PERGAMON

Biochemical Systematics and Ecology 29 (2001) 1–20

biochemical
systematics
and ecology

www.elsevier.com/locate/biochemsyseco

Character coding of secondary chemical variation for use in phylogenetic analyses

Todd J. Barkman*

Department of Biology, Pennsylvania State University, University Park, PA 16802, USA

Received 11 November 1999; accepted 24 January 2000

Abstract

A coding procedure is presented for secondary chemical data whereby putative biogenetic pathways are coded as phylogenetic characters with enzymatic conversions between compounds representing the corresponding character states. A character state tree or stepmatrix allows direct representation of the secondary chemical biogenetic pathway and avoids problems of non-independence associated with coding schemes that score presence/absence of individual compounds. Stepmatrices are the most biosynthetically realistic character definitions because individual and population level polymorphisms can be scored, reticulate enzymatic conversions within pathways may be represented, and down-weighting of pathway loss versus gain is possible. The stepmatrix approach unifies analyses of secondary chemicals, allozymes, and developmental characters because the biological unity of the pathway, locus, or character ontogeny is preserved. Empirical investigation of the stepmatrix and character state tree coding methods using floral fragrance data in *Cypripedium* (Orchidaceae) resulted in cladistic relationships which were largely congruent with those suggested from recent DNA and allozyme studies. This character coding methodology provides an effective means for including secondary compound data in total evidence studies. Furthermore, ancestral state reconstructions provide a phylogenetic context within which biochemical pathway evolution may be studied. © 2000 Elsevier Science Ltd. All rights reserved.

Keywords: *Cypripedium*; Orchidaceae; Secondary chemicals; Benzenoids; Chemosystematics; Cladistics; Character coding; Stepmatrix; Character-state tree

* Tel.: + 814-863-6413; fax: + 814-865-9131.
E-mail address: tjb20@psu.edu (T.J. Barkman).

PII: S0305-1978(00)00031-4

1. Introduction

Systematists often consider multiple sources of data in order to estimate evolutionary relationships of taxa for investigating character evolution, postulating biogeographic events, or constructing classifications. Most studies consider information from morphology and molecules even if the data are not ultimately combined in one analysis. Currently, although many methods exist, parsimony analyses of morphology and DNA sequence information are most commonly used for estimating phylogenies. In contrast, although vast amounts of secondary compound information became available concomitantly with the development of cladistic methodologies, few studies have incorporated it into cladistically based systematic studies. Frustration regarding the lack of cladistic analysis of micromolecular data was expressed by Jeffrey (1995); "Unfortunately, this recommendation (cladistic analysis of chemical data) has been far from universally followed, with the result that the taxonomic input of micromolecular data has been disproportionately and disappointingly small in comparison with its quantity (pp. 7–8)."

1.1. Previous approaches treating micromolecular data

Methods of analysis for secondary chemical data are diverse, including both non-cladistic and cladistic approaches that either consider variation within a biogenetic context or ignore it. Perhaps the most thorough cladistic treatment of secondary chemical variation was that of Seaman and Funk (1983) in which an additive binary coding method was used to score the presence or absence of a pathway and each compound within the pathway. A slight modification of their methodology was subsequently used only once (Culberson, 1986), in spite of the potential generality for all classes of secondary compounds. In some chemosystematic studies, compounds, classes of compounds, or substitution features were simply coded as present or absent (Bolick, 1983; Humphries and Richardson, 1980; Richardson, 1983; Nandi et al., 1998) and analyzed using parsimony algorithms. One cladistic modification of the flavonoid scoring system (Bate-Smith and Richens, 1973), produced branching diagrams based on the a priori decision of chemical character polarity with the level of compound oxidation determining the phylogenetic position of a taxon (Richardson and Young, 1982). A non-cladistic but related approach using minimum biosynthetic-step indices was pathway-based and assigned a distance measure to each pair of taxa sampled (Levy, 1977). Distance measures were also used by Figueiredo et al. (1995) to quantify the evolutionary advancement parameters related to oxidative levels and skeletal specializations of diterpenes (Gottlieb, 1989). Finally, some researchers have assessed secondary compound variation after a phylogeny has been estimated using other data (Miao et al., 1995; Plunkett et al., 1996). In these cases, chemical data were interpreted as congruent with patterns derived from DNA variation; however, no numerical analyses were performed, making such assertions difficult to interpret. While several of the methods mentioned above have advantages, most are not generalizable and suffer from problems of non-independent character definitions.

1.2. Non-independence of secondary compound variation

The most important challenge to using secondary chemical variation in phylogenetic studies is coding statistically independent characters — a requirement of any numerical analysis (Sneath and Sokal, 1973). The biogenetic relatedness of many compounds makes them statistically non-independent. As an example, consider two compounds differing only by a methylation substitution (i.e., eugenol and methyleugenol). These two compounds are likely produced via identical pathways, although the action of a methyl transferase converts eugenol to the substituted form (Wang and Pichersky, 1998). In this case, the production of one compound, methyleugenol, is present only because of the prior formation of eugenol. Due to the expected covariation of these two compounds, coding each as a separate character would violate the assumption of independence.

Another concern about secondary chemical variation is that homoplasy is widespread and therefore problematic for cladistic analyses (Richardson, 1983). This concern is well-founded especially since Bohlmann et al. (1998) have found independent evolution of the enzymes producing limonene in conifers and mints. While homoplasy of terpenoid production may exist among taxa as diverse as angiosperms, gymnosperms, algae, and fungi, it does not necessarily preclude the use of secondary chemical variation at lower taxonomic levels. Homology is ultimately determined by congruence of a particular character with others as determined by phylogenetic analysis (Patterson, 1988).

1.3. A method for character coding

A general coding methodology is presented below in an attempt to stimulate more widespread use of secondary compound data in phylogenetic analyses. The coding scheme represents independent biogenetic pathways as characters and their related enzymatic conversions as character states. The pathways are directly represented as multistate characters with either a character state tree or stepmatrix transformational definition, both of which are standard user-defined characters in PAUP (Swofford, 1991) or MacClade (Maddison and Maddison, 1992). The result is a set of statistically independent characters that are comprised of putatively homologous character states. No attempt is made to include quantitative variation due to the many non-phylogenetic factors that affect it (Harborne and Turner, 1984). This character definition loosely equates a biochemical pathway with an evolutionary pathway whereby biogenetic transformations are potential phylogenetic markers. This view, that the character of evolution is the biochemical pathway, is supported by the prediction that enzymes catalyzing steps later in a pathway evolved from enzymes performing reactions earlier in the pathway (Sacchettini and Poulter, 1997; see Wang and Pichersky, 1999 for an empirical example). An implication of this paradigm is that novel enzymatic steps evolve only if integration within the context of the biochemical pathway has occurred.

