Molecules under the microscope

Clive Stace looks at how DNA has altered the way we classify plants and argues it is a change for the better

RHS LINDLEY LIBRARY

Plate, dated 1793, from *Linnaeus's System of Botany* by William Curtis showing groups of related plants.

THE DESIRE TO CLASSIFY, whether people alphabetically in a telephone directory, elements according to atomic number in a periodic table, or living organisms in some taxonomic system, is universal, and a fundamental feature of human nature. The process is fully justified because it ensures retrieval of components and information far more easily and quickly than can random searching. It achieves order out of chaos. With regard to plants, ancient man probably employed primitive systems of classification in order to communicate, since plants were (and are) vital as sources of food and medicine, but the earliest written examples are those of the Greeks, dating back at least 2300 years. This can be seen as the start of the first of the three great phases of plant classification: phenetic or artificial; phyletic or natural; and molecular.

The earliest attempts that we know of used the most obvious features to classify plants, such as growth habit and gross features of flower structure and fruit types. More characters were utilized as finer distinctions were required, but even in the sixteenth century these same characteristics still held pride of place. The trend to increase the range of characters employed was bucked in the seventeenth century by Linnaeus, whose 'sexual system' can be considered the culmination of the artificial systems. It was based mainly on the number of stamens and pistils. Such systems are considered artificial because the most similar species or genera are not necessarily classified together. For example, Linnaeus' class *Diandria* (two stamens) included *Anthoxanthum* (*Poaceae*), *Syringa* (*Oleaceae*), *Salvia* (*Lamiaceae*) and *Circaea* (*Onagraceae*). These all share two stamens, but have scarcely anything else in common. In other words, saying a plant belongs to the *Diandria* tells one nothing more about it – an artificial classification is not predictive of further characters.

Linnaeus realized his system was artificial, as had others before him and everyone after him, and in 1764 he listed 58 'natural orders' (what we now call families), many of which are recognisable as families that we still accept, e.g. *Umbellatae* are largely our *Apiaceae*. These taxa (natural orders) are far more predictive than the classes in his sexual system; learning that a plant belongs to the *Umbellatae* immediately tells one a lot about that plant, without one ever having seen it let alone having investigated those characters in it. The main aim of taxonomists in the nineteenth and twentieth centuries has been to construct ever more predictive classifications, and in my view this sole criterion should be the primary judge of how good (useful) a particular system of classification is.

Once the more obvious (exomorphic) structural features of plants had been investigated, botanists increasingly turned their attention to cryptic (less obvious or even invisible) characters in order to achieve greater predictivity. These included vegetative, floral and fruit anatomy (often concentrating on particular aspects such as epidermis, trichomes, pollen, timber or leaf vasculature), phytochemistry, chromosomes and patterns of hybridization. All of these fields of investigation have provided many new insights

into plant relationships, and have therefore led to improved (more predictive) classifications. And we must not forget that these data are still of enormous importance in botanical investigation, and should not be abandoned just because a newer or more fashionable field of study has emerged. If taxonomists of that generation dared to believe that one day they would hit upon the magic data-set that revealed the ultimate über-predictive classification, they would have been frustrated. Such a revelation came only with the advent of molecular classifications.

The molecular revolution

Even before Darwin's time it was realized that patterns of past evolution must bear an intimate correlation with present similarity. Taxa that separated relatively recently will be more similar than those whose divergence was more ancient. The rise in popularity of phylogenetic systems can be seen as an obvious extension of the concept of natural classifications. But the problem was how to deduce the past evolutionary patterns and, with reliance on the same data-sets as earlier workers, no certain solution was reached. But once DNA was utilized, nowadays usually via the base sequence of very small portions of DNA, the ultimate goal, so eagerly sought by previous generations, had been attained. There can be no doubt about this. DNA sequences are indisputably a true measure of the course of evolution, and therefore of relationships. If our 'traditional methodology' (phenetics, phyletics) were an accurate predictor of relationships, the molecular methodology would simply confirm it. Where the molecular system contradicts the traditional one the phenetic characters have misled us, or we have misinterpreted them, or we have misconstrued the molecular evidence. The Angiosperm Phylogenetic Group (APG) system (based on DNA sequences) of family classification, now universally adopted, is certain to endure for centuries to come, as it will prove to be the most highly predictive system. These assertions are surely incontrovertible.

The main advantages and disadvantages of the molecular approach are listed in Table 1.

