DNA sequences and cactus classification – a short review

Root Gorelick
Dept. of Biology, Arizona State University, Tempe, AZ 85287-1501, USA (email: cycad@asu.edu).

Summary: There are several pitfalls when using DNA sequence data in classification of cacti or any other plants or animals. DNA sequence data has not and will not result in the "ultimate" cactus classification; it only provides additional characters by which phylogenies can be reconstructed. Only limited sequence data is currently available. Even if complete DNA sequences of all cactus species were known, we still could not fully use this data until we understand the mechanisms and tempos with which DNA sequences evolve.

Zusammenfassung: Die Benutzung von DNA-Sequenzen bei der Klassifikation von Kakteen und anderen Pflanzen bzw. Tieren ist mit verschiedenen Stolpersteinen verbunden. DNA-Sequenzen ergeben keine “letzte und richtige” Kakteen-Klassifikation, und werden das auch nie tun. Sie liefern lediglich zusätzliche Merkmale, mit welchen die Phylogenie rekonstruiert werden kann. Zur Zeit sind Sequzenzdaten nur in beschränktem Umfang verfügbar. Aber auch wenn vollständige Sequenzen für alle Kakteenarten bekannt wären, könnten wir diese Daten trotzdem nicht nutzen, bevor wir nicht über die Mechanismen und Geschwindigkeiten Bescheid wissen, mit denen DNA-Sequenzen sich ändern.

DNA sequences provide useful information regarding cactus classification, but is no magic bullet. In addition to old-fashioned morphological data, DNA sequences provide botanists with another useful and independent character by which they can attempt to infer the evolutionary relationships amongst cacti (Wallace, 1995). DNA sequences change incrementally over evolutionary time. However, they change in ways and at rates that nobody yet fully understands, much as any morphological trait might change over evolutionary time. In this review, I show why DNA sequence data cannot yet – and maybe never will – provide the ultimate classification (*cf*. Whiteley, 2002).

Êhough DNA sequences and other forms of high-tech molecular data have become fashionable in recent years with the advent of relatively cheap and fast technologies, their use to the exclusion of all classical morphological data is short-sighted (Axsmith *et al*., 1998). Why ignore data, especially when it often reflects selective pressures that influenced evolution of a species, genus, family, or other lineage?

Modern plant classifications exist in at least two forms. The first form is one of convenience. Field guides provide an excellent example of such a classification, where plants are first classified by flower colour, second by flower number, third, by flower shape, etc. Such classifications often lump together completely unrelated plants, while separating out very closely related ones. John Ray and Linnaeus produced such classifications. Since Charles Darwin’s time, there has been impetus to have classifications reflect evolution, i.e.: origin by descent, such as family trees. This is the form of classification that I discuss here and is the object of discussion by professional botanists.

Having a classification reflect evolutionary history is a highly non-trivial matter, especially when only having living descendants to examine. Some groups of plants contain fossil representatives, which provide valuable evidence as to which plants had the most recent common ancestors and which ones had more distant common ancestors (this is the key measure of evolutionary relatedness). Fossil evidence is usually only morphological in nature, although rare and very small bits of DNA fragments have been found in very specifically preserved fossilized leaves dating back to the early Miocene, approximately 20 million years ago (Rogers & Bendich, 1985; Golenberg *et al*., 1990). Unfortunately, except for the last 15,000 years (Van Devender, 1990), fossil cacti have never been found (Becker, 1960; Anderson, 2001) – not stems, roots, spines, seeds, nor DNA – leaving botanists in a more tenuous situation when classifying cacti.

Automated sequencing machines have made

DNA sequencing feasible, albeit still expensive. With only limited funding to collect DNA sequences of a wide swath of cactus species, only a few genes have been sequenced. Most funding for DNA sequencing has been for the human genome in the hope that cures for various human diseases can be found. The majority of non-human DNA sequencing has been with common laboratory animals (mice, rats, primates, fruit flies, and one species of nematode) and on agriculturally important crops (rice, maize, tobacco, yeast, and one species of mustard). Cacti take short shrift. Despite this bleak picture, work on cactus DNA sequencing in a few labs has been quite impressive, especially considering the limited level of funding and personnel working on such projects (Wallace, 1995; Hershkovitz & Zimmer, 1997; Porter *et al.*, 2000; Applequist & Wallace, 2001; Hartmann *et al.*, 2001; Nyffeler, 2002).

