J. theor. Biol. (1996) **183,** 255–267

Statistical Tests for Discrete Cross-species Data

ALAN GRAFEN AND MARK RIDLEY

*Department of Plant Sciences, South Parks Road, Oxford, OX*1 3*RA, U.K. and the Departments of Anthropology and Biology, Emory University, Atlanta, Georgia* 30322, *U.S.A.*

(*Received on* 16 *December* 1994, *Accepted in revised form on* 25 *April* 1996)

Four methods have been proposed that can be used to test for associations between the states of discrete characters in cross-species data and that do not suffer from non-independence due to overcounting of data points. The tests are those of Ridley (1983), Burt (1989), Grafen (1989), and a new test called the ICDE test. The aim of the paper is to measure the Type I error rates for these methods with simulated null distributions of discrete characters. The null data is generated by a model of discrete character evolution, using three shapes of phylogeny: tetratomous, dichotomous, and realistic. Ridley's and Burt's tests are both reasonably valid with the realistic phylogeny but biased with the tetratomous and dichotomous phylogenies. Grafen's phylogenetic regression is reasonably valid with all tree shapes. One version of the ICDE test was valid, the other less so. The invalid results are explained in terms of two kinds of statistical non-independence that arise in discrete data: non-independence due to the reconstruction of character states by parsimony, and the "family problem" in which similar patterns are found in null data in many separate radiations because all the radiations began from the same ancestral state.

1. Introduction

A number of methods have been proposed to test for associations between the states of discrete characters in cross-species data. An analysis of their statistical properties requires, as Ridley & Grafen (1996) argued, measuring their Type I and Type II error rates. We (Grafen & Ridley, 1996a) have devised a model of discrete character evolution that can generate null and non-null simulated datasets and the present paper will use simulated null datasets to look at the validity of four proposed comparative methods for discrete data. Some other methods have also been proposed but can be ruled out, as Grafen & Ridley (1996a) discussed, because they find more datapoints in radiations with more species than in radiations with less species; they therefore suffer from non-independence, of a kind analogous to pseudoreplication. The work that follows confines itself to the simplest case of two discrete characters with two states each (A/a and B/b).

2. The Tests

2.1. SPECIES COUNTS

We looked at the significance of naive species counts, using a chi-squared calculated from the number of species with each pair of character states. The results provide a useful point of comparison with the other methods. The test was performed by a Mathematica program.

2.2. THE INDEPENDENT CHARACTER EVOLUTION (ICE) TEST

This is the test proposed by Ridley (1983) and formalised by Grafen & Ridley (in prep. a). We shall call the test the independent character evolutions test or ICE test, because it is based on inferring the locations in the phylogenetic tree at which characters changed. The test begins with a known distribution of character states among the species, and a working phylogeny (Grafen, 1989) of those species. The first step is to reconstruct what we call the "character change tree"; a contingency table test (χ^2, G-test, or

Fisher exact) is then performed on the character changes recognized in that tree. The character change tree is reconstructed by parsimony; it is a tree in which every uniform region in the phylogeny is collapsed into a single node: the result is a tree showing only the character states that there have been changes among. Any one dataset usually has more than one equally parsimonious reconstruction, and more than one character change tree. In this study we dealt with the problem by calculating z for each reconstruction, averaging z, and then finding the p-value for that z (see Section 4 below). Ridley (1983) and Grafen & Ridley (in prep. a) provide further details about the method. It was implemented by a program written in Mathematica (Wolfram, 1988).

2.3. THE PHYLOGENETIC REGRESSION

Grafen (1989) originally devised and justified the phylogenetic regression for continuous characters; but it can be applied to discrete characters by entering them in 0/1 form in the regression. The method regresses the values of the two characters on each other within each radiation in the phylogeny and combines the regressions into the "phylogenetic regression"; each radiation contributes one datapoint. Grafen (1989) should be consulted to see how the method works. Because the justification there is for continuous characters, it is uncertain whether it will retain its validity with discrete data; the simulations below will reveal whether it does in particular cases. The phylogenetic regression is implemented by a program written in GLIM (Numerical Algorithms Group, 1987).

2.4. BURT'S METHOD

Burt (1989) was also mainly concerned with continuous characters. He provides, for continuous characters, a valuable method that makes no assumptions about branch lengths, and which (with the replacement of the sign of a covariance with the sign of a Spearman rank correlation) would be a truly nonparametric method for two continuous variables and a phylogeny with unresolved polytomies. Although the extension to discrete characters is only a brief and incomplete suggestion in the original paper, the logic and procedure are clear enough. In Burt's words, "phylogenetically independent contrasts are again identified, with the proviso that each one contains both values for both variables." We start at the terminal taxa, or tips, of the phylogeny and work up until we reach a node below which both characters vary. The species below that "contrast" node contribute only to that node and are excluded from other comparisons. Burt draws a path connecting the species below the contrast node, and defines nodes as "phylogenetically independent" if their "paths do not cross at any point."

Once the nodes providing independent contrasts have been found, Burt's test lists the contingency tables below the nodes, finds the sign of each, and executes a sign test. There are two ways to find the sign. Burt counted the numbers of the species in the contingency table: there are four numbers, n_{AB}, n_{aB}, n_{Ab}, n_{ab}, for two binary characters. He then found the sign of each contingency table by the sign of ($n_{AB}n_{ab}$-$n_{Ab}n_{aB}$); this is the first method of finding the sign. It uses raw species numbers and there may be a danger that large uniform blocks of phylogenetically related species will be overweighted. The randomisation scheme is designed to ensure the validity of the test, but this extra weighting might be expected to lead to reduced power. We therefore used a second method, in which all non-zero entries in the contingency table were reduced to one; these reduced values of n_{AB}, n_{aB}, n_{Ab}, n_{ab} are then all either one or zero. The sign can then again be found using the sign of ($n_{AB}n_{ab}$-$n_{Ab}n_{aB}$). The two procedures are identical except where there are non-zero entries in all four corners of the contingency table (and at least one of them is greater than one). Then Burt's method can have any sign, +, 0 or − according to the four numbers, whereas the method we used, that reduces all the non-zero values to one, will have sign 0). As things turned out in this study, it would almost certainly have made no difference which method we used, because the simulated datasets contained hardly any contingency tables with sign 0. The test was implemented by a Mathematica program.

