Pak. J. Bot., 43: 1-4, Special Issue, December, 2011 (Medicinal Plants: Conservation & Sustainable use)

DNA BARCODING AS A MEANS FOR IDENTIFYING MEDICINAL PLANTS OF PAKISTAN

M. SCHORI* AND A.M. SHOWALTER

Department of Environmental and Plant Biology, Molecular and Cellular Biology Program,
315 Porter Hall, Ohio University, Athens, OH 45701 USA.

Abstract

DNA barcoding involves the generation of DNA sequencing data from particular genetic regions in an organism and the use of these sequence data to identify or “barcode” that organism and distinguish it from other species. Here, DNA barcoding is being used to identify several medicinal plants found in Pakistan and distinguished them from other similar species. Several challenges to the successful implementation of plant DNA barcoding are presented and discussed. Despite these challenges, DNA barcoding has the potential to uniquely identify medicinal plants and provide quality control and standardization of the plant material supplied to the pharmaceutical industry.

Introduction

DNA barcoding is a method of identifying an organism based on sequence data from one to several gene regions. Many recent papers have been written about DNA barcoding in plants, including an elegant review by Hollingsworth *et al.,* (2011). Barcoding has multiple applications and has been used for ecological surveys (Dick & Kress, 2009), cryptic taxon identification (Lahaye *et al.*, 2008), and confirmation of medicinal plant samples (Xue & Li, 2011). Several chloroplast gene regions are typically used as plant barcodes, with maturase K (*matK*) and ribulose 1,5-bisphosphate carboxylase/oxygenase large subunit (*rbcL*) considered core barcodes (Hollingsworth *et al.,* 2009). The spacer between tRNA-His and photosystem II protein D1 (*trnH-psbA* spacer) and the nuclear internal transcribed spacer 2 (ITS2), however, are also widely used (Chen *et al.*, 2010; Gao *et al.*, 2010; Fu *et al.*, 2011). Barcoding works by matching sequence data from a query sample (an unknown specimen) to a reference sequence (from a voucher specimen). Our particular interest is in using DNA barcodes to confirm the identity of common medicinal plants of Pakistan. Medicinal plants are widely used in Pakistan, often in the form of packaged herbal preparations manufactured by the pharmaceutical industry. Raw materials are predominantly collected from the wild in the northern regions of the country and then transported to markets in Islamabad and other cities. The lack of cultivation, the possibility of inaccurate identification at the time of collecting, and the long supply chain from harvesting site to market provide opportunities for substitution or adulteration of the raw ingredients. In some cases substituting one species for another may have a minimal effect on a product’s efficacy, but in other cases the beneficial effect of a product may be lost entirely. Moreover, substitutions within certain plant families (especially Apiaceae and Solanaceae) could prove fatal, as medicinal and culinary plants may look very similar to poisonous species. Barcoding of suspect raw ingredients can confirm or disprove the identity of medicinal plants before they are processed, enabling the pharmaceutical industry and consumers to use plant species that are known to be effective.

Our research efforts are focused on developing and testing reference DNA barcodes for a particular set of medicinal plants from Pakistan. Table 1 lists the initial group of species that has been collected. The part of the plant that is used medicinally varies among species, but all reference barcodes are being generated from leaf tissue dried in silica gel to ensure that high-quality genomic DNA is used for analysis. Several species, including *Anethum graveolens*, *Foeniculum vulgare* and *Linum usitatissimum*, have a long history of cultivation and are widely grown around the world. Their identification may pose no problems at the time of collection, but in the case of *A. graveolens* and *F. vulgare*, it may be difficult to physically distinguish leaf material once it has been dried and packaged. Fruits of these species, erroneously referred to as seeds in culinary and medicinal literature, may be difficult to correctly identify if material from related species has been added. In cases where a visual assessment confirms the presence of more than one species, barcoding can be used to identify both the target species and the contaminants. Certain sequence data for the species in Table 1 are already available in GenBank. We examined how accurately these barcoding regions can identify the species. A discussion of challenges and limitations to barcoding is provided with our results.

Materials and Methods

Sequence data for 12 medicinal plants were downloaded from the GenBank database (http://www.ncbi.nlm.nih.gov/genbank/). The gene regions chosen were *rbcL*, *matK*, and *psbA-trnH*, for a total of 22 available sequences, as not all gene regions have been sequenced for each species. Each sequence was entered into GenBank’s BLAST search function (http://blast.ncbi.nlm.nih.gov/Blast.cgi), using the Megablast parameter, to assess similarities between barcoding sequences in the medicinal plants and related taxa. Percent similarity was examined for the closest matches.