2. Materials and methods

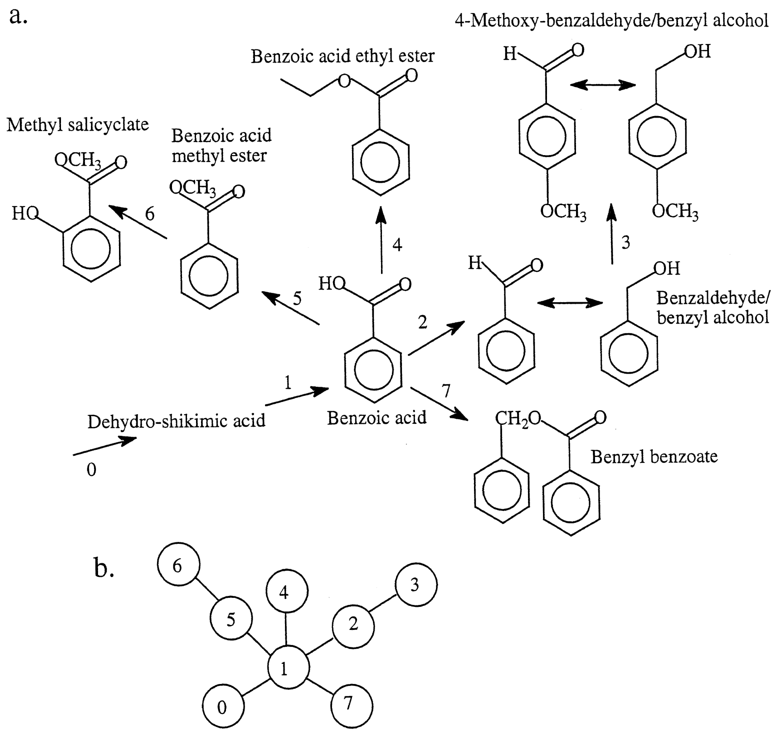
The coding procedure is outlined below with several problematic examples discussed. Two assumptions are involved in this coding procedure. First, it is assumed that the biogenetic pathways used for delimiting characters are correct. This is often not a limitation, because abundant experimental information exists for many secondary compounds. Second, it is assumed by analogy that the experimentally elucidated biogenetic pathways are the same in all study groups. Even if a presumed pathway does not have extensive experimental support for the specific taxa studied, the explicit statement of such, at the onset of a cladistic analysis, allows other researchers to evaluate the soundness of such assumptions and investigate alternatives. It is not clear, however, if violations of these assumptions could compromise results obtained in phylogenetic analyses.

Floral fragrance compound variation in *Cypripedium* (Barkman et al., 1997) was used to demonstrate the coding procedure. This data set contained several classes of compounds including terpenoids, phenyl propanoids, and fatty acid derivatives. The biogenetic pathways for these groups of compounds were elucidated from experimental studies including Schreier (1984) for all classes of volatiles, Gross (1981) for phenolic acids and Wheeler and Croteau (1986), Croteau and Karp (1991), Gray (1987), and Gambliel and Croteau (1984) for terpenoids.

2.1. Character coding

The first step required to code secondary chemical data is to identify a biogenetic pathway within which each compound was produced. For each assumed pathway, compounds may be arranged relative to assumed or sampled precursors based upon probable enzymatic conversions between them. The *Cypripedium* data were delimited into 8 characters (pathways) with numbers of character states (enzymatic conversions) ranging from 2 to 7 as shown in Appendix A. Assumed biogenetic pathways used to define the phenylalanine and benzoic acid characters are illustrated in Figs. 1 and 2. Assumed pathways used to define all other characters are shown in Appendices B and C or are available from the author upon request. The enzymatic reactions were ordered with respect to the common precursor of the entire pathway and represented directly by a character state tree or stepmatrix.

A character state tree defines a path, identical to the biochemical pathway, by which character state transformations occur. Fig. 1a shows an assumed biogenetic pathway for benzoic acid derivatives. The corresponding character state tree, defined in Fig. 1b, shows that various evolutionary transformations can occur only within the constraints of the defined path. For example, consider the steps required to lose production of 4-methoxy benzaldehyde and gain production of benzyl benzoate within the context of the character state tree. First, loss of 4-methoxy benzaldehyde must occur, then loss of production of benzaldehyde followed by a gain of the production of benzyl benzoate. In this case, a direct interconversion from 4-methoxy benzaldehyde to benzyl benzoate is not definable. By contrast, a stepmatrix



c.

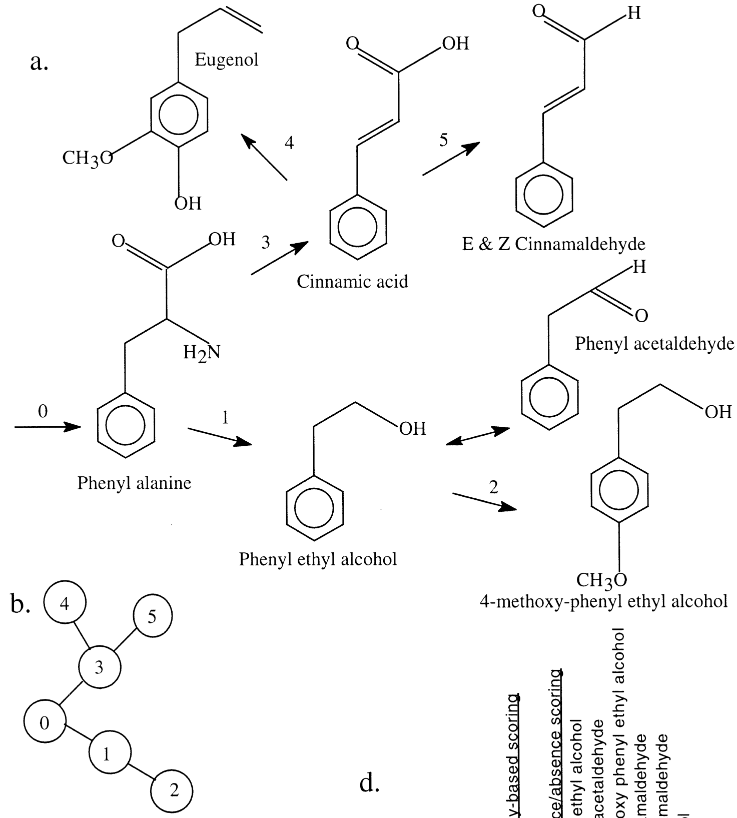
TO:	0	1	2	3	4	5	6	7
FROM: 0	-	1	2	3	2	2	3	2
1	1	-	1	2	1	1	2	1
2	2	1	-	1	2	2	3	2
3	3	2	1	-	3	3	4	3
4	2	1	2	3	-	2	3	2
5	2	1	2	3	2	-	1	2
6	3	2	3	4	3	1	-	3
7	2	1	2	3	2	2	3	-

d.

TO:	0	1	2	3	4	5	6	7
FROM: 0	-	1	2	3	2	2	3	2
1	1	-	1	2	1	1	2	1
2	1	1	-	1	2	2	3	2
3	1	1	1	-	2	2	3	2
4	1	1	2	3	-	2	3	2
5	1	1	2	3	2	-	1	2
6	1	1	2	3	2	1	-	2
7	1	1	2	3	2	2	3	-

Fig. 1. (a) Assumed biogenetic pathway for compounds derived from benzoic acid. Numbers above arrows represent enzymatic conversions to be coded in cladistic analyses for character 2 (benzoic acid, Appendix). (b) Character state tree representing transitions within character 2. (c) Symmetric stepmatrix with equal costs assigned to transitions within character 2. (d) Asymmetric stepmatrix with differential costs assigned to each transformation within character 2.

transformational definition allows all possible interconversions, with relative weighting representing the biogenetic complexity required for such changes. Rationale for assigning relative weights will be discussed more below. Stepmatrix definitions of the benzoic acid pathway in Fig. 1a are shown in Fig. 1c and d. Fig. 3 illustrates the difference between a character state tree and stepmatrix with respect to transformational constraint.



c.