2.5. THE INDEPENDENT CHARACTER DIFFERENCE EVOLUTIONS (ICDE) TEST

The ICDE test is described by Grafen & Ridley (in prep. b). It treats the two characters separately and for each character traces back from the terminal taxa until it reaches the nodes below which the character varies. The variation below each of these nodes has evolved independently. A contingency table is compiled for the character states below each such node. There are two ways of compiling the table, rather like the two ways of finding the contingency table sign in Burt's test. The two result in two versions of the ICDE test. The contingency table for a node initially contains the numbers of species below the node that have the four pairs of character states. One way is to retain the contingency tables with numbers of species; this is the "CSP" (for "counting species") version of the ICDE test. The other way is to reduce all non-zero values in the contingency tables to one;

this is the "C1" (counting one) version of the ICDE test. The motivation behind the C1 method is not to give too much weight to large uniform taxa. Note, however, that the randomisation procedure (to be discussed shortly) is designed to avoid phylogenetic overcounting, in both CSP and C1. These contingency tables are then added together, and a test statistic calculated from that summed table. A randomisation test is then performed to find the significance of that test statistic. The randomisation is applied independently to the separate contingency tables, and independently to the two characters. With equal probabilities it either switches or does not switch the two character states of a character in the contingency table for a node. These randomised contingency tables are summed and the test statistic calculated to provide a distribution of values with which the observed test statistic can be compared. We have obtained Type I error rates for both versions of the ICDE test. The main feature of the method is that it exploits the differences in the character states below the nodes, without attempting to reconstruct the ancestral character states at or above these nodes. This contrasts with the ICE test and other tests such as Maddison's (1990), which rely on reconstructed ancestral states through the tree. In this study the ICDE test was implemented using another GLIM program.

3. Null Data Generation

We looked at the behaviour of the methods when they analysed simulated null data. We generated null data using Grafen & Ridley's (1996a) model of discrete character evolution. This section gives details of the parameters, which are of technical interest. The general reader only needs to know that null data with a phylogenetic structure was generated: however, such a reader might note the main feature of our model (next paragraph), the shapes of phylogeny (next but one paragraph), and perhaps the realism of the datasets (penultimate paragraph, at end of section).

Grafen & Ridley's (1996a) model produces a pattern of states for two observed characters (A and B). The states of A and B, however, are controlled by evolution in two unobserved characters (C and D); the state of C determines the state of A and the state of D determines that of B. In the null case there is no causal relation between the unobserved characters and C and D each evolve through the phylogeny in the manner of a discrete branching Markov process in continuous time. The importance of the hidden variables is that they allow the probability of change in A/a and in B/b to vary between different regions of the tree. Read & Nee (1995) and Grafen & Ridley (1996a) argue this feature is biologically required. The feature is realized in the model by means of multiple states in the unobserved characters corresponding to each of the two states of the observed characters.

We studied three shapes of phylogeny. The number of species in each was 256. The three shapes are as follows. (i) tetratomous: the 256 (4^4) species are arranged in groupings of four at four hierarchical levels. (ii) dichotomous: all branches are dichotomous. The shape is a compatible dichotomous refinement of the tetratomous tree, obtained by the method described in Grafen & Ridley (1996a) for data generation. The tree is symmetrical on a broad scale but the resolution of each tetratomy may be asymmetrical or symmetrical. The exact tree is given in the Appendix to this paper. (iii) "Hennig": the phylogeny was intended to have a relatively realistic amount of asymmetry and proportions of dichotomous and polytomous nodes. It was abstracted from Hennig's (1981) phylogeny of the insects: a level that had approximately 256 taxa was selected and the "Hennig" phylogeny in this paper is the branching pattern from that level back to the common ancestor of the insects. Some manipulation was required to produce a tree with exactly 256 species. It is not intended to correspond precisely to any one real phylogeny but merely to be approximately realistic; the perfectly symmetric tetratomous phylogeny is unrealistic and the perfectly dichotomous phylogeny implies a precision of knowledge that is rarely available. The Hennig tree we used is also given in the Appendix.

In the specific version of the model used in this study, both C and D had six states although A and B had only two each; this enables the multiple determination of the observed character states. The transition matrix was the matrix exponential of

$$\begin{pmatrix} -0.01 & 0.01 & 0.0 & 0.0 & 0.0 & 0.0 \\ 0.2 & -0.82 & 0.6 & 0.02 & 0.0 & 0.0 \\ 0.0 & 0.6 & -0.62 & 0.02 & 0.0 & 0.0 \\ 0.0 & 0.0 & 0.02 & -0.62 & 0.6 & 0.0 \\ 0.0 & 0.0 & 0.02 & 0.6 & -0.82 & 0.2 \\ 0.0 & 0.0 & 0.0 & 0.0 & 0.01 & -0.01 \end{pmatrix}$$

Each entry may be read as a hazard rate, or instantaneous proportional rate of change. Thus, the

Table 1

Frequency distributions for the significances of the six tests, with null data simulated through three tree shapes

