3 Secondary Structure of the ITS1 Transcript and its Application in a Reconstruction of the Phylogeny of Boraginales[2]

Abstract

In this study, we present the possibilities for calculating systematic relationships on higher taxonomical levels based exclusively on ITS1. This is demonstrated for Boraginales (Boraginaceae s.str., Cordiaceae, Ehretiaceae, Heliotropiaceae, Hydrophyllaceae, and Lennoaceae). Secondary structure of the ITS1 region is more conserved than the primary structure (i.e., sequence itself) and is therefore a tool for optimising alignments. It increases the number of structural characters. Information inferred from the secondary structure enables us to construct well-resolved phylogenetic trees at higher taxonomic levels. A general secondary structure of ITS1 for Boraginales with four major helices, is proposed here. In each subtaxon, derivations from this structure are found. The paraphyly of Boraginaceae s.l. is evident from both a comparison of secondary structures and from bootstrap analysis. Boraginaceae s.str. are the sister group of a clade formed by Hydrophyllaceae, Cordiaceae, Ehretiaceae, Heliotropiaceae, and Lennoaceae. The last four taxa constitute a monophyletic group which is the sister group of Hydrophyllaceae. Lennoaceae are closely related to Ehretiaceae, and these two taxa in turn are the sister group of Cordiaceae.

3.1 Introduction

The highly variable first internal transcribed spacer (ITS1) lies between the 18S and the 5.8S genes of the nucleus. It was initially proposed for use at lower taxonomical levels (BALDWIN

[2] Published as GOTTSCHLING M., HILGER H.H., WOLF M. & **DIANE N.** (2001): Secondary structure of the ITS1 transcript and its application in a reconstruction of the phylogeny of Boraginales. - Pl. Biol. **3**: 629-636.

Own contributions: sequencing (partly), calculating secondary structure, alignment (all in cooperation with M. GOTTSCHLING), review and discussion of the manuscript.

1992) and has been extensively employed in phylogenetic studies. It has recently been argued that the secondary structure of the ITS1 region is conserved at higher taxonomical levels, while the primary sequence is highly variable (NUES *et al*. 1994, COLEMAN & MAI 1997, COLEMAN *et al*. 1998).

Some attempts have been made in different eukaryotic taxa to reconstruct phylogeny inferred from the secondary structure of ITS (e.g., Volvocales by COLEMAN & MAI 1997, *Ancylostoma* [hookworms] by CHILTON & GASSER 1999, Digenea [Trematoda] by SCHULENBURG *et al*. 1999). However, our knowledge of secondary structure outside these few groups, and especially in the spermatophytes, is still very scant.

Boraginales are currently considered as compromising Lennoaceae, Hydrophyllaceae, and Boraginaceae *sensu* GÜRKE (1893, "Boraginaceae s.l." with Cordioideae, Ehretioideae, and Heliotropiodeae, Boraginoideae). Paraphyly of "Boraginaceae s.l." is now undisputed since at least some groups of "Boraginaceae s.l." are more closely related to Hydrophyllaceae according to molecular data (OLMSTEAD *et al*. 1993, COSNER *et al*. 1994, BÖHLE & Hilger 1997, FERGUSON 1999, ALBACH *et al*. 2001).

In this study, we used ITS1 sequences of 50 representatives of Boraginales, and by mutual comparison of these, we have deduced a basic secondary structure common to the entire group. This procedure has been successfully used previously for ITS2 (AN *et al*. 1999). We then used the information from a secondary structure model for optimising alignments.

3.2 Materials and Methods

Representatives of 50 Boraginales have been investigated and are listed in Table **3-1** (appendix). Representatives of Solanaceae were used for a user-specified outgroup comparison. DNA extraction, PCR, purification, and sequencing followed standard protocols (DIANE *et al*. 2002a, GOTTSCHLING & HILGER 2001).

Common structural elements were initially recognized with the help of mFOLD (JAEGER *et al*. 1989, 1990) by screening for thermodynamically optimal and suboptimal secondary structures. Pairing regions (helices or "hairpins") were identified by mutual comparison of sequences in a manual alignment and by compensating substitutions (e.g., pairing GC into pairing AT). These

hairpins were taken as homologous or at least conserved sequence regions in the alignment of all taxa. Energy levels of the proposed secondary structures were then computed with mFOLD (JAEGER *et al*. 1989, 1990).