DNA sequencing should not be confused with DNA fingerprinting, even if both procedures use the same technology. DNA fingerprinting is based on the fact that no two individuals (except possibly identical twins) have the exact same DNA sequences. This allows identification of individuals from forensic evidence left at crime scenes, such as hair or skin flakes. However, it is not obvious how this will help in determining how closely two species are related to each other. The problem is that we do not often know how quickly (over evolutionary time) the DNA sequences of two lineages will diverge from each other.

The rate at which DNA sequences change over time has been likened to the rate at which the hands of a clock move, hence called molecular clocks. However, upon closer examination, molecular clocks based on DNA or protein sequences from a single (paralagous) gene move at varying speeds from one species to another, further confounding classification. Molecular clock speeds also vary between different genes in a single species, even if comparing essentially the same gene, especially when one copy of the gene is in the nucleus and the other is in chloroplasts or mitochondria (see below for details and references). Even some of the best-studied molecular clocks have substantial errors in their timing. For example, molecular clocks place the origin of all modern (metazoan) animal phyla at 1,200 million years ago, whereas the fossil record places their origin at only 600 million years ago (Bromham & Hendy, 2000). Similar problems plague the dating of the origin of flowering plants, where molecular clock data also place their divergence from cone-bearing plants at twice the age estimated from existing fossil data (Martin & Dowd, 1991; Sanderson & Doyle, 2001).

The tempo of evolution varies greatly between lineages, over time within a lineage, and between different traits in a single lineage. This is true for both morphological traits (Simpson, 1944) and molecular characters such as DNA sequences (Vawter & Brown, 1986; Palmer *et al.*, 2000; Zhang *et al.*, 2001). Until we understand how and why these changes in tempo occur and can estimate them, it will be nigh impossible to infer which cacti are most closely related to each other and which are more distantly related. On a more metaphorical level, this is why with our personal experiences of meeting strangers, we often confuse a parent for a grandparent or a cousin for a sibling. On a more technical level, a large portion of my research is on modelling and predicting rates of evolution of DNA sequences, which could eventually play a role in better classification of cacti and any other organisms.

Although DNA sequence data can be extraordinarily useful for classification, the next four paragraphs provide several reasons why DNA sequence data cannot provide an ultimate classification for cacti or any other organisms. Some of these reasons may eventually evaporate once a sufficient number of sequences are obtained, while others will almost certainly remain even if the complete genome is known from every cactus species.

When classifying any group of organisms, it is important to use a large suite of independent characters and to pick ones that will not arise via convergent evolution. Convergent evolution confounds classification by placing unrelated or very distantly related lineages together in the classification. To avoid this problem, botanists typically pick morphological characters that do not seem to provide the organisms with any selective advantage (e.g.: do not use presence of spines or succulent stems as a character for building a classification). Similar advice should be followed when deciding which DNA sequences to use in classification. Researchers should look at a large suite of independent genes (i.e.: not modified copies of a single gene) and only those sequences in which individuals with any given sequence have equal fitness. For this reason, people doing such genetics work often focus on so-called synonymous substitutions and non-coding regions of DNA. Synonymous substitutions occur when multiple DNA nucleotide triplets code for the same amino acid and can yield very different molecular clock dates than do non-synonymous substitutions (Sanderson & Doyle, 2001). By definition, synonymous substitutions are contained in coding regions of the genome. Typically regulatory sequences, which turn on and off coding genes, are also counted as being part of the coding region. Non-coding regions of DNA are parts of the genome that do not get transcribed into messenger RNA and subsequently translated into proteins (and are not regulatory sequences). Non-coding regions of DNA are particularly pervasive in plants, due to the prevalence of transposons, which are duplicate copies of so-called "junk" DNA.