P-values from	to	Tetratomous			Dichotomous			Hennig			E(360)
Species											
0	0.001		33			22			61		0.36
0.001	0.01		21			20			15		3.24
0.01	0.025		15			8			18		5.4
0.025	0.05		17			13			9		9
0.05	0.1		16			17			17		18
0.1	0.25		29			33			19		54
0.25	0.75		103			109			63		180
0.75	0.9		42			52			23		54
0.9	0.95		22			26			18		18
0.95	0.975		16			18			13		9
0.975	0.99		16			16			18		5.4
0.99	0.999		15			12			33		3.24
0.999	1		15			14			53		0.36
Totals:			360			360			360		360.00
The ICE test											
		min	mu	max	min	mu	max	min	mu	max	
0	0.001	0	0	0	1	1	1	1	1	1	0.36
0.001	0.01	0	2	5	1	1	1	3	4	6	3.24
0.01	0.025	0	2	5	0	0	0	6	6	10	5.4
0.025	0.05	3	2	5	6	6	6	7	9	14	9
0.05	0.1	2	4	9	3	3	4	16	21	24	18
0.1	0.25	8	11	14	6	6	5	43	54	60	54
0.25	0.75	24	37	46	27	29	31	173	178	166	180
0.75	0.9	20	25	36	32	31	29	57	47	48	54
0.9	0.95	20	23	20	27	26	26	25	18	16	18
0.95	0.975	13	16	33	30	30	30	13	10	5	9
0.975	0.99	22	26	37	29	30	30	9	8	7	5.4
0.99	0.999	63	82	70	78	77	77	7	4	3	3.24
0.999	1	185	130	80	120	120	120	0	0	0	0.36
Totals:		360	360	360	360	360	360	360	360	360	360.00
Phylogenetic regression											
0	0.001		0			0			2		0.36
0.001	0.01		3			4			4		3.24
0.01	0.025		8			1			6		5.4
0.025	0.05		9			13			4		9
0.05	0.1		15			20			14		18
0.1	0.25		62			48			38		54
0.25	0.75		150			190			236		180
0.75	0.9		61			50			39		54
0.9	0.95		26			17			3		18
0.95	0.975		17			12			2		9
0.975	0.99		5			3			7		5.4
0.99	0.999		4			2			2		3.24
0.999	1		0			0			3		0.36
Totals			360			360			360		360.00
Burt's test											
		min		max	min		max	min		max	
0	0.001	0		0	0		0	0		12	0.36
0.001	0.01	0		1	0		1	0		8	3.24
0.01	0.025	0		0	1		2	4		13	5.4
0.025	0.05	0		2	0		1	7		7	9
0.05	0.1	1		11	2		5	10		25	18
0.1	0.25	9		26	4		22	32		66	54
0.25	0.75	71		104	78		123	170		178	180
0.75	0.9	42		94	60		69	65		31	54
0.9	0.95	60		43	41		55	28		15	18
0.95	0.975	36		29	32		28	10		4	9
0.975	0.99	47		24	60		31	16		0	5.4
0.99	0.999	47		24	48		14	4		1	3.24
0.999	1	47		2	34		9	14		0	0.36
Totals		360		360	360		360	360		360	360.00

continued overleaf

Table 1—*continued*

P-values from	to	Tetratomous	Dichotomous	Hennig	E(360)
The ICDE test: *C*1 *version*					
0	0.001	0	0	0	0.36
0.001	0.01	0	0	4	3.24
0.01	0.025	2	3	6	5.4
0.025	0.05	9	5	8	9
0.05	0.1	12	6	14	18
0.1	0.25	30	47	58	54
0.25	0.75	147	175	162	180
0.75	0.9	61	61	66	54
0.9	0.95	35	31	21	18
0.95	0.975	22	18	15	9
0.975	0.99	17	8	5	5.4
0.99	0.999	23	5	0	3.24
0.999	1	2	1	1	0.36
Totals		360	360	360	360.00
The ICDE test: *CSP version*					
0	0.001	1	0	0	0.36
0.001	0.01	3	1	3	3.24
0.01	0.025	3	3	5	5.4
0.025	0.05	4	7	13	9
0.05	0.1	17	16	11	18
0.1	0.25	67	51	54	54
0.25	0.75	196	181	186	180
0.75	0.9	44	81	58	54
0.9	0.95	15	14	17	18
0.95	0.975	7	3	6	9
0.975	0.99	3	1	6	5.4
0.99	0.999	0	2	1	3.24
0.999	1	0	0	0	0.36
Totals		360	360	360	360.00

There were 360 datasets for each tree shape, and the expected numbers for random data are given in the E(360) column. Higher numbers in the lower half of the distribution mean the test is biased against the ancestral state: the test finds too many results on the non-ancestral diagonal.

chance per unit time that a lineage in state 1 (top row) leaves state 1 is proportional to 0.01 (the diagonal element is 0.01, and is negative because it is a chance of leaving that state). When it does leave state 1, the lineage joins state 2. The chance that a lineage in state 2 (second row) leaves state 2 is proportional to 0.82. In a fraction 0.2/0.82 of cases, the lineage joins state 1; in a fraction 0.6/0.82, the lineage joins state 3; while in a fraction 0.02/0.82 the lineage joins state 4. This transition matrix has two states, 1 and 6, that a lineage rarely leaves once it has entered it. There is considerable interchange between the other states.

The transition matrix for a character was operated down a phylogeny, obtaining character states at each node from the state of the parent node and the length of the branch segment connecting them. The root was always set to state ($C = 4$, $D = 4$). The branch lengths were determined by establishing a "height" for each

Table 2. *z-values for the six tests scrutinised by null data in three tree shapes*

	Tree shapes		
Test	Tetratomy	Dichotomy	Hennig
ICE	−2.275 ± 1.533	−2.263 ± 1.429	0.02375 ± 1.050
ICDE C1	−0.5399 ± 1.130	−0.2840 ± 0.9693	−0.06588 ± 1.007
ICDE CSP	0.09311 ± 0.8871	−0.04202 ± 0.8709	−0.004584 ± 0.9519
Phylogenetic regression	−0.06541 ± 1.053	0.01201 ± 0.9717	0.06029 ± 0.9370
Burt	−1.153 ± 1.052	−1.162 ± 1.046	−0.001849 ± 1.002
Species	0.2740 ± 2.366	0.01284 ± 1.985	−0.06696 ± 3.144

The reader may like to bear in mind that the 95%, 99% and 99.9% confidence intervals for means are ±0.1033, ±0.1358 and ±0.1734, while for standard deviations they are (0.9269, 1.0731), (0.9045, 1.0967) and (0.8789, 1.1243). These confidence intervals are based on the *z*-scores truly coming from a Normal distribution with a mean of zero and a standard deviation of one.

node. The height of each node depended on the number of species in the clade below the node, say n, and was given by the natural logarithm of n, except for the Hennig phylogeny where it was given by $(n-1)$. The duration of character evolution of a character between two nodes with heights h_m and h_d was

$$\mathrm{RATE}^*\left(\left(\frac{\mathrm{h_m}}{\mathrm{h_R}}\right)^{\mathrm{RHO}}-\left(\frac{\mathrm{h_d}}{\mathrm{h_R}}\right)^{\mathrm{RHO}}\right)$$

where h_R is the height of the root, and RATE and RHO were constants chosen to produce a suitable number of character changes with a suitable distribution of those changes over the different heights of the tree. For the tetratomous and dichotomous cases, we used RATE = 0.4, RHO = 1.0. For the Hennig case we used RATE = 3.6, RHO = 1.0.

Once the underlying characters had been generated for each species in the phylogeny, they were converted to the observable characters. States 1, 2 and 4 of character C determined state 1 of A, otherwise A was in state 2. The same rule converted character D to character B. The calculations were performed using a Mathematica program. We generated 360 random replicates per phylogeny.