*Author for correspondence E-mail: schori@ohio.edu

Table 1. General information on fourteen species of medicinal plants from Pakistan collected for DNA barcoding in this study. Species in bold have not been sequenced for two or three of the primary barcoding regions.

Species	Family	Common Name	Part used
Acacia nilotica	Fabaceae	Gum arabic	Bark, leaves
Achyranthes aspera	Amaranthaceae	Prickly chaff flower	Roots, leaves
Amaranthus caudatus	Amaranthaceae	Foxtail amaranth	Leaves, seeds
Anethum graveolens	Apiaceae	Dill	Leaves, fruits
Calotropis procera	Apocynaceae	Milkweed	Roots, leaves, flowers
Carthamus oxyacantha	Asteraceae	Jeweled distaff thistle	Leaves, fruits
Foeniculum vulgare	Apiaceae	Fennel	Leaves, fruits
Justicia adhatoda	Acanthaceae	Vasak	Leaves
Linum usitatissimum	Linaceae	Flax	Seeds
Nigella sativa	Ranunculaceae	Black cumin	Seeds
Rosa x *damascena*	Rosaceae	Damask rose	Petals
Solanum surattense	Solanaceae	Mamoli	All parts
Trachyspermum ammi	Apiaceae	Ajwain	Fruits
Trigonella foenum-graecum	Fabaceae	Fenugreek	Leaves, seeds

Results and Discussion

Several potential challenges can prevent DNA barcoding from being a fail-safe method of identifying plants. The first challenge is often in generating reference DNA sequences. Medicinal plants contain biologically useful secondary compounds, including tannins, alkaloids, and polysaccharides, all of which can inhibit DNA extraction and amplification by co-precipitating with or binding to DNA. If clean genomic DNA is obtained, appropriate primers are needed to amplify the targeted gene region. Certain gene regions, including *rbcL*, *trnL*, and the *trnL-F* intergenic spacer, have universal primers that work for most plants. Other regions like *matK* may be more variable and require custom primer design. Primers might also work adequately during the amplification phase (PCR), but may not be specific enough to work during cycle sequencing to generate fluorescently-labeled DNA. If the primer sequence is not an exact match to the primer region, cycle sequencing may fail despite evidence that PCR was successful. We are examining a wide range of medicinal plants, representing 10 distantly related families of flowering plants, so we expect to encounter issues of primer infidelity, especially for *matK*.

The second challenge is in authenticating the utility of barcoding by analyzing raw plant material purchased at markets. Reference sequences are generated from leaf tissue of voucher specimens, which are carefully collected and dried to minimize DNA degradation. Plant material in the markets may have a very different recent history. The age and condition of the plant at the time of collection is unknown, as are the conditions of drying, processing, and transporting the plant. The quality of the DNA is likely to be lower than from plants that were collected specifically for reference purposes. Depending upon which part of a plant is collected, it may be difficult to extract DNA for comparisons, especially in the case of bark or sap. Authentication is a critical aspect of our research, because we need to demonstrate that DNA from roots, seeds and fruits of medicinal plants can successfully be compared to reference sequences.

The third and fourth challenges, sampling and discrimination power, are related to authentication. It is not enough to know that DNA sequences match between a market sample and the reference barcode(s) for a given species. We must also verify that the barcode(s) can uniquely identify the sample. For example, if a market sample is verified as *Trachyspermum ammi*, its barcodes must match the reference for that species and they must not match *T. clavatum* or *T. baluchistanicum*. Barcoding needs to include sister species and other closely related taxa to ensure species-level specificity. Table 2 lists the number of congeneric species found in Pakistan for each medicinal plant from Table 1. A combination of two to three gene regions is typically used in barcoding. Within a given genus, the *rbcL* sequence might be identical for all species, which means that while *rbcL* might verify the genus, it cannot verify the species. Variability within a given gene region may be quite different from one genus or family to the next, so the best way to increase the likelihood of a positive identification is to adequately sample related species for every target plant. A pair of species in Fabaceae may be distinguished by *matK* and the *psbA-trnH* intergenic spacer, while a pair in Apiaceae is distinguished by *rbcL* and the *trnL-F* intergenic spacer. The combination of barcode regions that discriminates among related species will have to be determined separately for each medicinal plant. Even with increased sampling, it is likely that some raw material will not be matched to any reference barcodes. A recent paper by Stoeckle *et al.*, (2011) compared barcode sequence data from teas to their listed ingredients. In multiple cases, they sequenced DNA that had no matches in GenBank and did not correspond to any of the species listed as ingredients. If a reference sequence has not been generated, it is not always possible to identify an unknown or suspect plant. Thus, while it may be possible to determine to which family or genus such a plant belongs, it is unlikely that a species' identification will be confirmed.