The above paragraph described the ideal situation in which botanists have a rich set of DNA sequences from which to choose. In reality, nowadays, this is far from the case. Very few genes have been sequenced for most genera of cacti, and botanists must simply take what they can get. The most common plant gene for sequencing is *rbc*L (e.g.: Rettig *et al.*, 1992), which is a coding region of the chloroplast genome. ITS (internal transcribed spacer) of nuclear ribosomal DNA is also commonly sequenced (e.g.: Hershkovitz & Zimmer, 1997; Hartmann *et al.*, 2001), but is non-coding; it is transcribed into messenger RNA, but edited out and hence never translated into proteins. Note, however, that there is lingering doubt as to whether different sequences in non-coding regions can confer selective advantages (Eyre-Walker, 1999), even when the sequences are extremely variable, as is the case with ITS. Most sequence data is only for a short portion of the genome, and not for the entire gene or genome segment. It is unknown whether classifications based on *rbc*L and ITS sequences are commensurate with classifications based on other genes or other non-coding regions of the genome. Until we can address these issues, the best that can be done is cross our fingers and assume that these initial pieces of DNA sequence data provide robust evidence regarding evolution by descent.

Further confounding use of DNA sequence data in classifications are the different implications of looking at nuclear versus chloroplast versus mitochondrial genomes. In flowering plants, chloroplast and mitochondrial DNA appears to only be inherited from the mother (this is not universally true; some conifers and mollusks have paternally inherited organelle DNA), whereas nuclear DNA is inherited from both parents. Do we want classifications to be based on only the maternal parents? Do we want classifications to reflect both parents? How can we incorporate hybridization between different species or genera into classifications? Hybrid origins of species and probably genera are reality in cacti (Walkington, 1966; Pinkava *et al.*, 1998; Anderson, 2001), so this last question should not be dismissed lightly. Hybridization causes so-called reticulate evolution and therefore begs for use of nuclear DNA sequences in classification, which means that the most common plant DNA sequence data – from the *rbc*L gene and other portions of the chloroplast genome – is of little use.

A recent classification based on a non-coding segment of chloroplast DNA placed *Blossfeldia liliputana* in its own tribe as the most basal (i.e.: the most ancestral) member of the subfamily Cactoideae (Nyffeler, 2002). Yet, morphological studies show that *Blossfeldia* is probably the most highly derived of all cacti: having fewer stomata than in any photosynthetic terrestrial plant (cacti or otherwise), lacking almost all epidermal and hypodermal protection, and being the only cactus with the ability to become highly desiccated without dying (Barthlott & Porembski, 1996). Why should chloroplast DNA data yield diametrically opposite taxonomic placements for *Blossfeldia* within the Cactoideae when compared with well-accepted classifications based on morphological data? The problem is that *Blossfeldia liliputana* is a hexaploid species (Ross, 1981), i.e.: contains three times the usual diploid number of chromosomes, hence must be of hybrid origin. The maternal parent that produced the first *Blossfeldia* was probably a basal member of the Cactoideae. The paternal parent could have been a highly derived member of the Cactoideae, yet it would not have contributed any chloroplast DNA. *Blossfeldia* beautifully illustrates the problem of using chloroplast DNA to reconstruct phylogenies for species that are of hybrid origin.

When carefully and thoughtfully used, DNA sequence data can be as useful as any character in classifying cacti or any other group of organisms. Unfortunately, we are still severely limited by lack of DNA sequence data and a cogent theory as to how these sequences change over evolutionary time. Therefore, cactus classification will remain in flux for many years to come, but with the hope that we are – with each year – coming closer to estimating the true evolutionary relatedness of taxa (Benton, 2001).

Acknowledgments

Thanks to Colin Walker for providing an invaluable suggestion.

References

Anderson, E. F. (2001). *The cactus family*. Timber Press, Portland.

Applequist, W. L. & Wallace, R. S. (2001). Phylogeny of the portulacaceous cohort based on *ndh*F sequence data. *Syst. Bot.* **26**: 406-419.

Axsmith, B. J., Taylor, E. L. & Taylor, T. N. (1998). The limitations of molecular systematics: a palaeobotanical perspective. *Taxon* **47**: 105-108.

Barthlott, W. & Porembski, S. (1996). Ecology and morphology of *Blossfeldia liliputana* (Cactaceae): a poikilohydric and almost astomate succulent. *Bot. Acta* **109**: 161-166.

Becker, H. F. (1960). Epitaph ? to *Eopuntia douglassii. Cact. Succ. J.* (*US*) **32**: 28-29.

Benton, M. J. (2001). Finding the tree of life: matching phylogenetic trees to the fossil record through the 20th century. *Proc. Roy. Soc., Ser. B. Biol. Sci.* **268**: 2123-2130.