The approximate realism of the parameter values is indicated by the number of events in the phylogenies. With the tetratomous, dichotomous, and Hennig phylogenies, the mean number of character changes, reconstructed by ICE, and standard deviation, were 29.64 (SD 4.777), 31.36 (SD 5.011), and 27.67 (SD 5.970) respectively, among the 256 species.

The extent of evolution and effects of the hidden variables can be exhibited. The probability distribution over the six states of the hidden variable C for any one species is {0.0000222, 0.000794, 0.00710, 0.802, 0.182, 0.00797} for the tetratomous and dichotomous trees. (The same probabilities apply to the hidden variable D.) Thus, few species are in the "frozen" states 1 and 6, though those few are likely to be concentrated as large blocks of species in a few datasets. For the Hennigian phylogeny, the probabilities are {0.00645, 0.0195, 0.0346, 0.395, 0.328, 0.216}, so here a sizeable fraction of species are in the frozen states. The probability distributions for the four combinations of states of the two observed variables A and B are {{0.645, 0.158}, {0.158, 0.0387}} for tetratomous and dichotomous cases, and {{0.177, 0.244}, {0.244, 0.335}} for the Hennigian case. These probabilities are calculated analytically using matrix exponentiation and the RATEs in the different phylogenies.

4. Results

Table 1 gives the probability distributions of one-tailed p-values for the four methods (and two versions of the ICDE test), the species counts, and the three phylogenetic shapes. For ICDE and the species counts, the probability distributions are straightforward. For ICE, there is a mean ("mu" in Table 1), a max, and a min. The range of results arises because there can be more than one minimum evolution reconstruction (character change tree) for a dataset. The average number of reconstructions ± standard deviation for the 360 data sets was 101.1 ± 749.9 for the tetratomy, 3.764 ± 7.652 for the dichotomy, and 31.69 ± 176.7 for the Hennig phylogeny. The high standard deviations arise from a few datasets with very many reconstructions, the maximum being 13824, 128, and 3072 for the tetratomous, dichotomous and Hennigian phylogenies, respectively. Each phylogeny had the bulk of its datasets with 50 or fewer reconstructions (297, 359, and 332 out of 360, respectively). For Burt's test the probability distributions have a max and a min. The "min" is the probability calculated by the test that a more extreme result would have been observed by chance. The "max" is the probability that a more extreme or equally extreme result would have been observed by chance. The distributions are summarized as z-values in Table 2.

5. Discussion

5.1. THE SPECIES COUNTS

The distribution of p-values for species is grossly invalid; there are far too many entries in the tails. In the Hennig tree, for example, 220/360 datasets are in the two 5% tails combined, as compared with an expected frequency of 36/360. The simulations reillustrate the point that "significance" tests with species counts in phylogenetically structured data are too likely to find significant results (Grafen, 1989). The comparison of the species results with those for the other tests show that those other tests are all improvements, and have therefore at least some merit. The only apparent exceptions are the ICE and Burt's tests, which with the dichotomous and tetratomous trees have very asymmetric distributions. Here, however, the tests would only sensibly be used if the ancestral state favoured the hypothesis and the relevant p-values are then in the top half of the distribution. There they are very conservative in both cases. The species count probability distributions are unbiased; the average z-value is not significantly different from the correct value of zero (Table 2). It is not expected that z will be biased. [Indeed the proof in Grafen & Ridley (in prep. b) that the ICDE test is unbiased also applies to species counts.]

5.2. THE ICE TEST

The two main features of the results for the ICE test in Tables 1 and 2 are its invalidity with dichotomous and tetratomous phylogenies and its validity with the Hennig phylogeny. Its invalidity is due to a kind of non-independence discussed by Grafen & Ridley (1996b) and Ridley & Grafen (1996). The ICE test reconstructs character states throughout the tree by parsimony. However, with parsimony it is impossible for adjacent nodes to have the same reconstructed character state. Suppose, for example, that a higher node has been reconstructed to have the state *AB*. This node has a number of neighboring nodes in the character change tree. With parsimony, none of those nodes can be *AB*; if one did it would have been merged with the node in question. The neighboring nodes can therefore only have the three different states (*Ab*, *aB*, *ab*). This creates non-independence because the nodes of the character change tree form the entries of the contingency table from which a chi-squared is calculated.

This kind of non-independence is particularly vicious for the dichotomous and tetratomous phylogenies because they produce relatively "star-shaped" character change trees. A star-shaped character change tree has the ancestral state in the centre, and all the other nodes are only one step away from it. This contrasts with a "string" shaped character change tree in which the nodes are arranged in a chain, and there can be many steps from one node to another. In a perfectly star shaped character change tree all the changes in the tree are forced to be away from the ancestral state; the ICE test is then highly biased: it is conservative when the ancestral state favours, but liberal when the ancestral state counts against, the hypothesis (Table 1). In practice, if the ancestral character state permeates the phylogeny, it is only sensible to use the ICE test when the ancestral character state favours the hypothesis under test.

Another way to understand the conservativeness of the ICE test is as follows. Recall that reconstructions in which adjacent nodes have the same state are excluded. If these excluded reconstructions contributed to each possible contingency table in the same ratio as included states, then the p-values would be valid. However, the excluded states are found particularly in the extreme contingency tables, (3,0,0,3) and (0,3,3,0). With the contingency table (3,0,0,3) there are only two types (*AB* and *ab*) to be allocated to all the nodes, and the chance that two states picked at random are equal is two fifths. Whereas, if we have a less extreme contingency table such as (2,1,1,2) then there are four types being allocated, and the corresponding chance is two fifteenths. This argument shows that the lower the p-value in an ICE test, the more conservative the method is. Another consequence of the argument is that the conservativeness is not expected to disappear as a tree increases in size. In general, the probability that a tree with data ($2n$,0,0,$2n$) will have two adjacent nodes the same is always going to be considerably greater than when the data is (n,n,n,n).