Table 2. Target medicinal species and number of species in the same genus recorded in Flora of Pakistan.

Species	Family	# of Congeneric species in Pakistan
Acacia nilotica	Fabaceae	27
Achyranthes aspera	Amaranthaceae	1
Amaranthus caudatus	Amaranthaceae	10
Anethum graveolens	Apiaceae	0
Calotropis procera	Apocynaceae	1
Carthamus oxyacantha	Asteraceae	*
Foeniculum vulgare	Apiaceae	0
Justicia adhatoda	Acanthaceae	11
Linum usitatissimum	Linaceae	3
Nigella sativa	Ranunculaceae	1
Rosa x *damascena*	Rosaceae	14+**
Solanum surattense	Solanaceae	14
Trachyspermum ammi	Apiaceae	2
Trigonella foenum-graecum	Fabaceae	16

*The Flora of Pakistan (http://www.efloras.org/flora_page.aspx?flora_id=5) does not yet include all genera of Asteraceae

**Fourteen species of *Rosa* are included in the Flora of Pakistan but other species and hybrids are also cultivated in the country

DNA sequences that are currently available in GenBank demonstrate the challenge of discrimination power. Table 3 lists barcoding sequences from 12 of the 14 medicinal plants and indicates how well each sequence identifies the plant. The results vary by species and by gene region. For example, *rbcL* from *Acacia nilotica* does not vary among 14 different vouchers and distinguishes the species from other *Acacia* accessions in the database, but *psbA-trnH* varies by 1-2% among 12 different vouchers and is 98% similar to *Vachellia* (*Acacia*) *farnesiana*, which grows in Pakistan. This suggests that *rbcL* is a better barcode for *Acacia nilotica* than *psbA-trnH*. *Anethum graveolens* and *Foeniculum vulgare* are 99% similar in both their *matK* and *rbcL* sequences; two vouchers of *F. vulgare* were 99% similar to each other for these gene regions, which suggests that *matK* and *rbcL* may not adequately discriminate between *A. graveolens* and *F. vulgare*. *Anethum graveolens* has not yet been sequenced for *psbA-trnH*, so its utility as a barcode cannot be determined for this pair of species. *MatK*, *rbcL*, and *psbA-trnH* sequences for *Linum usitatissimum* were a 99-100% match both to *L. usitatissimum* vouchers and to *L. bienne*, indicating that none of these barcodes may be able to distinguish these two species. Other species had no close matches for barcoding regions, but this is likely to be an artifact of poor sampling. *Justicia adhatoda* had no similar matches for *matK*, but no other species of *Justicia* have *matK* sequences in GenBank. Other studies have indicated that *matK* and *rbcL* are not always useful as barcodes for certain groups of plants (Roy *et al.,* 2010 – *Berberis*; Fu *et al.,* 2011 – *Tetrastigma*).

Table 3. Sequence data from GenBank for 12 species of medicinal plants.

