundamental cladograms. It is simply due to the fact that sequences within an area, be they from the same species or not, are almost identical. ITS sequences have been studied in many angiosperms even in cases of reticulation. And we know that additive polymorphisms are usually detected in those cases, i.e., two nucleotides for the same site, in those sites in which the two hybridising species differ (e.g. SANG *et al.* 1995). So the question is: is it still sound our interpretation for the discrepancy between ITS pattern and taxonomic arrangement even if additive polymorphic sites are really scarce in our study?

There is a mechanism responsible for the homogenisation of multicopy sequences such as those contained in the ribosomal DNA where the ITS are. This mechanism, known as *concerted evolution*, takes place not only within individuals but within species or, more precisely, within reproductive groups. And such mechanism can be very active. If it were active enough, the sharing of a sequence could be due not only to common ancestry but also to gene flowfollowed by active homogenisation. The fact is that we have found evidence for this rapid homogenization in experimental F_2 hybrids of *Armeria* (FUERTES *et al.* 1999a). Therefore, the above question can be answered affirmatively. The sampling of this study has been thoroughly enlarged both taxonomically and geographically and the geographical structure holds (unpubl.).

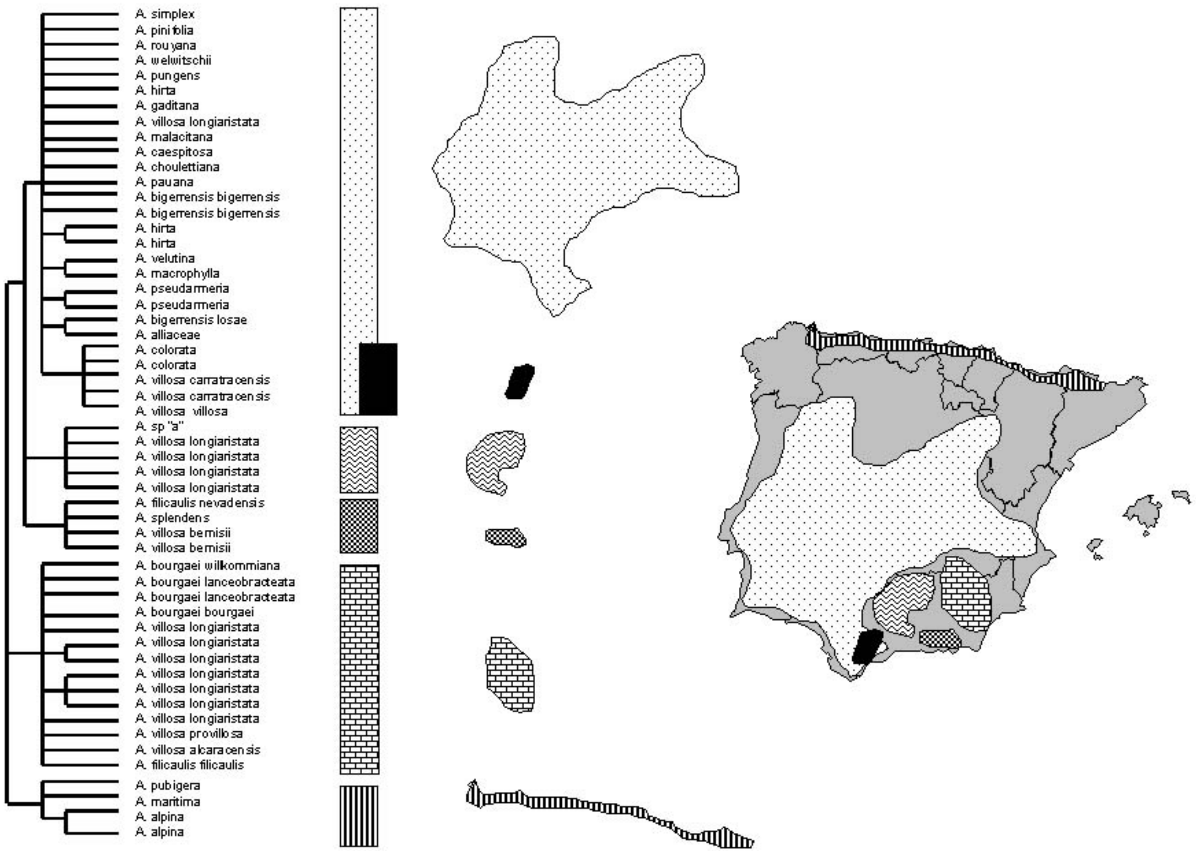

Fig. 1.- Strict consensus tree from 15 fundamental trees in *Armeria* based on ITS1+5.8S+ITS2 data (L=103; C.I.=0.84; R.I.=0.96). Patterns indicate the correspondence between each one of the major clades and the geographical area where samples falling in the clade belong.

Clarification of the discrepancy is possible in this second example, through inputting independent experimental evidence, something that is very clearly in the original spirit of biosystematics.

CONCLUSION

May we conclude anything at all from these two very different examples? Let us refer to a modern paradox related to the topic of our paper and specifically to the use of molecular data: While congruence among patterns remains the strongest argument for a single true explanation (common ancestry), discrepancies do also constitute an avenue for biological insights. In other words, while consistent groupings of species revealed by independent data sets is the best evidence for common ancestry, the discrepancies may be telling something relevant about the organisms that is not simply noise (WENDEL & DOYLE 1998).