The bias disappears, and the probability distribution transforms into validity, when the data are generated through the Hennig phylogeny. The character change tree is less star shaped with the Hennig data, and the ancestral state's sphere of influence is reduced. Now it is more likely that there is a change away from the ancestral state high enough up in the tree for changes back to the ancestral state lower in the phylogeny to be recognized in the parsimonious reconstruction. There are two related effects at work to make the character change tree less stellar. One is the asymmetry of the Hennig tree; the other is that we used a different height rule ($(n-1)$ rather than $\log_e n$) and higher value (3.6 rather than 0.4—a nine-fold increase) for the rate of evolution in the Hennig data generation process (see Section 3: "Null data generation"). The higher rate was needed to produce an approximately similar number of reconstructed evolutionary events as in the dichotomous and tetratomous trees, given that all trees were to have the same number (256) of species. The comparison among tree shapes is therefore partly confounded by the difference in rate: and it may be that with the same rate as the Hennig tree, the ICE test would be valid for the dichotomous and tetratomous trees too. However, the Hennig rate gives an unrealistic number and distribution of changes with the parameter values used for the dichotomous and tetratomous trees. There is a trade-off between the theoretical interest of the comparison between tree shapes at a constant rate and the biological interest of the comparison among trees for a similar, realistic number and distribution of events. We have concentrated on the latter purpose in this paper.

The underlying variable that controls the behaviour of the test is the shape of the character change tree: for a similar number of events, the character change tree is less stellar for the Hennig, than the dichotomous and tetratomous, data. We have not explored the interesting question of how the validity of the ICE test can be tuned by varying the character change tree's shape. With the perfect star it is invalid, and some way toward the full string the test comes to behave itself. In between there will be character change trees with a number of central nodes, each of which acts in the same manner as the single central node in a perfect star. They all force changes to be away from their state. In the case of a double star, with two central nodes that are joined to each other, the two central nodes must have different states. The results will be forced disproportionately into two sets made up of the three states that are complementary to those of the two central nodes; there will then be far too many contingency tables in the tails. There will also be complicated results due to the interaction of triple stars, double stars with a string in the middle, and so on. Our simulations provided several hints of the rich possibilities in varying tree shape. Pursuing these hints may allow us to discover practical rules of the form "The ICE test has approximately correct Type I error rates provided either (i) the character change tree is stellar but the hub state supports the hypothesis or (ii) the character change tree is sufficiently non-stellar". The crucial question is how to measure stellarity.

5.3. THE PHYLOGENETIC REGRESSION

The main features of the results (Tables 1 and 2) are that the frequency distributions of z-values have means and standard deviations that are within the 95% confidence limits around the true values. The results with none of the trees are significantly biased, and the standard deviations of ±1.05, 0.97, and 0.94 are within the range (0.88, 1.12) which are the 95% confidence limits around the true value of one. This result is expected from the second order theory of the phylogenetic regression (Grafen, 1989).

However, the distribution of z can deviate from Normal, as Table 1 reveals for the Hennig phylogeny. The standard deviation for z is correct, but this is attained by having a high number in the middle class (236, cf. the 180 expected) and too many in both tails (at and beyond 1%), compensated by a deficiency in between. The non-Normality arises from the data's discreteness, which gives the error a two-point distribution, and the deviation will decrease for conventional Central Limit Theorem reasons, as sample size increases. The deviation implies that p-values more extreme than a threshold of 5% may not be as significant as they appear with this data generation process. Also, we do not know what the threshold would be with different phylogenies or models of character change. However, the satisfactory means and variances do suggest that the phylogenetic regression is unlikely to be far in error with discrete data.

The results are in one respect not as good as they may appear. We fixed the value of rho, and the branch lengths, at their correct values. Rho = 1 for all tree sizes; but the branch lengths differed because of the $(n - 1)$ height rule in the Hennig tree and $\log_e n$ rule in the dichotomous and tetratomous tree (Section 3). In real applications, branch lengths are unknown and the phylogenetic regression program itself makes a maximum likelihood estimate of, or the biologist specifies, a value of rho, and neither method will normally supply as informed a value as we were able to—because we knew the data generation process. The valid results in Table 1, therefore, do not rule out the possibility (indeed the probability) of difficulties in estimating rho and branch lengths in some real cases. Further work could easily be done on the matter, because any branch lengths can be used in the test, and its validity could therefore be investigated across whatever range of branch lengths the investigator was interested in. Our reason for fixing the correct value of rho was to focus on a direct comparison between continuous and discrete data, to see whether discreteness alone was enough to invalidate substantially the phylogenetic regression; in our simulations it did not.

The influence of branch lengths is an interesting area of difference between discrete and continuous methods. For continuous data, all methods except Burt's rely on branch lengths and the first step is to show that a method works properly when the branch lengths are right. That is the first step we have taken here. A possible second step is then to compare methods that succeed when the branch lengths are right, to investigate how robust they are to error in the phylogeny or branch lengths. For discrete data, the tests do not have to depend on branch lengths. Some tests will not, others will, in ways that may not be immediately obvious. The phylogenetic regression, the ICDE test, and Burt's test (with discrete data) depend on branch length assumptions either in their execution, or justification, or both. The validity of the ICE test, however, may not depend on branch lengths in either its execution or its justification, and in that respect it differs from all the other proposed tests for discrete, and all except Burt's tests for continuous, data. Harvey & Pagel (1991) and Pagel (1994) interpreted the ICE test in terms of branch length assumptions; but we disregard that interpretation.

5.4. BURT'S DISCRETE TEST

5.4.1. *Comparison of results for Hennig phylogeny and those for dichotomous and binary phylogenies*

The results have two main features to discuss, the invalidity of the test in the tetratomous and dichotomous phylogenies, and its validity with the more realistic Hennig phylogeny. The results can be understood in terms of another kind of non-independence discussed by Grafen & Ridley (1996b) and Ridley & Grafen (1996); we call it the family problem. Burt's test finds variable nodes by tracing up from the terminal taxa until it reaches a node below which both characters vary. These nodes can be of three kinds (Fig. 1). Below the node each character will have changed only once. Suppose the ancestral state of the node is *AB*. If the characters change in separate branches that are not below one another, the species below the node will produce a contingency table containing species with the character states *AB*, *Ab*, and *aB*. If the changes happen in successive branches in a lineage, the contingency table will have *ab*, and either *Ab* or *aB*, and maybe some unchanged *AB* descended from the ancestor. Any character state may be found below the node, but Grafen & Ridley (1996b) demonstrate that there is a bias in favour of nodes containing the species with single changes from the ancestor and against species containing the double change. Contingency tables with entries on the "non-ancestral" diagonal are more frequent than those with entries on the "ancestral" diagonal. It will be convenient to discuss the results in terms of the sign of the contingency table. We stay with the convention that the ancestral state is *AB*. Then a contingency table for a radiation that contains some species in the ancestral character state and other species that have a single change from that state will be of form *x*, *x*, *x*,0. Let the values of the four entries in general be n_{AB}, n_{Ab}, n_{aB}, *and* n_{ab}; the sign of ($n_{AB}n_{ab}$-$n_{Ab}n_{aB}$) is then negative. A contingency table containing species in the ancestral state and with a double change from the ancestral state will have a positive sign. The extent of the bias depends on the mix of variable nodes of the three sorts in Fig. 1. The "scattered" sort [Fig. 1(c)] has the worst bias. It always generates a contingency table in which there are no species with a double change; its sign has to be negative.