TO:	0	1	2	3	4	5
FROM: **0**	-	1	2	1	2	2
1	1	-	1	2	3	3
2	1	1	-	2	3	3
3	1	2	3	-	1	1
4	1	2	3	1	-	2
5	1	2	3	1	2	-

d.

	Pathway-based scoring	Presence/absence scoring	Phenyl ethyl alcohol	Phenylacetaldehyde	4-methoxy phenyl ethyl alcohol	E-cinnamaldehyde	Z-cinnamaldehyde	Eugenol
C. parviflorum	0		0	0	0	0	0	0
C. pubescens	2		1	1	1	0	0	0
C. candidum	1		1	1	0	0	0	0
C. arietinum	0		0	0	0	0	0	0
C. kentuckiense	1		1	0	0	0	0	0
C. acaule	0		0	0	0	0	0	0
C. reginae	0		0	0	0	0	0	0
C. guttatum	4		0	0	0	0	0	1
C. macranthum	5		0	0	0	1	1	0

Fig. 2. (a) Assumed biogenetic pathway for compounds derived from phenylalanine. Numbers above arrows represent enzymatic conversions to be coded in cladistic analyses for character 1 (phenylalanine, Appendix). (b) Character state tree representing transitions within character 1. (c) Asymmetric stepmatrix representing transitions within character 1. (d) Data matrix contrasting coding of same chemical data using the pathway-based approach advocated here and a simple presence/absence scoring.

2.1.1. Problems of non-independence

Fig. 2a shows the assumed pathway for compounds derived from phenylalanine. A single enzymatic reaction was coded for the production of the stereoisomers E- and Z-cinnamaldehyde via reduction from cinnamic acid. The evolution of the ability to

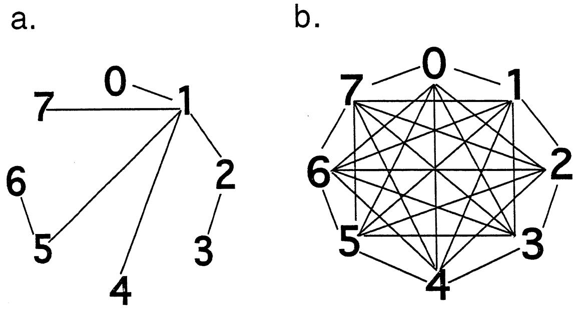


Fig. 3. Comparison of transformational constraint between (a) a character-state tree and (b) a stepmatrix. Note that only a subset of the stepmatrix transformations are defined for a character state tree. Variable weights may be applied to each transformation depending on relative biogenetic complexity in a stepmatrix whereas equal weights must be applied to all transformations in a character-state tree.

produce E-cinnamaldehyde did not likely occur independently of the ability to produce Z-cinnamaldehyde, and therefore the coding is justified. Indeed, experimental evidence has demonstrated that some stereoisomers may be produced by a single enzymatic reaction (Gambliel and Croteau, 1984; Bohlmann et al., 1997). *Cypripedium macranthum* Sw., therefore, was coded as possessing character state 5 for the phenylalanine character rather than the two non-independent characters (E- and Z-cinnamaldehyde) assigned in the compound-based coding method shown in Fig. 2d.

The phenylalanine pathway provides another example of biogenetically linked compounds and how they are coded in a pathway-based procedure (Fig. 2, Appendix). Consider, *Cypripedium pubescens* Willd. and *C. candidum* Muhl. ex Willd. These two taxa both produce phenyl ethyl alcohol and phenyl acetaldehyde. However, *C. pubescens* also produces the derivative, 4-methoxy-phenyl ethyl alcohol, which obviously relies on the presence of the precursor, phenyl ethyl alcohol, for its biogenesis. In this case, *C. candidum* was coded as possessing character state 1, while *C. pubescens*, was scored for state 2. This method avoids coding two non-independent characters for *C. pubescens* (both compounds) while preserving the homology of the character states shared by the two taxa. In this case it is obvious that the 4-methoxy derivative could not have been produced without the earlier enzymatic step represented by 1. Fig. 2d shows a comparison of the coding procedure for the phenylalanine character (1) using the method described here and one possibility under a compound-based approach whereby all compounds found within the pathway are scored as present/absent. An important point is to notice the reduction in total character number from 6 to 1 for the group of compounds assumed to be produced via the phenylalanine pathway.

For cases in which knowledge of biogenesis is uncertain or lacking, one may assume the compounds are independently synthesized and code them as independent binary characters or use *ad hoc* criteria in order to justify a specific coding procedure. One such criterion might be based upon the common presence of several putatively biogenetically related compounds within a species or the repeated occurrence of each in several species. For instance, in Fig. 1a the presence/production of benzyl alcohol and benzaldehyde is treated as a single character state due to the fact that both were always recovered together in all taxa sampled. It does not seem useful to code this enzymatic conversion as a character state because of the ubiquity of the transformation across the species sampled. In some cases, no pathways were found in the literature for compounds such as 1,4-dimethoxy benzene, and therefore these were treated as independent presence/absence characters.

2.1.2. Central importance of experimental evidence

Experimentally determined pathways are important for assessing homology of characters. One case in particular concerned the terpenoid class of compounds which were delimited as mono- or sesquiterpenes in *Cypripedium*. The mono- and sesquiterpene pathways are largely independent of each other because plants produce both in different pathways and subcellular compartments (Bohlmann et al., 1998). Further divisions within these two classes of terpenoids are difficult to make, however, as they are all derived via GPP for monoterpenes, or farnesyl pyrophosphate (FPP) for sesquiterpenes. In the analyses utilizing a stepmatrix, all monoterpenes were considered as a single character. Analyses using character state trees assigned monoterpene compounds to two pathways (hydrocarbons and alcohols) although they are all derived from GPP. This decision was made because several taxa produced both monoterpene hydrocarbons and alcohols and polymorphic taxa cannot be analyzed when using a character state tree. It should be noted that the term "polymorphic" in this case applies to an individual rather than a population that exhibits more than one character state. In cases where extensive pathway polymorphism is observed for character state tree definitions (individual or population level), a possible alternative may be to simply code each subpathway or enzymatic reaction separately in spite of some degree of non-independence among them.

2.1.3. Relative weighting of stepmatrices

The representation of character state transformations by a stepmatrix requires the assignment of relative weights to all possible interconversions. A symmetric stepmatrix assigns equal weight to the gain or loss of a particular transformation, whereas an asymmetric stepmatrix has different weights applied to the gain or loss of a particular transformation. If certain transformations are biologically impossible, they could be assigned an infinite weight so that the direct evolutionary transformation would not be allowed. A discussion of the presumed events involved in the evolutionary gain or loss of compound production provides rationale for the relative weights assigned to the stepmatrices used in this paper.