Species	Gene	% Match in GenBank
Acacia nilotica	*matK*	98-99% to multiple *Acacia* spp.
	psbA-trnH	98-99% to 11 vouchers and *Vachellia* (*Acacia*) *farnesiana*
	rbcL	100% to 13 different vouchers
Achyranthes aspera	*matK*	98% to 2 vouchers, 97% to *Pupalia*
	psbA-trnH	100% to 1 voucher, but few *Achyranthes* spp. sequenced
	rbcL	99% to 1 voucher, 99% match to 9 other genera
Amaranthus caudatus	*matK*	98% to 4 *Amaranthus* spp.
Anethum graveolens	*matK*	100% to 1 voucher, 98-99% to *Ammi*, *Apium*, *Foeniculum*, *Petroselinum*, and *Ridolfia*
	rbcL	99% to 1 voucher, 99% match to *Ammi*, *Apium*, *Conium*, *Coriandrum*, *Foeniculum*, *Peucedanum*, and *Petroselinum*
Calotropis procera	*rbcL*	100% to 1 voucher, 99% to many other genera
Carthamus oxyacantha	*matK*	97-98% to *Carthamus*, *Cenaturea*, and *Aegialophila*
Foeniculum vulgare	*matK*	99% to 1 voucher, 99% to *Anethum*, *Apium*, and *Ridolfia*
	psbA-trnH	No other vouchers in GenBank, *Anethum*, *Apium*, etc. not sequenced
	rbcL	99% to 1 voucher, 99% to many other genera
Justicia adhatoda	*matK*	No other vouchers or species of *Justicia* sequenced
Linum usitatissimum	*matK*	No other voucher, 99% to *L. bienne*
	psbA-trnH	100% to 9 vouchers, 99-100% to *L. bienne* vouchers
	rbcL	No other voucher, 99% match to *L. bienne*
Rosa x *damascena*	*rbcL*	98% to other *Rosa* spp.
Trachyspermum ammi	*psbA-trnH*	No other vouchers or species of *Trachyspermum* sequenced
Trigonella foenum-graecum	*matK*	No other voucher, 97-98% to other *Trigonella* spp., *Melilotus* spp.
	psbA-trnH	94% to 1 voucher and to *T. gladiata*

Sequences were downloaded for *matK*, *psbA-trnH*, and *rbcL*. The % match indicates how closely the barcode sequences matched other accessions in GenBank, including other voucher sequences for the same species, if present

This leads to the question of what quality of match is required to use barcodes for identification. A match of 100% between a query sequence and a reference sequence is unambiguous at one level – each base pair is exactly matched. However, if the query sequence is 150 base pairs long, and the reference sequence is 2000 base pairs long, the 100% match might not be as meaningful. The match might be along a part of the gene region that is highly conserved, with little to no variation among many species. Although *rbcL* and *matK* are relatively long (approximately 1430 and 1550 bp respectively), not all portions evolve at the same rate, and submissions of reference sequences to GenBank do not always include the complete gene region. A partial sequence from a less variable portion of a gene may lead to a high match percentage that does not reflect an accurate identification of the query sequence. If a match of less than 100% is accepted for identification, it is important to recognize that a 99% match to a gene region that is 1500 bp long could include 15 mismatches, while a 99% match to a region 150 bp long reflects only a single mismatch. The *psbA-trnH* spacer may be much shorter (200-650 bp, Kress *et al.*, 2005) than *rbcL* and *matK*, so a 99% match using it as a barcode may be more accurate than a 99% match using either of these gene regions. However, variability within barcoding regions from a single species can pose matching problems. Whitlock *et al.*, (2010) demonstrated that the *psbA-trnH* spacer was variable within individual species of *Gentiana* due to the presence of polymorphic inversions. The utility of a given barcoding region needs to be evaluated and confirmed for each different species whose identity is being verified. The data in Table 3 show that while *rbcL* is a good barcode for *Acacia nilotica*, with a 100% match to multiple vouchers (representing different subspecies), it is not useful for distinguishing *Anethum graveolens* and *Foeniculum vulgare*, which are 99% matches to each other and to other genera within Apiaceae. Other gene regions, such *rpl16*, may be more useful for identifying these two species (Downie *et al.*, 2000). The *psbA-trnH* spacer does not appear to be a good barcode for *Acacia nilotica* or *Linum usitatissimum* because it matches a sister species (*Vachellia farnesiana*, *L. bienne*) as well as it matches other voucher sequences for each species. The 94% match between two *psbA-trnH* sequences for *Trigonella foenum-graecum* may represent intraspecific variation or misidentification of one voucher. Multiple voucher sequences should be established for each different barcode, especially for barcoding regions that are known to vary within some species. Despite these challenges, DNA barcoding is proving to be an exciting and powerful tool for identifying and verifying plant specimens.

References

Chen, S., H. Yao, J. Han, C. Liu, J. Song, L. Shi, Y. Zhu, X. Ma, T. Gao, X. Pang, K. Luo, Y. Li, X. Li, X. Jia, Y. Lin and C. Leon. 2010. Validation of the ITS2 region as a novel DNA barcode for identifying medicinal plant species. *PLoS ONE*, 5(1): e8613.