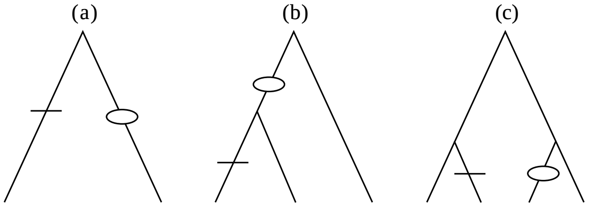

FIG. 1. Three kinds of node below which both characters vary. The line indicates a change in one of the characters (e.g., *A*/*a*) and the squashed circle a change in the other (*B*/*b*). (a) Both characters vary immediately below the same node. The category includes the case in which both characters change in the same branch as well as the illustrated case with changes in sister branches. (b) Staggered variation. The node at which one character varies is above the node at which the other character varies. The category includes the case in which the change below the high node is in the long branch as well as the illustrated case: either way there is variation in the character at the top node. (c) Scattered variation. The two characters vary in different lower branches that do not [as in (a)] connect directly to the same node. Only the elemental branches of the three patterns are shown: any number of uniform species could be added to each.

The invalidity with tetratomous and dichotomous phylogenies exists because, in our datasets, most of the doubly variable nodes were of the Fig. 1(c) type. We inspected a number of datasets to find out which comparisons the test was using and in the few we looked at, nearly all the comparisons were of the scattered Fig. 1(c) type; we found no nodes in which one character changed above another to produce a (*x*, 0, 0, *x*) contingency table (where the ancestral state is either *ab* or *AB* and *x* is any number bigger than zero). The reason is the rate of change in the data generation. A high rate—probably an unrealistically high rate—would be required to obtain the full array of Fig. 1; in our data generation the rate was lower. When the rate is low, nodes like Fig. 1(c) become relatively common compared with the Fig. 1(a) and Fig. 1(b) node types. The rate we used was not unrealistically low; there were an average of about 8–10 doubly variable nodes in the 256 species tree. No method could sensibly excuse itself on the grounds that it was not designed to handle this kind of rate of change.

In the datasets we analysed, some of the positive contingency tables did arise by double changes within the node [like Fig. 1(a) and (b)]. But most of them arose when a character had changed between the ancestor of the tree as a whole and the local node: then a single change in each character produces a contingency table on the ancestral diagonal for the tree as a whole (though locally it is the non-ancestral diagonal). For example, if the ancestral state for the whole tree is *AB* and there has been a change above a doubly variable node to *Ab*, then the two changes will generate *ab* and *AB* species. The contingency table for the node either contains *AB*, *Ab*, and *ab* (if some of the descendants retain the locally ancestral state) or *AB* and *ab* alone (if none do): the sign is positive.

The degree of bias in Burt's test therefore depends on the numbers of doubly variable nodes above which there has been either a change, or no change, from the ancestral state. The former produce positive signs; the latter negative. In the dichotomous and tetratomous phylogenies the latter predominate and invalidate the test. The Hennig phylogeny is more asymmetric, or "ladder-like", and doubly variable nodes are more often beneath a branch containing a change from the ancestral state of the tree (the same process makes the character change tree more string-like, as discussed in relation to the ICE test above). There is now a balance between about half the doubly variable nodes having a local ancestor that is on the non-ancestral and half on the ancestral diagonal. The probability distribution with null data is then more or less symmetric and has valid frequencies (Table 1). The comparison between the two shapes of phylogeny reveals the workings of the family problem. Burt's test treats the states above the doubly variable nodes as if they were randomised. In reality they will not be, but they can be more or less so, and the test is proportionally more or less valid. In the tetratomous and dichotomous phylogenies the family problem is strong, in the Hennig phylogeny it is relaxed.

Burt's test does not owe its validity with the Hennig phylogeny to the reasons that originally inspired the test. With continuous data, the slope of a relation between two variables within a node can be +, 0, or −. The test aimed to test the significance of a relation by seeing whether the slopes were similar in many nodes. If there were no association, there would be a random distribution of slopes. However, with the Hennig phylogeny and discrete data, the distribution of signs do not have this source. Most of the contingency tables are constrained to (tend to) have the sign of the diagonal of the locally non-ancestral character states. The valid behaviour results when half the signs are + and half are − because the locally ancestral states are on each diagonal 50% of the time: but below each node the slope is constrained.

The source of the bias is the use of the signs of the contingency tables. Consider a tetratomous node with one change in each character. With null data the chance that both changes are in the same lineage is 25% (in which case the contingency table is positive) and there is a 75% chance that the two changes will be in different lineages and produce a negative contingency table. The contingency tables (2,1,1,0) and (3,0,0,1) have frequencies 3:1, and the signs are biased 3:1 in favour of the negative. Burt's main exposition of his method was for continuous characters and he wrote of phylogenetically independent contrasts that the "variances and covariances of characters within contrasts, which are assumed to depend only on events occurring since the last common ancestor, will thereby be independent of variances and covariances in other contrasts." For continuous characters this is true. But the relation between phylogenetic and statistical independence is more difficult with discrete characters. The evolutionary events indeed happened independently below each node, but that does not guarantee statistical independence: the signs suffer from the family problem. However, the underlying numbers are unbiased. A rare extreme positive contingency table (3,0,0,1) balances a common less extreme negative one (2,1,1,0). If we calculate the covariance estimates of the two they are $-1/16$ and $+3/16$ and balance the frequencies. The value of the true covariance is therefore zero, and unbiased. It may be that the way forward for Burt's test with discrete data is not to use the signs, but the covariances, of the contrasts. With null data the covariance is zero and the test could look for deviations from zero. The covariances are not normally distributed, however, and a *t*-test is inapplicable.