Examples of major taxonomic changes in the British flora demanded by adopting a molecular classification are the amalgamation of *Anagallis*, *Centunculus*, *Glaux* and *Trientalis* into *Lysimachia* (loosestrifes), and the splitting of *Gnaphalium* (cudweeds) and *Chenopodium* (goosefoots) each into five segregate genera. Such radical changes to our system of classification are not welcome, but are inescapable. Even greater consternation arises in those (so far rather few) cases where genera must be considered separate on molecular grounds, but are not distinguishable on morphological characters. Examples, again from the British flora, are the ragworts *Senecio* and *Jacobaea*, the orchids *Orchis* and *Anacamptis*, and the vetches *Vicia* and *Ervilia*. In such cases the generic attribution of a new species could not be made without a DNA analysis. The previous classification (e.g. *Jacobaea* included in *Senecio*) arose because the gross features of these plants over-rode the less obvious but vastly more numerous characters indicating the true relationships. Unfortunately, we simply have to learn to live with these problem taxa.

Limitations of molecular data

Despite the above assertions, which are based on the overwhelming reliability of molecular data, and on the illogicality of ignoring the latter when the going gets tough, DNA does not provide all the answers in every case. Molecular data cannot be applied in a mechanical, inflexible way, but taxonomic judgement is still important, and is vital in many cases. I want to mention five situations.

Incongruity

This is actually simply a euphemism for 'something's gone wrong'. It is a term used when two analyses, using different parts of DNA, produce different data sets leading to different classifications. It is »

Table 1. Positive and negative aspects of the adoption of a DNA-based classification.

GOOD IDEA	BAD IDEA
Not affected by environment	Characters are cryptic
Not affected by gene expression	Very expensive
Constant across all life-forms	Dependent on high level of expertise
Will not change with time unless faulty data-collection or -manipulation has occurred	Some radical taxonomic changes, some of which are 'unacceptable' by some users

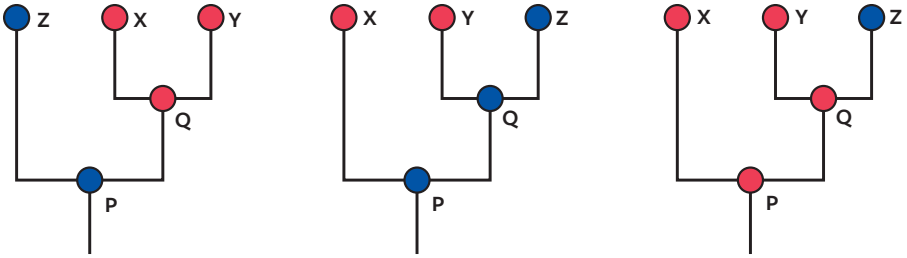

Fig. 1 Patterns of relationship revealed by DNA analysis.

most commonly encountered when nuclear and chloroplast DNA data are compared. However, various reasons for such differences are well understood, and there can be little doubt that after investigation in each case these will be ascertained and the true classification will be revealed. This is our experience so far. Apparent incongruity can also emerge when the facts have been misrepresented; perhaps the sampling was inadequate, or the data-analysis was faulty. The remedy in these situations is obvious, and it is a warning to taxonomists that they should not adopt new classifications until their molecular basis has been fully investigated and understood.

'No difference'

There are several cases in the literature where DNA analyses have not revealed any difference in the base sequences of two similar species. But the conclusion that such pairs of species are molecularly identical and must be amalgamated, which has been expressed by some taxonomists in the past, is surely erroneous. Only a tiny fraction of the DNA has ever been sequenced, so we have no idea of the total level of difference between the two species. Examples from the British flora for which this has been claimed are the butterfly-orchids *Platanthera chlorantha* and *P. bifolia*, and the gentians *Gentianella amarella* and *G. anglica*. In each case the two species clearly *are* different, despite what the incomplete results of DNA analysis suggest; very similar, evidently, but identical, no.

Level of discrimination

Despite some assertions to the contrary, DNA sequences (or any other taxonomic characters) do not determine the taxonomic level at which differences should be recognized. Molecular (or any other) data tell us the scope of a taxon, but not its rank. A good example is furnished by the two well-known emergent aquatics *Typha* (bulrush) and *Sparganium* (bur-reed). These two genera have no close relatives and are often placed in two separate monogeneric families, *Typhaceae* and *Sparganiaceae*. Molecular data confirm that they have a common ancestor, which is shared by no other genera. Hence in the derived cladogram they are separated by a simple bifurcation. Whether one places these two genera in one or two families (or, indeed, in one or two genera) is a matter of subjective judgement. In other words, the concept of splitters and lumpers is as relevant today as it ever was. The authors of the APG system state that they have opted for the lumping approach, and so recognise the *Typhaceae* with two genera, but the two-family alternative is equally valid as it is equally supported by the molecular evidence. Another example in the British flora is the amalgamation or separation of *Lamium* (deadnettles) and *Lamiastrum* (yellow archangel), and on a world scale the splitting or not of the *Boraginaceae* into 12 separate families.