Dick, C.W. and W.J. Kress. 2009. Dissecting tropical plant diversity with forest plots and a molecular toolkit. *BioScience*, 59: 745-755.

Downie, S.R., D.S. Katz-Downie and M.F. Watson. 2000. A phylogeny of the flowering plant family Apiaceae based on chloroplast DNA *rpl16* and *rpoC1* intron sequences: towards a suprageneric classification of subfamily Apioideae. *Am. J. Bot.*, 87(2): 273-292.

Fu, Y.M., W.M. Jiang and C.X. Fu. 2011. Identification of species within *Tetrastigma* (Miq.) Planch. (Vitaceae) based on DNA barcoding techniques. *J. Syst. Evol.*, 49(3): 237-245.

Gao, T., H. Yao, J. Song, Y. Zhu, C. Liu and S. Chen. 2010. Evaluating the feasibility of using candidate DNA barcodes in discriminating species of the large Asteraceae family. *BMC Evol. Biol.*, 10: 324-330.

Hollingsworth, P.M., L.L. Forrest, J.L. Spouge, M. Hajibabaei, S. Ratnasingham, M. van der Bank, M.W. Chase, R.S. Cowan, D.L. Erickson, A.J. Fazekas, S.W. Graham, K.E. James, K.-J. Kim, W.J. Kress, H. Schneider, J. van AlphenStahl, S.C.H. Barrett, C. van den Berg, D. Bogarin, K.S. Burgess, K.M. Cameron, M. Carine, J. Chacón. A. Clark, J.J. Clarkson, F. Conrad, D.S. Devey, C.S. Ford, T.A.J. Hedderson, M.L. Hollingsworth, B.C. Husband, L.J. Kelly, P.R. Kesanakurti, J.S. Kim, Y.-D. Kim, R. Layahe, H.-L. Lee, D.G. Long, S. Madriñán. O. Maurin, I. Meusnier, S.G. Newmaster, C.-W. Park, D.M. Percy, G. Petersen, J.E. Richardson, G.A. Salazar, V. Savolainen, O. Seberg, M.J. Wilkinson, D.-K. Yi and D.P. Little. 2009. A DNA barcode for land plants. *Proc. Nat. Acad. Sci.*, 106(31): 12794-12797.

Hollingsworth, P.M., S.W. Graham and D.P. Little. 2011. Choosing and using a plant DNA barcode. *PLoS ONE*, 6(5): e19254.

Kress, W.J., K.J. Wurdack, E.A. Zimmer, L.A. Weigt and D.H. Janzen. 2005. Use of DNA barcodes to identify flowering plants. *Proc. Nat. Acad. Sci.*, 102(23): 8369-8374.

Lahaye, R., M. Van der Bank, D. Bogarin, J. Warner, F. Pupulin, G. Gigot, O. Maurin, S. Duthoit, T.G. Barraclough and V. Savolainen. 2008. DNA barcoding the floras of biodiversity hotspots. *Proc. Nat. Acad. Sci.*, 105(8): 2923-2928.

Roy, S., A. Tyagi, V. Shukla, A. Kumar, U.M. Singh, L.B. Chaudhary, B. Datt, S.K. Bag, P.K. Singh, N.K. Nair, T. Husain and R. Tuli. 2010. Universal plant DNA barcode loci may not work in complex groups: a case study with Indian *Berberis* species. *PLoS ONE*, 5(10): e13674.

Stoeckle, M.Y., C.C. Gamble, R. Kirpekar, G. Young, S. Ahmed and D.P. Little. 2011. Commercial teas highlight plant DNA barcode identification successes and obstacles. *Sci. Rep.*, 1, 42; DOI:10.1038/srep00042.

Whitlock, B.A., A.M. Hale and P.A. Groff. 2010. Interspecific inversions pose a challenge for the *trnH-psbA* plant DNA barcode. *PLoS ONE*, 5(7): e11533.

Xue, C.Y. and D.Z. Li. 2011. Use of DNA barcode *sensu lato* to identify traditional Tibetan medicinal plant *Gentianopsis paludosa* (Gentianaceae). *J. Sys. Evol.*, 49(3): 267-270.

(Received for publication 14 October 2011)