5.4.2. *Møller and Birkhead's method of pairwise comparisons*

The bias is removed in the method of Møller & Birkhead (1992). They suggested picking pairs of populations within a species, or pairs of species within a genus, such that the pair differed for one of the characters. The relation with some other character could then be examined in each pair. In their example, they picked pairs that differed in whether the population (or species) was solitary or colonial, and compared copulatory frequencies between the two. The character used to define the pairs was practically discrete; the other character, copulatory frequency, was continuous, but the same method could be used if it were discrete. The main advantage of the method, as Møller & Birkhead argued, is that the influence of other, unobserved variables on the relation between the characters under study should be minimised. Other things are more likely to be equal for two species in a genus than for two species from two classes, or two phyla.

However, the method has the additional advantage that, by picking *two* species from a genus, it avoids the bias that is otherwise introduced by focusing on low level nodes below which both characters vary. The lineages leading to a pair of species will have a shared region between the ancestor of the whole tree and the common ancestor of the species pair, and then separate regions after their common ancestor. Provided there is the same chance of change from the common ancestor of the pair to each of the species, there must be a 50% chance of a + relation between the characters and a 50% chance of a − relation with null data. Both characters change in the pair of lineages. If character A/a has changed in one of the separate lineages, the chance that the character B/b also changed in that lineage equals the chance B/b

changed in the other lineage. The chance of + and − relations are therefore equal. The only important assumption is that there is an equal chance of change in the lineages leading to the two species. It does not have to be true, but for much of evolution it is a reasonable assumption in a null hypothesis. It is interesting to note that to remove the bias, it is essential to pick exactly two species from the variable node: if three were picked, for example, the bias would creep back. Using all the species, as Burt's test does, gives the full bias.

We did not scrutinise Møller & Birkhead's method in our simulations. In our 256 species tree there was no "intraspecific" variation and our values of rho and rate gave very few "genera" with a pair of species varying in both characters: there might be two or so per tree. The method was practical in their study, which concerned relatively well studied characters in a relatively well studied group containing thousands of species (birds): even then they found only 13 pairs for comparison. The main drawback of the method is that it throws away data when it focuses on intrageneric and intraspecific pairs. This introduces an element of arbitrariness into the test. Why stop at genera? Why not go up to families? Suppose a study including families gave one relation and a study excluding them gave another: what should be concluded? The throwing away of data would then matter. It also matters for the more obvious reason of statistical power: its power may be reasonable with continuous data and a broad taxonomic sweep: but there is more to comparative biology than that. We are not arguing the method is without merit; indeed the way it avoids the family problem is clearly meritorious. In summary, the method is valid, and in particular avoids the family problem, but will be practicable only in restricted circumstances.

5.5. THE ICDE TEST

The main features of the results (Table 1) are that (i) the CSP version, in which numbers of species are used within each randomisation node, is reasonably well behaved with all shapes of phylogeny. It appears to be a little conservative. (ii) The C1 version, in which all non-zero numbers of species are reduced to unity, gives invalid significances, and is biased against the ancestral diagonal, except with the Hennig phylogeny. Thus CSP is generally better behaved than C1. Grafen & Ridley (in prep. b) demonstrate that CSP will be unbiased on two assumptions, which were met in our simulations; but they did not formally analyse its validity. Here we have shown with simulated null data that it is indeed unbiased and has phenomenologically valid Type I error rates. The bias in the C1 version of the test is due to the family problem discussed in Section 5.4.1 above. With null data, the test finds too many contingency tables with entries on the non-ancestral diagonal. The bias is strongest in the tetratomous and dichotomous trees and disappears in the Hennig phylogeny. The reason is that the Hennig tree is asymmetrical and, as we noticed above when discussing the ICE test (Section 5.2), it is commoner than in the dichotomous or tetratomous trees for one change to be below another change in the same character. Now the local ancestral state above a change may be in any position in the contingency table, and the entries contributed by the randomisations may also be in any position, rather than being constrained into the neighbourhood of the ancestral state. Thus in the Hennig phylogeny the family problem is relaxed; the assumption that changes in different parts of the tree are independent applies; and the ICDE C1 test is approximately valid.

CSP works with the same randomisation nodes as C1: why does it not suffer from the same bias due to the family problem as C1 does? Here is at least part of the answer. In CSP the numbers in the contingency table are again constrained in the triangular pattern of C1 (with entries for the ancestral state and the two single change non-ancestral states but none for the double change); but the effect is not so extreme. Focus (for concreteness) on the corner opposite the ancestral state; it is *ab* when the ancestral state is *AB*. It will be observed to be low; and the expected value is calculated from the row and column totals. The highest number of species will be *AB*; but in C1 each chunk of ancestral-state species is reduced to one and the *AB* total will be lower than in CSP. When the expected number for *ab* is calculated in CSP, the expectation is low because it is predicted from two small fractions (the non-ancestral row and column totals/*N*), whereas in C1, *N* is lower and the predictive fractions higher, and the expected value for *ab* higher too. Thus, C1 spotlights the non-independence in the data due to the family problem. In CSP the lower expected number of species with *ab* appears to compensate for the bias in the way the test looks at the data, with the result that the family problem type non-independence does not bias the test statistic. The power of ICDE remains to be evaluated.

6. Conclusion

Grafen & Ridley (1996a) proposed understanding the problems of discrete phylogenetic data in terms of non-independence between path segments, and not just between species tips, in the process that gives rise to the data. Above we encountered the two further types of non-independence described by Grafen & Ridley (in prep. b): non-independence that arises in our reconstructions (between adjacent nodes in the character state tree in ICE) and analyses (the "family problem"). These various kinds of independence are a formalisation of the statistical problems of discrete data, and it is through identifying and characterising their nature and effects that progress will be made.

We now summarise the main results; we also offer some interim recommendations concerning which tests to use, though the analysis is too incomplete to allow a final judgement. The ICE test has reasonably valid Type I error rates with a realistically shaped phylogeny, though the results for the tetratomous phylogeny show how it can become severely biased under some conditions. The things to watch out for in real cases are the extent to which the ancestral state at the root is retained throughout the tree and whether that ancestral state counts for or against the hypothesis. If the ancestral state is retained through much of the tree, such that most changes are only one step away from it, it is only safe to use the test if the ancestral state counts for the hypothesis; the ICE test is then conservative (and can become unusably conservative). The phylogenetic regression, at least when branch lengths are known, has approximately valid Type I error rates and therefore can be used with discrete as well as continuous data. Burt's test shows a similar pattern to the ICE test: it has reasonable validity for a realistically shaped tree but is vulnerable to bias in the dichotomous and tetratomous trees. The results, however, are not due to the reason that inspired the test, and this brings into question the whole rationale of the test with discrete data. The new ICDE test supplied valid results in its CSP version, but the analysis here is preliminary and would not justify a recommendation for general use.