On the basis of a framework dictated by secondary structure, the sequences were then manually aligned (bi-directional, complementary) using Se-Al v2.0a72 (RAMBAUT 2001). The complete data matrix is available in NEXUS format on request. Parsimony analyses were performed using PAUP* 4.0b1 (SWOFFORD 1998) for Power Macintosh PC and NONA 2.0 (GOLOBOFF 1993). Heuristic searches were run in PAUP without changes from the default settings. All characters were weighted equally, and character state transitions were treated as unordered. Bootstrap resampling (FELSENSTEIN 1985) was performed with 1000 replicates and a heuristic search with the following settings: addseq = random, nreps = 100, and mulpars = off for each bootstrap replicate. For both methods, heuristic search and bootstrap analysis, gaps were considered as informative and calculated as fifth base, to prevent the loss of information on secondary structure.

3.3 Results

Secondary structure—Several derivations from the secondary structure of the ITS1 found in Boraginales are proposed in Figures. **3-1** to **3-7**. These are the basis for reconstructing a general secondary structure, as shown in Figure **3-8**. At least four helices of pairing regions (I to IV) can be identified: two long ones (up to 25 bp long) and two short ones (about 6 to 10 bp and 5 to 6 bp, respectively). Additional helices are found at the 5'- and 3'-end of the ITS1 and are indicated with lowercase letters. These helices are not universally present in all species investigated. The two short helices are comparatively conservative (especially helix III), while the long helices I and IV show high primary sequence variability. Statistics of the derivations from the secondary structure are given in Table **3-2** (appendix): helix III is nearly uniform in length (5 to 6 bp), while helix II shows distinct size classes (10 to 14 bp *versus* 6 to 7 bp). Helices I and IV are more or less uniform in length (17 to 25 bp), with the notable exception of *Heliotropium* (5 bp in helix I). The GC content differs in the sequences from 46% to 66%. Proportions of pairing G—U in the helices are usually low (5% to 18%); *Borago officinalis* is an exception with 27%.

The major helices are flanked by A-rich regions in the stem loop (Fig. **3-8**). One part of the stem loop (between helix III and IV) is conserved throughout all species under investigation, with the motif 5'—AAGGAA—3' (exceptions are only found sporadically in, e.g., *Ehretia acuminata* or *Pholisma arenaria*). The others are variable at species level.

Tree—The aligned ITS1 data set was 383bp in length after taking secondary structure into consideration. The alignment of helix III and IV is shown in Figure **3-9** as an example. 293 bp (77%) were informative (5,9 bp per taxon). The heuristic search found five most parsimonious trees, one of which is shown in Figure **3-10**. Using different programs (PAUP and NONA) yielded trees with identical topology and equivalent bootstrap support values.

Systematic relationships—All ingroup taxa constitute a monophyletic group with respect to the outgroup (Fig. **3-10**). Within the parsimony tree, six clades can be identified, five of which are supported by bootstrap support values (BS) above 50 %: Boraginaceae s.str. (100% BS), Hydrophyllaceae (87% BS), Heliotropiaceae (54% BS), Cordiaceae (71% BS), Ehretiaceae (<50% BS), and Lennoaceae (100% BS). Ehretiaceae and Lennoaceae constitute a well-supported monophylum (96% BS).

The relationships between the subtaxa identified in Boraginales are also supported by bootstrap analysis. Boraginaceae s.str. are the sister group of a clade formed by the remaining Boraginaceae s.l. (Cordiaceae, Ehretiaceae, Heliotropiaceae) plus Hydrophyllaceae and Lennoaceae (100% BS). Cordiaceae, Ehretiaceae, Heliotropiaceae, and Lennoaceae constitute a well-supported monophylum (85% BS). Lennoaceae are the sister group of Ehretiaceae or are closely related to a part of it; this problem could not readily be resolved unequivocally by this study. However, they both are the sister group of Cordiaceae (87% BS).

3.4 Discussion

Deducing secondary structure from sequences—It has been proposed that the basic secondary structure of ITS1 in Chlorobionta (i.e., green algae and land plants) is a circle with some protruding helices ("hairpins") of pairing bases (LIU & SCHARDL 1994, COLEMAN *et al*. 1998, MAYOL & ROSSELLÓ 2001). Several authors have emphasized that the secondary, but not the

primary, structure of ITS1 is conserved at higher taxonomical levels (NUES *et al*. 1994, COLEMAN & MAI 1997, COLEMAN *et al*. 1998).

mFOLD (JAEGER *et al*. 1989, 1990) yields highly divergent secondary structures, even from the same sequence, depending on the settings (e.g., temperature). From the phylogenetic point of view, the thermodynamically optimal structure does not necessarily reflect the *in vivo* structure since evolution does not always proceed *via* the shortest route. Satisfactory general secondary structures can be recognized through mutual sequence comparison in an alignment, like deduction of a phylogenetic tree from the character table.