An evolutionary gain is more complicated than a loss on a physiological and molecular level. Consider, *Clarkia breweri*, the best studied example of an evolutionary gain of fragrance production (Pichersky et al., 1994; Raguso and Pichersky, 1995; Wang and Pichersky, 1998; Dudareva et al., 1998; Wang et al., 1997; Wang and Pichersky, 1999; Ross et al., 1999; Raguso and Pichersky, 1999). In this species, the initial steps of fragrance evolution required gene duplication and divergence. In the case of isoeugenol *O*-methyl transferase (IEMT), only 7 amino acid substitutions were required to cause a diverged function from the precursor, caffeic acid *O*-methyl transferase (COMT) (Wang and Pichersky, 1999). Transcriptional and translational regulatory integration into a tissue specific biochemical pathway must have then occurred (Wang et al., 1997; Nam et al., 1999). Finally, evolution of osmophores might occur, but there is evidence that this may not be necessary, at least in *Clarkia breweri* (Raguso and Pichersky, 1995). This rather complicated set of events is probably required for the evolution of a novel scent compound within any species; however, some differences might exist. For example, Cseke et al. (1998) found that linalool synthase arose not as a duplicated gene, but by recombination between two different pre-existing terpene synthases.

In contrast to scent gain, the evolutionary loss of scent production is less complicated on a physiological and molecular level and may have multiple causes. Experimental studies have shown that absence of scent in otherwise scent producing species or their close relatives may occur due to complete gene absence (Yuba et al., 1996), gene presence but lack of transcription (Wang et al., 1997), and gene presence with transcription but lack of enzyme function due to point mutations of critical amino acids (Dudareva, pers. comm.). Apparent loss of scent production could also be due to lack of precursor formation or inability to secrete scent compounds. It is evident that multiple causes of scent loss exist, and it would be difficult to develop differential weights for the point mutations that cause lack of gene transcription versus those that cause lack of appropriate enzymatic activity. The important point is that scent loss could occur within a single generation of an individual with one or few mutational changes, whereas scent gain would likely take many generations of mutation subject to selection. For this reason, one justifiable stepmatrix weighting would assign less weight to losses than to gains. For all asymmetric stepmatrices analyzed below, single-step losses were coded for entire pathways, regardless of their length, based on the assumption that loss of an early enzymatic step in any pathway would result in loss of all subsequent steps. This model of loss of pathway expression assigns low weights to pathway loss, whereas the symmetric stepmatrix model assigns higher weights that are equal for pathway loss or gain. Below, symmetrically and asymmetrically weighted stepmatrices will be compared in phylogenetic analyses.

Fig. 1a shows the benzenoid pathway with the corresponding user-defined character state tree representing it in Fig. 1b, symmetric stepmatrix in Fig. 1c, and asymmetric stepmatrix in Fig. 1d. In this case, the use of an ordered character state tree and symmetric stepmatrix resulted in the same number of steps for gains and losses of compound production. As an alternative, an asymmetric stepmatrix was used to allow single-step losses of pathway expression with multistep gains for alternative branches of a particular pathway. This feature is evident upon inspection of the asymmetric stepmatrix of Fig. 1d. As an example, consider the number of steps required to lose enzymatic conversion 3 and to gain conversion 4. A single-step loss of the pathway leading to conversion 3 with a single-step gain of conversion 4 results in 2 steps total. This is one step less than would be specified by the character state tree or symmetric stepmatrix. Alternative weights could be applied to pathway losses that would range between zero and the number of steps required for pathway gain. Although a single-step model is assumed in the analyses performed below, alternative weights may be preferable when other models of pathway evolution are assumed. The advantages and disadvantages of the asymmetric stepmatrix character state transformations defined here will be discussed below.

The Barkman et al. (1997) data set was chosen because independent phylogenetic estimates exist for *Cypripedium*, which serve as bases of comparison for the results obtained in this paper. For *Cypripedium*, a reference cladogram based on 5s rDNA sequence variation (Cox, 1995) and a phenogram derived from genetic distances estimated from allozyme variation (Case, 1994) were used. The reference cladogram of Cox (1995) was modified by pruning taxa which were not sampled for the chemical data coded in this paper. These reference trees were compared with cladograms estimated using four different coding methods. The four coding methods were: (1) a pathway-based approach using a character state tree, (2) a pathway-based approach using asymmetric stepmatrices, (3) a pathway-based approach using symmetric stepmatrices, and (4) simple presence/absence scoring of each compound. All analyses were performed using PAUP (Version 3.0s Swofford, 1991).

Similarity between the reference trees and those generated using the four coding methods was quantified using the symmetric-difference metric (Penny and Hendy, 1985) available in PAUP 3.0s (Swofford, 1991). The reference trees were derooted before symmetric-difference metrics were calculated because no outgroup was available for the fragrance compound data set. Trees obtained using the asymmetric stepmatrix coding method were derooted for subsequent comparisons. The reference phenogram calculated by Case (1994) is not a rooted cladogram, but for the purposes of this paper, only the branching relationships were of interest. Taxa not sampled by Case (1994) were excluded from trees estimated by the four competing coding methods before symmetric-difference metrics were calculated.

3. Results

Fig. 4 shows trees obtained for *Cypripedium* including: (a) Cox, (1995, pruned), (b) Case (1994), (c) floral fragrance data coded using character state tree, (d) floral fragrance data coded using the asymmetric stepmatrix approach, (e) floral fragrance data coded using the symmetric stepmatrix approach, and (f) floral fragrance data coded using presence/absence compound-based approach. Table 1 lists symmetric-difference metrics obtained in comparisons of the competing trees with the two reference trees. Coding method 2, utilizing the asymmetric stepmatrix, had the smallest symmetric-difference metric relative to reference tree 1, whereas analyses utilizing the character state tree had the smallest values relative to reference tree 2. The simple presence/absence compound-based coding procedure produced a topology (Fig. 4f) that had the highest symmetric-difference metrics relative to both reference trees.