ACKNOWLEDGEMENTS

This work has been supported by grants DGES PB97-1146 and DGES PB96-0849 of the Spanish Dirección General de Enseñanza Superior e Investigación Científica. The authors thank J. A. Rosselló and J. Fuertes Aguilar with whom the molecular study on *Armeria* is being conducted, and P. Vargas for helpful comments.

REFERENCES

BAKER, H. G. (1966). The evolution, functioning and breakdown of heteromorphic incompatibility systems, I. The Plumbaginaceae. *Evolution* 20: 349-368.

BULL, J. J., J. P. HUELSENBECK, C. W. CUNNINGHAM, D. L. SWOFFORD & P. J. WADDELL (1993). Partiotioning and combining data in phylogenetic analysis. *Syst. Biol.* 42: 384-397.

CAMP, W. H. & C. L. GILLY (1943). The structure and origin of species. *Brittonia* 4: 323-385.

CAVILLIER, F. (1907). Étude sur les *Doronicum* a fruits homomorphes. *Annuaire Conserv. Jard. Bot. Genève* 10: 177-251.

CAVILLIER, F. (1911). Nouvelles études sur le genre *Doronicum*. *Annuaire Conserv. Jard. Bot. Genève* 13-14: 195-368.

DOYLE, J. J. (1992). Gene trees and species trees: molecular systematics as one-character taxonomy. *Syst. Bot.* 17: 144-163.

EERNISSE, D. & A. G. KLUGE (1993). Taxonomic congruence versus total evidence, and amniote phylogeny inferred from fossils, molecules, and morphology. *Mol. Biol. Evol.* 10: 1170-1195.

FARRIS, J. S. (1983). The logical basis of phylogenetic analysis. In: Platnick & Funk (eds.), *Advances in Cladistics* vol. 2 pp. 7-36.

FARRIS, J. S., M. KÄLLERSJO, A. G. KLUGE & C. BULT (1995). Testing significance of incongruence. *Cladistics* 10: 315-319.

FUERTES AGUILAR, J., J.A. ROSSELLÓ & G. NIETO FELINER (1999a). Nuclear ribosomal DNA (nrDNA) concerted evolution in natural and artificial hybrids of *Armeria* (Plumbaginaceae). *Mol. Ecol.* 8: 1341-1346.

FUERTES AGUILAR, J., J.A. ROSSELLÓ & G. NIETO FELINER (1999b). Molecular evidence for the compilospecies model of reticulate evolution in *Armeria* (Plumbaginaceae). *Syst. Biol.* 48 (in press).

HEDBERG, O. (1997). Progress in biosystematics: an overview. *Lagascalia* 19: 307-316.

HULL, D. L. (1988). Science as a progress. Univ. of Chicago Press, Chicago. xiii + 586.

JOHNSON, L. A. & D. E. SOLTIS (1998). Assessing congruence: empirical examples from molecular data. In: D. E. Soltis, P. S. Soltis & J. J. Doyle (eds.), Molecular Systematics of Plants II. DNA sequencing, pp. 297-348. Kluwer Academic Publishers, Boston.

LLEDÓ, M. D., M. B. CRESPO, K. M. CAMERON, M. F. FAY & M. W. CHASE (1998). Systematics of Plumbaginaceae based upon cladistic analysis of rbcL sequence data. *Syst. Bot.* 23: 21-29.

MIYAMOTO, M. M. & W. M. FITCH (1995). Testing species phylogenies and phylogenetic methods with congruence. Syst. Biol. 44: 64-76.

NIETO FELINER, G. (1990). Armeria. In S. CASTROVIEJO *et al.* (eds.), Flora Iberica 2, pp. 642-721. CSIC, Madrid.

NIETO FELINER, G. (1997). Natural and experimental hybridization in Armeria (Plumbaginaceae): A. salmantica. *Int. J. Pl. Sci.* 158: 585-592.

NIETO FELINER, G., A. IZUZQUIZA & A.R. LANSAC (1996). Natural and experimental hybridization in *Armeria (Plumbaginaceae): A. villosa* subsp. *carratracensis. Pl. Syst. Evol.* 201: 163-177.

PATTERSON, C., D. M. WILLIAMS & C. J. HUMPHRIES (1993). Congruence between molecular and morphological phylogenies. *Ann. Rev. Ecol. Syst.* 24: 153-188.

SANG, T., D.J. CRAWFORD & T.F. STUESSY (1995). Documentation of reticulate evolution in peonies (*Paeonia*) using internal transcribed spacer sequences of nuclear ribosomal DNA: implications for biogeography and concerted evolution. *Proceedings of the National Academy of Sciences of the USA* 92: 6813-6817.

STACE, C. A. (1980). Plant taxonomy and biosystematics. Edward Arnold, London. viii + 279 pp.

WENDEL, J. F. & J. J. DOYLE (1998). Phylogenetic incongruence: Window into genome history and molecular evolution. In: D. E. Soltis, P. S. Soltis & J. J. Doyle (eds.), Molecular Systematics of Plants II. DNA sequencing, pp. 265-296. Kluwer Academic Publishers, Boston.

WIENS, J. J. (1998). Combining data sets with different phylogenetic histories. *Syst. Biol.* 47: 568-581.