In Boraginales, all sequences under investigation have at least four homologous helices (I to IV, Fig. **3-8**). This assumption was supported by numerous compensating base substitutions. For example, the conserved motif 5'—CGGGGC-(nN)-GCCCCG—3' of position 153 to 175 (helix II) has changed to 5'—CG**A**GGC-(nN)-GCC**U**CG—3' in *Bourreria* (Ehretiaceae). Similarly, the highly conserved motif 5'—GGCGC-(nN)-GCGCC—3' of position 204 to 220 (helix III) has changed to 5'—GGC**A**C-(nN)-G**U**GCC—3' in *Cordia* sect. *Varronia* (Cordiaceae).

Our helices II and III, as well as the conserved sequence 5'—AAGGAA—3' in association with helix III, were also found by LIU & SCHARDL (1994). With few exceptions, this motif appears to be unique to land-plants (SLOTTA 2000) and may have a key function in the processing of rRNA gene transcripts (LIU & SCHARDL 1994). LIU & SCHARDL (1994) found two major helices near 5'- and 3'-ends for *Arabidopsis thaliana* (Brassicaceae). These helices are probably homologous to helices I and IV in Boraginales because they are long (26 bp and 21 bp, respectively) and also found in other plant species (LIU & SCHARDL 1994). The secondary structures of ITS1 in species of *Quercus* proposed by MAYOL & ROSELLÓ (2001) inferred from sequences of MANOS *et al*. (1999) are also similar to those we found in Boraginales: their helices IV to VII seem to be homologous with our helices I to IV.

The helices are flanked by A-rich motifs, while the helices themselves have a high GC content. The bonds between C and G are stronger than between A and U, and it is more likely that pairing regions are comprised of bases with strong bonds, while unpaired regions are rich in bases with lower bond paring potential. Furthermore, the amount of G—U pairings in the proposed helices are low (in *Borago officinalis* they are slightly higher), indicating high stability of the helices. The GC content of the sequences varies from 46% to 66%; varying GC content of ITS1 is also known from other taxa (e.g., BALDWIN 1992, RITLAND *et al*. 1993).

Secondary structure as a tool for recognising relationships—The identification of secondary structure may help to recognize relationships. For example, the large deletion found in some representatives of *Heliotropium* of the Old World (Heliotropiaceae) between position 61 and 111 in the alignment of DIANE *et al*. (2002a) can now be interpreted as a radical abridgement of helix I (Fig. **3-6**). This appears to be a structural apomorphic feature not found elsewhere and argues for the monophyly of the corresponding group.

Helix II also provides systematic insights: it is 10 to 14 bp long and shows high sequence variability (Figs. **3-1** and **3-2**) in both Boraginaceae s.str. and the outgroup representatives. This condition must be regarded as plesiomorphic within Boraginales. In contrast, Hydrophyllaceae, Ehretiaceae, Cordiaceae, Heliotropiaceae, and also parasitic Lennoaceae have a shorter helix II with fixed pairing sequences (typically: 5'—CGGGGC *versus* GCCCCG—3', Fig. **3-3** to **3-7**). This derived feature argues for the monophyly of the corresponding groups.

Phylogeny of Boraginales—Using information from secondary structure of the ITS1 region yields well-supported phylogenetic hypotheses which do not contradict former systematics of Boraginales (Fig. **3-10**). Additionally, it was possible to increase bootstrap support values on internal nodes in comparison with other studies. For example, the monophyly of Hydrophyllaceae was supported with 30 % BS in FERGUSON (1999) and is supported with 87% BS in this study.

The classification of "Boraginaceae s.l." has been controversial. While most recent authors retain GÜRKE'S (1893) division into four subfamilies (e.g., Al-SHEHBAZ 1991, VERDCOURT 1991), some segregate Cordiaceae, Ehretiaceae, and/or Heliotropiaceae from it (e.g., HUTCHINSON 1959, KHALEEL 1985, PARMAR 1991). Our molecular results confirm the paraphyly of "Boraginaceae s.l." (Fig. **3-10**). Futhermore, the idea of Lennoaceae being associated with Boraginales, previously proposed by YATSKIEVYCH *et al*. (1986), is supported by our molecular results.

To evaluate the general phylogeny of Boraginales, a representative set of species from all major groups were analyzed. This should provide a reliable reference system for the placement of other taxa, especially some enigmatic taxa whose exact relationships are difficult to ascertain on the basis of a direct comparison of sequences. We thus hope to elucidate the precise systematic position of taxa such as *Codon* (Hydrophyllaceae), *Trigonotis* (Boraginaceae s.str.), and *Wellstedia* (Wellstediaceae) in further studies.

The results of this study of the ITS1 region in Boraginales indicate that an investigation of secondary structure is a helpful tool to make full use of the phylogenetic information in DNA-sequences and can help to vastly improve the quality of alignments.