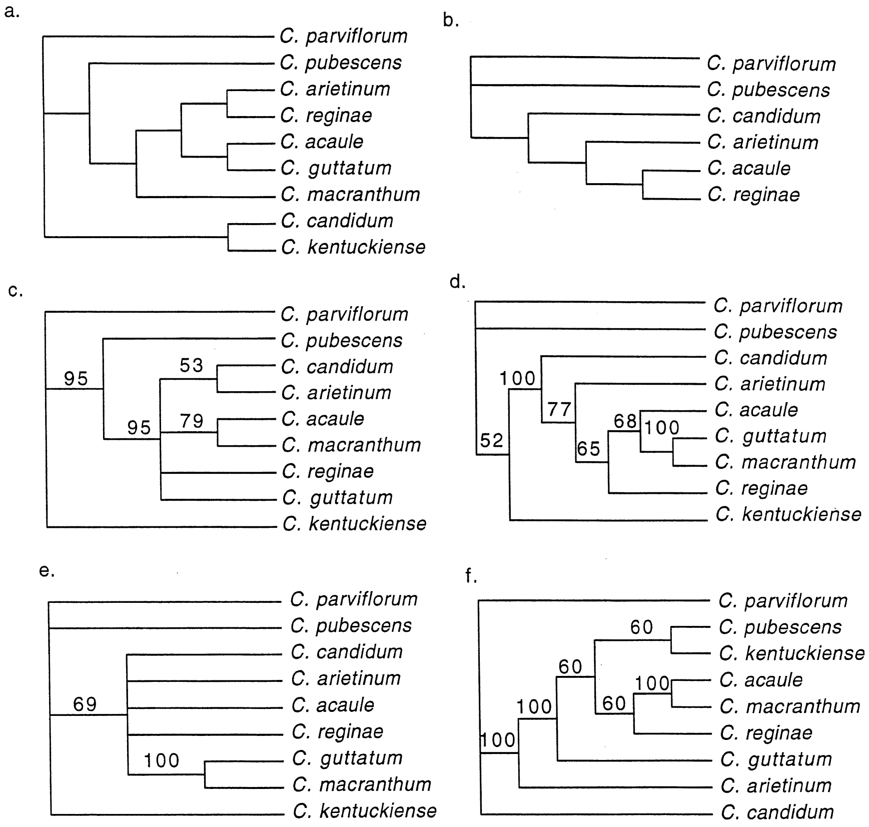


Fig. 4. Comparison of reference trees for *Cypripedium* derived from: (a) 5srDNA (Cox, 1995), (b) UPGMA estimated phenogram based on Nei's genetic identities (Case, 1994), (c) secondary compound variation using character state tree, (d) secondary compound variation using asymmetric stepmatrix, (e) secondary compound variation using symmetric stepmatrix, and (f) secondary chemical variation with each compound coded as present/absent. C–F are majority rule consensus trees with numbers above branches representing percentages of trees with that topology.

Table 1
Average symmetric difference metrics calculated between reference trees in Fig. 4a and b and competing trees in Fig. 4c–f obtained from four different coding methods for floral fragrance compound data in *Cypripedium*

	5srDNA	UPGMA
Fig. 4c character-state tree	10.6	2.4
Fig. 4d asymmetric stepmatrix	9.4	2.9
Fig. 4e symmetric stepmatrix	10.4	3
Fig. 4f presence/absence	11.2	5

4. Discussion

The pathway-based stepmatrix or character state tree coding procedures appear to be the most effective methods for studying secondary compound variation for several reasons. First, analyses using the stepmatrix or character state trees resulted in cladograms that were fairly similar to independent estimates of the evolutionary history in *Cypripedium*. The main point of incongruence between the reference tree in Fig. 4a and the consensus tree found in Fig. 4d regards the placement of *C. macranthum*. A probable cause for the putatively incorrect position of *C. macranthum* is the shared presence of pathways between this taxon and *C. guttatum* Sw. and *C. acaule* Ait. The position for *C. macranthum* in Fig. 4d is assumed to be erroneous due to the shared presence of these enzymatic conversions because the taxa have very different morphologies and have never been suggested as closely related by any other data set. These putatively homoplastic characters had a strong effect on the topology recovered largely because of the small number of characters used in this analysis. Ultimately, the combination of the secondary chemical data with other sources of data will provide a larger number of characters to more effectively estimate these relationships. The inability of the presence/absence coding method to recover the currently accepted phylogeny of *Cypripedium* suggests that non-independence among characters may be problematic. In addition, the coding of multiple non-independent characters results in an under-representation of the homology of represented compounds, therefore, researchers should avoid such non-pathway based coding procedures. The lack of "known" reference trees could have biased the comparisons performed in this paper. It should be noted, however, that the reference trees generally agree with long standing hypotheses about intrageneric relationships of *Cypripedium* (Atwood, 1984).

Second, from a theoretical perspective, the stepmatrix and character state tree methods used in this paper consider the pathway as the unit of evolution, thereby removing problems of character non-independence associated with other methods. While the stepmatrix and character state trees both utilize pathway-based information, the former has several advantages. Stepmatrices can be weighted to reflect biogenetic assumptions about pathway gains/losses, accommodate reticulations in biochemical pathways, and score polymorphisms if population- or individual-level variation exists. A character state tree, as a user-defined character in MacClade, cannot have differential weighting among transformations, does not allow reticulations and cannot accommodate polymorphic taxa. For these reasons, the asymmetric stepmatrix approach may be the preferable representation for biochemical pathway data; however, it does have two drawbacks. Analyses utilizing the asymmetric stepmatrix require greater amounts of computation time, which may be a problem when working with large data sets. Perhaps a more serious problem is the forced rooting that results from these analyses. This rooting occurs with or without outgroup taxa specified. The only solution is to define a rooted constraint topology to be imposed during tree searches. The character state tree transformational definitions appear more restrictive than stepmatrix definitions as shown in Fig. 3 because the path of transformations are constrained. It should be noted that, in theory, a stepmatrix can

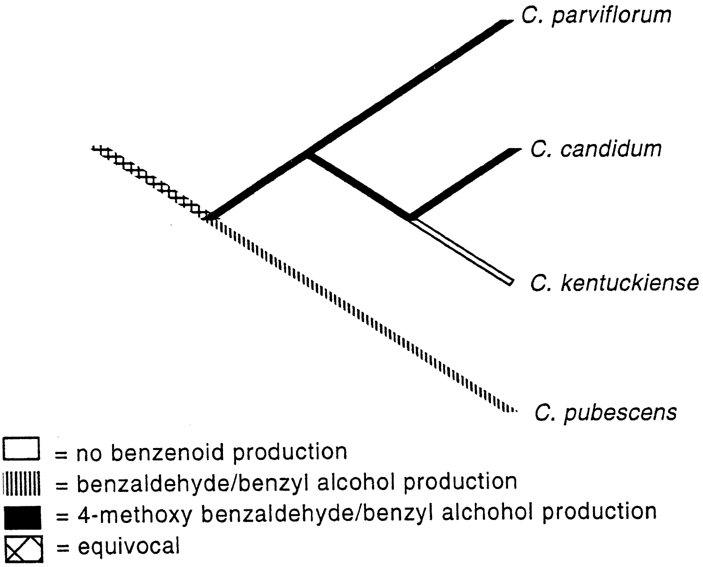

C. parviflorum
C. candidum
C. kentuckiense
C. pubescens
□ = no benzenoid production
▥ = benzaldehyde/benzyl alcohol production
■ = 4-methoxy benzaldehyde/benzyl alchohol production
☒ = equivocal

Fig. 5. Reconstruction of ancestral states and evolutionary changes in the benzenoid pathway based on the phylogenetic estimate of Cox (1995) for *Cypripedium*. Note that the ancestor of *C. candidum* and *C. kentuckiense* likely produced benzaldehyde and 4-methoxy benzaldehyde.

be defined to match the constraints of any character state tree and if both are based on the same biogenetic assumptions, paths of transformation will generally be the same.

Third, interpretation of biochemical evolution is significantly enhanced through the use of the stepmatrix or character state tree. The evolutionary modification of a biochemical pathway may be easily observed from character state reconstructions on a rooted phylogenetic tree. In contrast, little information about evolutionary change is obtained when reconstructing presence/absence of a compound on a tree. Although origin or loss of a compound may be identified, no information is gained about precursors or pathway evolution in this latter case because pathway intermediates are not included in character definitions. As an example, Fig. 5 shows the inferred evolution of benzenoid derivatives in *Cypripedium*. The complete loss of benzenoid expression is inferred in *C. kentuckiense* C. F. Reed because its ancestor possessed the capability to produce benzenoids. By contrast, an equivocal ancestral condition is inferred for the ancestor of *C. kentuckiense* when reconstructing presence/absence characters of multiple benzenoid compounds (data not shown).