The paraphyly conundrum

This topic remains the most controversial area of disagreement in molecular taxonomy. Figure 1 shows the three main patterns of relationship revealed by DNA analysis. The red and blue circles at the top level are present-day species; those in the lower levels are their ancestors. Red and blue represent two different sets of morphological characters that have led in the past to the recognition of two genera (red and blue).

In the left-hand cladogram species X and Y are monophyletic, i.e. they represent all the products of a single ancestor (Q). Such groups are considered the ideal taxon; their constituents (X and Y) are more closely related and more similar to each other than either is to any other species.

In the central cladogram X and Y are

polyphyletic, i.e. they do not have a common ancestor sharing the same characters. X and Y evolved their red-ness by separate evolutionary routes from different ancestors (P and Q). Although X and Y resemble each other in red-ness, Y is more closely related to Z than to X and is likely to share with it more characters (although not red-ness!). Polyphyletic taxa are not acceptable, because they would significantly reduce predictivity. The two ways in which the polyphyletic classification could be rendered monophyletic are either by recognizing X, Y and Z as three separate genera, or by placing all three into one genus with the common ancestor P. In real situations a decision is often needed as to which alternative to adopt. Space does not permit discussion of this here, but in some cases (e.g. the mallows, *Malva* / *Lavatera*) lumping seems more useful, and in others (e.g. the goosefoots, *Chenopodium*) splitting has been favoured. This dilemma was recently discussed more fully by Alan Paton in this journal in relation to the amalgamation or not of *Rosmarinus* (rosemary) and *Perovskia* (Russian sage) into *Salvia* (sage); predictably, for a Kew botanist, he opted for lumping. This is another example of molecular evidence not providing the definitive answer to a taxonomic problem; taxonomic judgement is also required.

In the right-hand cladogram X and Y are paraphyletic, i.e. they represent products of a single ancestor (P), but not all of the products of P, which, by the evolution of blue-ness, also gave rise to Z. Species Y of the paraphyletic genus XY is more closely related to species Z than it is to species X. Opinion is divided over whether or not paraphyletic taxa should be accepted. From the foregoing the basic arguments for and against are obvious, and both are persuasive. Here I want to present my reasons for strongly supporting the 'for' argument.

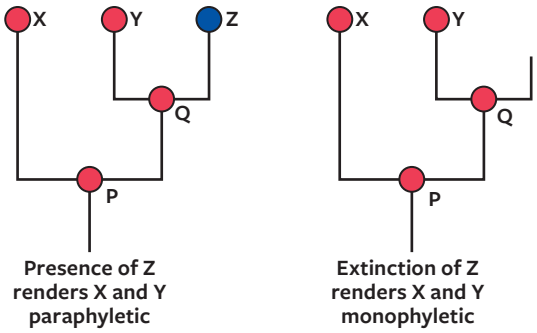

Fig. 2 The potential effect of extinction on paraphyly.

In Figure 2 the paraphyletic cladogram from Figure 1 is repeated on the left. Evolution certainly involves the formation of new species, but equally significant is extinction. We have no idea how common extinction has been in the past, but there is absolutely no reason to believe it has been rare. In the right-hand cladogram species Z has become extinct; by this one event the genus XY has been converted from a paraphyletic to a monophyletic state. Quite apart from the existence or not of species Z, the relationships and similarity / differences between X and Y are identical whether they are monophyletic or paraphyletic. We have no way of telling whether an apparently monophyletic group has always been so since its origin, or was once paraphyletic but has been rendered apparently monophyletic by extinction(s). For that reason I think that it is inescapable that paraphyletic taxa should be acceptable. I would add 'in certain circumstances' (again, space does not permit discussion), because often it is more convenient to amalgamate the offending taxon (taxa) to form a monophyletic group (see under *Salvia* and *Malva* above). But where the offending taxon is well demarcated by distinctive features, or has some particular ecological or chemical or economic characteristics, or contains a large number of species, or has been recognized as distinct for centuries, then I contend that there should be no reason not to recognize it as a distinct taxon. A very topical example is the grass genus *Spartina* (cord-grass). This is monophyletic, but if removed from *Sporobolus* the latter is paraphyletic. But *Spartina* is very distinctive, and very important ecologically and economically, and there is a large body of opinion calling for its continued recognition.

The problem with polyploids

Polyploids are not suited to cladistic analyses of DNA base sequences. The problem is not a small one. The proportion of angiosperms that are polyploid is uncertain, and estimates vary (with the use of more sophisticated techniques the figure is tending to rise); it is likely that around 50% of angiosperms have a polyploid origin. Cladistic analysis involves the construction of cladograms, which are branching 'family trees' showing the position of past evolutionary divergences.