Bromham, L. D. & Hendy, M. D. (2000). Can fast early rates reconcile molecular dates with the Cambrian explosion? *Proc. Roy. Soc., Ser. B. Biol. Sci.* **267**: 1041-1047.

Eyre-Walker, A. (1999). Evidence of selection on silent site base composition in mammals: potential implications for the evolution of isochores and junk DNA. *Genetics* **152**: 675-683.

Golenburg, E. M., Giannasi, D. E., Clegg, M. T., Smiley, C. J., Durbin, M., Henderson, D. & Zurawski, G. (1990). Chloroplast DNA sequence from a Miocene *Magnolia* species. *Nature* **344**: 656-658.

Hartmann, S., Nason, J. D. & Bhattacharya, D. (2001). Extensive ribosomal DNA genic variation in the columnar cactus *Lophocereus. J. Mol. Evol.* **53**: 124-134.

Hershkovitz, M. A. & Zimmer, E. A. (1997). On the evolutionary origins of the cacti. *Taxon* **46**: 217-232.

Martin, P. G. & Dowd, J. M. (1991). Studies of angiosperm phylogeny using protein sequences. *Ann. Miss. Bot. Gard.* **78**: 296-337.

Nyffeler, R. (2002). Phylogenetic relationships in the cactus family (Cactaceae) based on evidence from *trnK/matK* and *trnL-trnF* sequences. *Amer. J. Bot.* **89**: 312-326.

Palmer, J. D., Adams, K. L., Cho, Y. R., Parkinson, C. L., Qiu, Y. L. & Song, K. M. (2000). Dynamic evolution of plant mitochondrial genomes: mobile genes and introns and highly variable mutation rates. *Proc. Nat. Acad. Sci.* **97**: 6960-6966.

Pinkava, D. J., Rebman, J. R. & Baker, M. A. (1998). Chromosome numbers in some cacti of western North America - VII. *Haseltonia* **6**: 32-41.

Porter, J. M., Kinney, M. S. & Heil, K. D. (2000). Relationships between *Sclerocactus and Toumeya* (Cactaceae) based on chloroplast *trn*L-*trn*F sequences. *Haseltonia* **7**: 8-23.

Rettig, J. H., Wilson, H. D. & Manhart, J. R. (1992). Phylogeny of the Caryophyllales: gene sequence data. *Taxon* **41**: 201-209.

Rogers, S. O. & Bendich, A. J. (1985). Extraction of DNA from milligram amounts of fresh, herbarium and mummified plant tissues. *Plant Mol. Biol.* **5**: 69-76.

Ross, R. (1981). Chromosome counts, cytology, and reproduction in the Cactaceae. *Amer. J. Bot.* **68**: 463-470.

Sanderson, M. J. & Doyle, J. A. (2001). Sources of error and confidence intervals in estimating the age of angiosperms from *rbcL* and 18S rDNA data. *Amer. J. Bot.* **88**: 1499-1516.

Simpson, G. G. (1944). *Tempo and mode in evolution.* Columbia University Press, New York.

Van Devender, T. R. (1990). Late Quaternary vegetation and climate of the Sonoran Desert, United States and Mexico. Pages 134-163 in J. L. Betancourt, Van Devender, T. R. & Martin, P. S. *Packrat middens: the last 40,000 years of biotic change.* University of Arizona Press, Tucson.

Vawter, L. & Brown, W. M. (1986). Nuclear and mitochondrial DNA comparisons reveal extreme rate variation in the molecular clock. *Science* **234**: 194-195.

Walkington, D. L. 1966. Morphological and chemical evidence for hybridization in some species of *Opuntia* in southern California. Ph.D. dissertation. Claremont Graduate School, Claremont.

Wallace, R. S. (1995). Molecular systematic study of Cactaceae: using chloroplast DNA variation to elucidate cactus phylogeny. *Bradleya* **13**: 1-12.

Whiteley, D. (2002). What happened to "genetic fingerprinting" for the Cactaceae. *Brit. Cact. Succ. J.* **20**: 27.

Zhang, L. Q., Pond, S. K. & Gaut, B. S. (2001). A survey of the molecular evolutionary dynamics of twenty-five multigene families from four grass taxa. *J. Mol. Evol.* **52**: 144-156.