4.1. Advantages to the general method

The pathway-based coding method presented here is relatively simple to use, though it requires assumptions about biogenetic pathways for the compounds sampled. This is not problematic because biogenetic pathways have been studied for flavonoids, sesquiterpene lactones, terpenoids, glucosinolates, iridoids, anthocyanins, and alkaloids, i.e., many of the important compounds studied in a systematic context for higher plants. This methodology should be generalizable to any class of compounds because, although compound types and substitution features of them may vary tremendously, they are all derived via enzymatic conversions that can be identified as independent or not and then coded appropriately. An advantage to the use of biochemical data over other types of data is that when enzymatic conversions are known to produce particular compounds, delimitation of the character state is unambiguous. The direct correspondence between gene and enzyme allows comparison of genetic differences between taxa studied. Rarely is a comparable understanding of the genetic basis of morphological characters available.

4.2. Limitations of secondary compound data

A problem inherent to secondary compound data, particularly floral fragrances, is that the spectrum of compounds sampled within an individual can vary quantitatively and qualitatively with differing environmental conditions, developmental stage, circadian rhythms and experimental protocols (Tollsten, 1993; Jakobsen and Olsen, 1994; Hills, 1989; Altenburger and Matile, 1988; Matile and Altenburger, 1988; Loughrin et al., 1991). Although careful sampling methods may control for abiotic conditions, this is difficult in field studies. Many methods exist for the delimitation of continuous characters. However, the character coding method suggested in this paper does not incorporate quantitative variation because it seems to be affected by non-phylogenetic factors. Furthermore, the distinction between qualitative absence and accumulation of only small quantities of a compound becomes blurred if end-product quantities are too low to be detected by analytical instrumentation, even though the enzymatic conversions may occur. In addition, because enzymatic conversions are the information of interest for coding, it is irrelevant whether the plant accumulates large or small quantities of the compound. Quantitative variation is more important for ecological and physiological studies, however.

4.3. Unresolved problems

Although unique difficulties may arise when considering any biochemical pathway, this method deals with some of the more common problems. Perhaps the most difficult problem remaining to be dealt with concerns extremely diverse groups of compounds derived from single precursors. For instance, consider that at least 2000 sesquiterpenes are produced from one or a few precursors. Interestingly, Steele et al. (1998) found 54 sesquiterpenes were produced from a single enzymatic reaction underscoring the role experimental data will have in assisting future coding efforts of such complicated chemical groups.

Another potential problem may be encountered if single taxa express all the compounds within a delimited pathway and therefore cannot be assigned a single character state. While the stepmatrix approach can be used in cases of extensive polymorphism, the coding procedure of Mabee and Humphries (1993) may be a useful alternative.

4.4. Analogy to ontogenetic data

De Queiroz (1985) stressed that the appropriate delimitation of morphological characters is the ontogenetic transformation rather than the "instantaneous morphologies" used by most systematists. The instantaneous morphology is the character state possessed by an organism when it is sampled. Using a biosynthetic pathway as a character, as stressed in this paper, is analogous to considering an ontogenetic transformation as a character. In the case of biochemical pathways, the individual compounds accumulated are the instantaneous morphologies of De Queiroz. In organisms sampled for morphology or secondary chemicals, the stage of development can greatly affect the character state observed. Ontogenetic stages are transitory, however, and the morphological character states possessed will differ between juveniles and adults. In contrast, a single individual sampled for secondary chemicals may possess multiple character states simultaneously, because when sampled, an individual may produce end-products as well as precursor compounds. The coding of a biochemical pathway as a character, like an ontogenetic sequence, allows inference of ancestral states that are important for inferring types and frequencies of

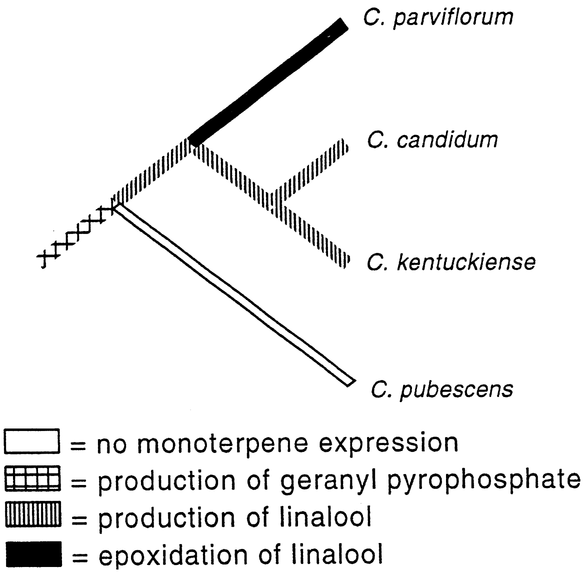

C. parviflorum
C. candidum
C. kentuckiense
C. pubescens
□ = no monoterpene expression
▦ = production of geranyl pyrophosphate
▥ = production of linalool
■ = epoxidation of linalool

Fig. 6. Reconstruction of ancestral states and evolutionary changes in the monoterpene pathway based on the phylogenetic estimate of Cox (1995) for *Cypripedium*. It is inferred that *C. parviflorum* evolved the ability to produce linalool oxide from an ancestor that only produced linalool.

evolutionary changes. Ontogenetic paedomorphosis may cause taxa to possess "juvenile" morphological characteristics through terminal deletion, and in biogenetic pathways, loss of enzymatic steps may occur in nested species giving the appearance of a "primitive" secondary chemical profile. As an example, Fig. 5 shows loss of benzenoid pathway expression in *C. kentuckiense* from an ancestor that expressed several benzenoid pathway reactions. A paedomorphic pattern was also suggested by Armbruster (1996), whereby the suppression of triterpenoid synthesis has resulted in the expression of the less biochemically-derived monoterpenoids. Also, ontogenetic terminal additions inferred from a rooted cladogram (Mabee, 1993) are analogous to chemical pathways or parts of them in which a precursor is further derivatized by the addition of a substitution group such as an epoxidation modification of linalool. This type of evolutionary modification may be found in *Cypripedium* because, as shown in the rooted cladogram of Fig. 6 (derived from Cox, 1995), *C. parviflorum* Salisb. expresses linalool oxide whereas the ancestor only produced linalool. Additional evolutionary inferences of this sort await further studies of secondary chemical variation considered in a phylogenetic framework.

Acknowledgements

The author would like to express his gratitude to D.A. Straney for initial conversations concerning coding of cladistic characters, Rodney Croteau, Natalia Dudareva, Robert Raguso, and Eran Pichersky, for discussions about secondary compound biogenesis, Douglas Gage and J.H. Beaman for support throughout the study, and John Clement, Joel McNeal, Beryl Simpson, Norris Williams, and an anonymous reviewer for valuable suggestions to improve the manuscript.