However, polyploidy equally involves convergences, where two lines have come together again via hybridization. When two taxa (with different base sequences in the portion of DNA being analysed) hybridize, both sequences are represented in the F1 hybrid. This is frequently observed in recent »

hybrids (especially those made artificially), but in ancient hybrids, particularly allopolyploids where the chromosome number has doubled, it is often found that only one sequence persists. This is thought to be due to concerted evolution, whereby one of the two sequences is gradually excluded (homogenization). One study has indicated that, in the bittercress *Cardamine* (*Brassicaceae*), over a period of 70 years, 20 out of the 23 different base positions had become homogenized. It seems that homogenization more often takes place to the sequence of the female parent, but not always, and there is evidence that different individuals of some allopolyploid (amphidiploid) species can homogenize in different (male or female) directions. Homogenization to the female parent would receive 'confirmation' of the position in the cladogram from a chloroplast analysis, because chloroplasts are inherited from the female parent. If base sequence analysis picks out only one sequence in a polyploid, the cladogram will place that taxon close to its female parent, and not to its true intermediate position. Many analyses in the literature must have falsely claimed the position of polyploids.

In the past some taxonomists have proposed that, in complex polyploid situations, generic limits should be defined by genome constitution, e.g. AB, AC, AD and BC polyploids should be placed in separate genera. Áskell Löve's name is particularly associated with such a move in the wheat tribe *Triticeae* (*Poaceae*). Perhaps base sequences could indicate genomic homologies more clearly than traditional genome analysis. But in large, ancient and complex polyploid groups the distribution of the haploid genomes is so diffuse and exists in so many combinations that a great many genera would need to be accepted, some of them unrecognizable morphologically. Moreover, some species that would be placed in different genera are very closely related. For example, the two British couchgrasses *Elymus repens* and *E. atherica*, which are very similar, often hybridize, and are often misidentified even by experienced field botanists, have different genomic constitutions and would demand different genera if based on strictly molecular data.

There is always, of course, the alternative solution, i.e. amalgamation. This has been proposed formally in the tribe *Maleae* (*Pyrineae*) of the *Rosaceae*, where well in excess of 1,000 species, formerly belonging to such genera as *Malus*, *Pyrus*, *Sorbus*, *Cotoneaster*, *Photinia*, *Stranvaesia*, *Aronia* and *Chaenomeles*, have recently been lumped into the genus *Pyrus*. A similar solution has been (so far informally) proposed for the *Triticeae*. This would involve genera as diverse as *Agropyron*, *Elymus*, *Hordelymus*, *Leymus*, *Secale*, *Aegilops* and *Hordeum* (wheats, couches, barleys, ryes) all being transferred into a huge genus *Triticum*. It must be admitted that intergeneric hybrids are rather common in both these groups (e.g. × *Sorbopyrus* and × *Agrohordeum*), but I can see no merit in creating such vast and unwieldy genera, which would require careful construction of useful infrageneric classifications.

CONCLUSIONS

Molecular taxonomy based on base sequences of DNA is a fantastically valuable tool that has realized the taxonomist's dream – a data-set that reliably interprets the past evolutionary history of species. Since the evolutionarily most closely related species must, by definition, have the greatest number of characters in common, expression of these data in a classification is mandatory, for it produces a very highly predictive system. The advantages listed in Table 1 are not mirrored in any other data-sets (such as morphology), and so our belief in and primary reliance upon molecular classifications must be unequivocal.

Nevertheless, some taxonomists remain unswayed by molecular evidence when it disagrees with traditional morphology-based classifications. If our classifications are to be rigorous and meaningful, however, we cannot tolerate polyphyletic genera or families; in such taxa morphology has misled us into accepting false relationships. The opposite extreme to an entirely morphology-based classification is a cladistically extreme molecular-based classification, where, for example, all our cotoneasters become pears. In my opinion both extremes, each championed by dogmatists who wish to impose their beliefs rigidly without much concern for the outcome, and without a defensible code of taxonomic practice, are unacceptable. Compromise is needed. Probably the greatest number of contentious cases, causing arguments between the two groups of dogmatists, involve paraphyly. I hope that I have made it clear above that there is no genetic (or other) reason not to recognize paraphyletic taxa, but that sometimes their acceptance produces the most satisfactory outcome. Sometimes the wisdom of their acceptance or rejection is equivocal; there is still plenty of room for healthy debate and taxonomic judgement. ○

Professor Clive Stace is author of the *New Flora of the British Isles* and numerous other botanical works.