Appendix A

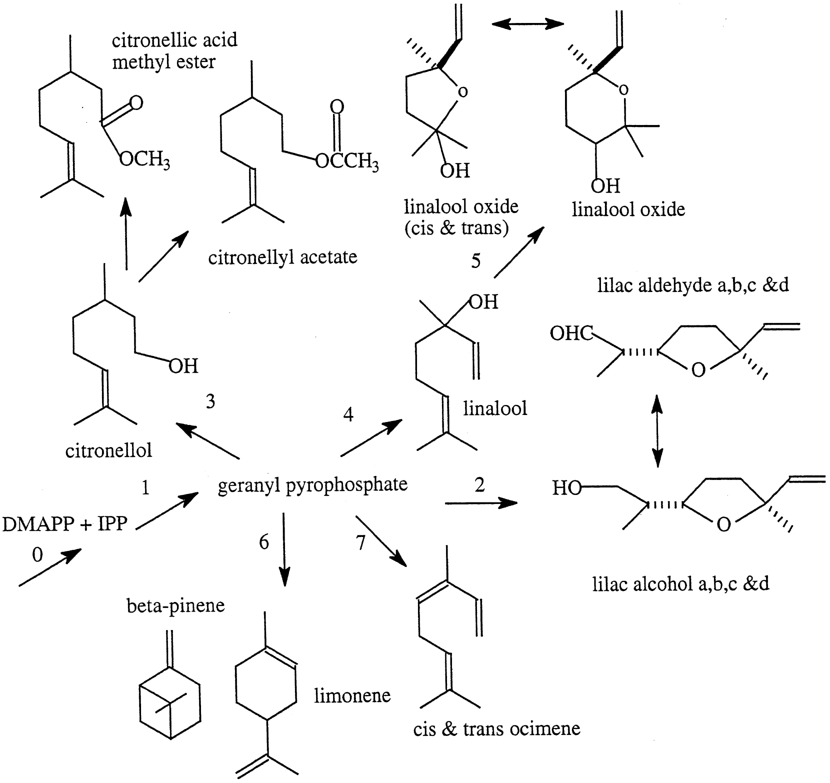
Data matrix for *Cypripedium*. Characters 1–4 are described in Figs. 1 and 2 and Appendices B and C. Characters 5–8 are not illustrated but are available upon request.

Taxa	1. Phenylalanine	2. Benzoic acid	3. Mono-terpenes	4. Sesqui-terpene	5. Fatty acid, methyl and ethyl ester	6. Fatty-acid acetates, alcohols	7. 1,4-Di-methoxy-benzene	8. Anthra-nilic acid
C. parviflorum	0	3	5 & 7	4	0	1	1	1
C. pubescens	2	2	7	0	0	1	1	0
C. candidum	1	3	4	0	0	0	0	0
C. arietinum	0	3	0	0	0	0	0	0

C. kentuckiense	1	0	4 & 7	4	0	2	0	0
C. acaule	0	4	2 & 7	0	2	0	0	1
C. reginae	0	6	0	0	0	0	0	0
C. guttatum	4	2	3	3	1	0	0	0
C. macranthum	5	7	2 & 7	2	0	0	0	0

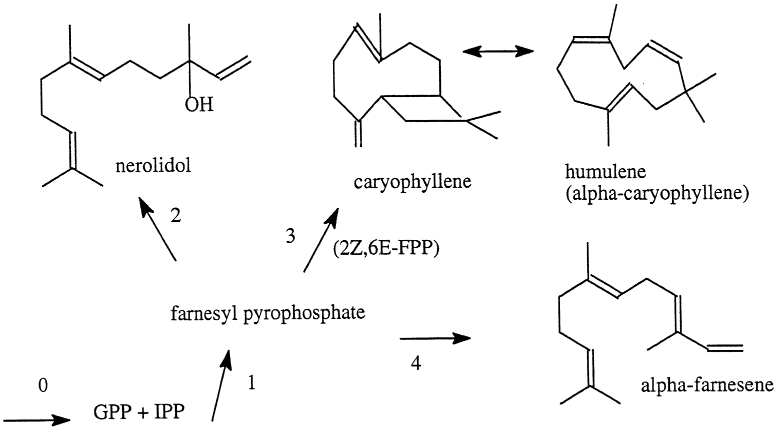
Appendix B

Assumed biogenetic pathway for monoterpenes sampled from *Cypripedium* that served as the basis for defining an independent multistate character.



Appendix C

Assumed biogenetic pathway for sesquiterpenes sampled from *Cypripedium* that served as the basis for defining an independent multistate character.



References

Altenburger, R., Matile, P., 1988. Circadian rhythmicity of fragrance emission in flowers of *Hoya carnosa*. R. Br. Planta 174, 248–252.

Armbruster, W.S., 1996. Exaptation, adaptation, and homoplasy: evolution of ecological traits in *Dalechampia* vines. In: Sanderson, M.J., Hufford, L. (Eds.), Homoplasy: the Recurrence of Similarity in Evolution. Academic Press, San Diego, pp. 227–243.

Atwood, J.T., 1984. The relationships of the slipper orchids (subfamily Cypripedioideae Orchidaceae): Selbyana 7, 129–247.

Barkman, T.J., Beaman, J.H., Gage, D.A., 1997. Floral fragrance variation in *Cypripedium* (Orchidaceae): implications for evolutionary and ecological studies. Phytochemistry 44, 875–882.

Bate-Smith, E.C., Richens, R.H., 1973. Flavonoid chemistry and taxonomy in *Ulmus*. Biochem. Syst. Ecol. 1, 141–146.

Bohlmann, J., Steele, C.L., Croteau, R., 1997. Monoterpene synthases from grand fir (*Abies grandis*). J. Biol. Chem. 272, 21784–21792.

Bohlmann, J., Meyer-Gauen, G., Croteau, R., 1998. Plant terpenoid synthases: molecular biology and phylogenetic analysis. Proc. Natl. Acad. Sci. USA 95, 4126–4133.

Bolick, M.R., 1983. A cladistic analysis of the Ambrosiinae Less. and Engelmanniinae Stuessy. In: Platnick, N.I., Funk, V.A. (Eds.), Advances in Cladistics, Vol. 2. Columbia University Press, New York, pp. 125–141.

Case, M.A., 1994. Extensive variation in the levels of genetic divergence and degree of relatedness among five species of *Cypripedium*. Am. J. Bot. 81, 175–184.

Cox, A., 1995. Molecular systematics of *Cypripedium*: phylogenetic utility of the 5s rDNA spacer. Ph.D. Thesis, University of Reading.

Croteau, R., Karp, F., 1991. Origin of natural odorants. In: Muller, P.M., Lamparsky, D. (Eds.), Perfumes; Art, Science, and Technology. Elsevier Applied Science, New York, pp. 101–126.

Cseke, L., Dudareva, N., Pichersky, E., 1998. Structure and evolution of linalool synthase. Mol. Biol. Evol. 15, 1491–1498.

Culberson, C.F., 1986. Biogenetic relationships of the lichen substances in the framework of systematics. Bryologist 89, 91–98.

De Queiroz, K., 1985. The ontogenetic method for determining character polarity and its relevance to phylogenetic systematics. Syst. Zool. 34, 280–299.

Dudareva, N., Raguso, R.A., Wang, J., Ross, J.R., Pichersky, E., 1998. Floral scent production in *Clarkia breweri* III. Enzymatic synthesis and emission of benzenoid esters. Plant Physiol. 116, 599–604.

Figueiredo, M.R., Auxiliadoro, M., Kaplan, C., Gottlieb, O.R., 1995. Diterpenes, taxonomic markers?: Pl. Syst. Evol. 195, 149–158.

Gambliel, H., Croteau, R., 1984. Pinene cyclases I and II. J. Biol. Chem. 259, 740–748.

Gottlieb, O.R., 1989. The role of oxygen in phytochemical evolution towards diversity. Phytochemistry 28, 2545–2558.

Gray, J.C., 1987. Control of isoprenoid biosynthesis in higher plants. In: Preston, A.D. (Ed.), Advances in botanical research, Vol. 14. Academic Press Limited, New York, pp. 25–91.

Gross, G.G., 1981. Phenolic acids. In: Stumpf, P.K., Conn, E.E. (Eds.), The biochemistry of plants, Vol. 7. Academic Press Inc, New York, pp. 301–316.

Harborne, J.B., Turner, B.L., 1984. Plant Chemosystematics. Academic Press, New York.

Hills, H.G., 1989. Fragrance cycling in *Stanhopea pulla* (Orchidaceae, Stanhopeinae) and identification of trans-limonene oxide as a major fragrance component. Lindleyana 4, 61–67.

Humphries, C.J., Richardson, P.M., 1980. Hennig's methods and phytochemistry. In: Bisby, F.A., Vaughan, J.G., Wright, C.A. (Eds.), Chemosystematics: principles and practice. Academic Press, London, pp. 353–378.

Jakobsen, H.B., Olsen, C.E., 1994. Influence of climatic factors on emission of flower volatiles in situ. Planta 192, 365–371.

Jeffrey, C., 1995. Compositae systematics 1975–1993 developments and desiderata. In: Hind, D.J.N., Jeffrey, C., Pope, G.V. (Eds.), Advances in Compositae Systematics. Royal Botanic Gardens, Kew, pp. 3–21.

Levy, M., 1977. Minimum biosynthetic-step indices as measures of comparative flavonoid affinity. Syst. Bot. 2, 89–97.

Loughrin, J.H., Hamilton-Kemp, T.R., Andersen, A., Hildebrand, D.F., 1991. Circadian rhythm of volatile emission from flowers of *Nicotiana sylvestris* and *N. suaveolens*. Physiol. Plant. 83, 492–496.

Mabee, P.M., 1993. Phylogenetic interpretation of ontogenetic change: sorting out the actual and artefactual in an empirical case study of centrarchid fishes. Zool. J. Linn. Soc. 107, 175–291.

Mabee, P.M., Humphries, J., 1993. Coding polymorphic data: examples from allozymes and ontogeny. Syst. Biol. 42, 166–181.

Maddison, W.P., Maddison, D.R., 1992. MacClade: Analysis of phylogeny and character evolution. Version 3.0. Sinauer Associates. Sunderland, MA.

Miao, B., Turner, B.L., Mabry, T., 1995. Molecular phylogeny of *Iva* (Asteraceae, Heliantheae) based on chloroplast DNA restriction site variation. Plt. Syst. Evol. 195, 1–12.

Nam, K.H., Dudareva, N., Pichersky, E., 1999. Characterization of benzylalcohol acetyltransferases in scented and non-scented *Clarkia* species. Plant Cell Physiol 40, 916–923.

Nandi, O.I., Chase, M.W., Endress, P.K., 1998. A combined cladistic analysis of angiosperms using *rbc*L and non-molecular data sets. Ann. Missouri Bot. Gard. 85, 137–212.

Patterson, C., 1988. Homology in classical and molecular biology. Mol. Biol. Evol. 5, 603–625.

Penny, D., Hendy, M.D., 1985. The use of tree comparison metrics. Syst. Zool. 34, 75–82.

Pichersky, E., Raguso, R.A., Lewinsohn, E., Croteau, R., 1994. Floral scent in *Clarkia* (Onagraceae) I. localization and developmental modulation of monoterpene emission and linalool synthase activity. Plt. Phys. 106, 1533–1540.

Plunkett, G.M., Soltis, D.E., Soltis, P.S., 1996. Higher level relationships of Apiales (Apiaceae and Araliaceae) based on phylogenetic analysis of *rbc*L sequences. Am. J. Bot. 83, 499–515.

Raguso, R.A., Pichersky, E., 1995. Floral volatiles from *Clarkia breweri* and *C. concinna* (Onagraceae): recent evolution of floral scent and moth pollination. Pl. Syst. Evol. 194, 55–67.

Raguso, R.A., Pichersky, E., 1999. A day in the life of a linalool molecule: chemical communication in a plant-pollinator system. Part 1: linalool biosynthesis in flowering plants. Pl. Sp. Biol. 14, 95–120.

Richardson, P.M., 1983. Flavonoids and phylogenetic systematics. In: Platnick, N.I., Funk, V.A. (Eds.), Advances in Cladistics, Vol. 2. Columbia University Press, New York, pp. 115–123.

Richardson, P.M., Young, D.A., 1982. The phylogenetic content of flavonoid point scores. Bioch. Syst. Ecol. 10, 251–255.

Ross, J.R., Nam, K.H., D'auria, J.C., Pichersky, E., 1999. S-adenosyl-L-Methionine-salicylic acid carboxyl methyltransferase, an enzyme involved in floral scent production and plant defense, represents a new class of plant methyltransferases. Arch. Biochem. Biophys 367, 9–16.

Sacchettini, J.C., Poulter, C.D., 1997. Creating isoprenoid diversity. Science 277, 1788–1789.

Schreier, P., 1984. Chromatographic Studies of Biogenesis of Plant Volatiles. Dr. Alfred Hüthig Verlag, Heidelberg.

Seaman, F.C., Funk, V.A., 1983. Cladistic analysis of complex natural products: Developing transformation series from sesquiterpene lactone data. Taxon 32, 1–27.

Sneath, P.H., Sokal, R.R., 1973. Principles of Numerical Taxonomy. W.H. Freeman and Co, San Francisco.

Steele, C.L., Crock, J., Bohlmann, J., Croteau, R., 1998. Sesquiterpene synthases from grand fir (*Abies grandis*). J. Biol. Chem. 273, 2078–2089.

Swofford, D.L., 1991. PAUP Phylogenetic analysis using parsimony. Version 3.0s. Computer program distributed by the Illinois Natural History Survey. Champaign, ILL.

Tollsten, L., 1993. A multivariate approach to post-pollination changes in the floral scent of *Platanthera bifolia* (Orchidaceae). Nord. J. Bot. 13, 495–499.

Wang, J., Pichersky, E., 1998. Characterization of *S*-adenosyl-L-methionine: (iso)eugenol *O*-methyltransferase involved in floral scent production in *Clarkia breweri*. Arch. Biochem. Biophys. 349, 153–160.

Wang, J., Pichersky, E., 1999. Identification of specific residues involved in substrate discrimination in two plant *O*-methyltransferases. Arch. Biochem. Biophys. 368, 172–180.

Wang, J., Dudareva, N., Bhakta, S., Raguso, R.A., Pichersky, E., 1997. Floral scent production in *Clarkia breweri* (Onagraceae) II. Localization and developmental modulation of the enzyme *S*-adenosyl-L-methionine: (iso)eugenol *O*-methyltransferase and phenylpropanoid emission. Plant Physiol. 114, 213–221.

Wheeler, C.J., Croteau, R., 1986. Terpene cyclase catalysis in organic solvent/minimal water activity media: demonstration and optimisation of (+)-alpha-pinene cyclase activity. Arch. Biochem. Biophys. 248, 429–434.

Yuba, A., Yazaki, K., Tabata, M., Honda, G., Croteau, R., 1996. cDNA cloning, characterization, and functional expression of 4*S*-(−)-limonene synthase from *Perilla frutescens*. Arch. Biochem. Biophys. 332, 280–287.