REFERENCES

Burt, A. (1989). Comparative methods using phylogenetically independent contrasts. *Oxford Surv. Evol. Biol.* **6,** 33–53.

Grafen, A. (1989). The phylogenetic regression. *Phil. Trans. R. Soc. Lond.* **B 326,** 119–157.

Grafen, A. & Ridley, M. (1996a). A new model of discrete character evolution. *J. theor. Biol.* (in press).

Grafen, A. & Ridley, M. (1996b). Non-independence in statistical tests for discrete cross-species data. *J. theor. Biol.* (submitted).

Grafen, A. & Ridley, M. (in prep. a). A formalisation of the comparative method of Ridley (1983).

Grafen, A. & Ridley, M. (in prep. b). Independent character difference evolution: a new statistical test for discrete cross-species data.

Harvey, P. H. & Pagel, M. D. (1991). *The Comparative Method in Evolutionary Biology*. Oxford: Oxford University Press.

Hennig, W. (1981). *Insect Phylogeny*. Chichester, U.K.: John Wiley.

Maddison, W. P. (1990). A method for testing the correlated evolution of two binary characters: are gains or losses concentrated on certain branches of a phylogenetic tree? *Evolution* **44,** 539–557.

Møller, A. P. & Birkhead, T. R. (1992). A pairwise comparative method as illustrated by copulation frequency in birds. *Amer. natur.* **139,** 644–656.

Numerical Algorithms Group [NAg] (1987). *The Generalised Linear Interactive Modelling System*. Oxford, UK and Downers Grove, Illinois: Numerical Algorithms Group.

Pagel, M. (1994). Detecting correlated evolution on phylogenies: a general method for the comparative analysis of discrete characters. *Proc. R. Soc. London* **B 255,** 37–45.

Read, A. F. & Nee, S. (1995). Inference from binary comparative data. *J. theor. Biol.* **173,** 99–108.

Ridley, M. (1983). *The Explanation of Organic Diversity*. Oxford: Clarendon Press.

Ridley, M. & Grafen, A. (1996). How to study discrete comparative methods. In: *Phylogenies and the Comparative Method in Animal Behavior*, Martins, E. P. (ed.), p. 76–103. New York: Oxford University Press.

Wolfram, S. (1988). *Mathematica*: *A System for Doing Mathematics by Computer*. New York: Addison-Wesley.

APPENDIX

Dichotomous: {{{{{{{{1, {{2, 3}, 4}}, {{{5, {6, 7}}, 8}, {{9, {10, 11}}, 12}}}, {{13, {14, 15}}, 16}}, {{{{{17, 18}, 19}, 20}, {21, {22, {23, 24}}}}, {{25, {26, {27, 28}}}, {29, {30, {31, 32}}}}}}}, {{{{33, {{34, 35}, 36}}, {37, {{38, 39}, 40}}}, {{{{41, 42}, 43}, 44}, {{{45, 46}, 47}, 48}}}, {{{49, {50, {51, 52}}}, {53, {{54, 55}, 56}}}, {{{57, {58, 59}}, 60}, {{61, {62, 63}}, 64}}}}}, {{{{{{{{65, 66}, 67}, 68}, {69, {70, {71, 72}}}}, {{73, 74}, {75, 76}}}, {{77, {78, 79}}, 80}}, {{{{81, 82}, {83, 84}}, {{{{85, 86}, {87, 88}}, {{89, 90}, {91, 92}}}, {{93, 94}, {95, 96}}}}, {{{97, {98, {99, 100}}}, {{{101, 102}, {103, 104}}, {{105, 106}, {107, 108}}}}, {{109, 110}, {111, 112}}}}}, {{{113, 114}, {115, 116}}, {{{{117, 118}, {119, 120}}, {121, {{122, 123}, 124}}}, {{125, 126}, {127, 128}}}}}}, {{{{{{{129, 130}, {131, 132}}, {{133, 134}, {135, 136}}}, {{137, 138}, {139, 140}}}, {{141, 142}, {143, 144}}}, {{{{{{145, 146}, {147, 148}}, {{149, {150, 151}}, {153, {{154, 155}, 156}}}, {{157, 158}, {159, 160}}}, {{{{165, 166}, {167, 168}}, {{169, 170}, {171, 172}}}, {{161, {{162, 163}, 164}}, {{173, 174}, {175, 176}}}}}}, {{{{177, 178}, {179, 180}}, {{181, 182}, {183, 184}}}, {{185, {186, {187, 188}}}, {189, {190, {191, 192}}}}}}}, {{{{193, {194, 195}}, 196}, {{{197, {{198, 199}, 200}}, {{201, 202}, {203, 204}}}, {{205, {206, 207}}, 208}}}, {{{{209, {210, {211, 212}}}, {{{{213, 214}, 215}, 216}, {{{217, {218, 219}}, 220}, {{221, {222, 223}}, 224}}}}, {{{{225, 226}, 227}, 228}, {{{229, {230, 231}}, 232}, {{{{233, 234}, 235}, 236}, {{237, {238, 239}}, 240}}}}}, {{{241, {242, {243, 244}}}, {{245, 246}, {247, 248}}}, {{249, {{250, 251}, 252}}, {253, {{254, 255}, 256}}}}}}}}

Hennig: {{1, 2}, {3, {{4, 5, {6, 7}}, {8, 9, 10}}, {11,{{{{{{{{12, 13}, {14, 15}}, {{{{16, 17}, 18}, 19}, 20}}, {{{{21, 22}, 23}, {{24, 25}, {26, {27, 28}}}}, 29}}, 30}, 31}, 32}, {{33, 34, 35}, {{{{36, 37, {38, {39, 40}}, {41, {42, {{{{43, 44}, 45}, 46}, {{{47, 48, 49}, 50}, {{{{51, 52}, 53}, 54}, {{{55, 56, 57}, {{{58, 59}, 60}, {{{{{{{{61, 62}, 63}, 64}, 65}, 66}, {{67, 68}, {69, 70, 71}}}, {{72, 73}, 74}}, {75, 76}}, 77}}, {{78, 79}, 80}}}}}}}}, 81}, {82, {{83, {84, 85}}, {{{86, 87}, {{88, 89}, {90, {91, 92}}}}, {{93, {94, 95}}, {{{96, {97, 98}}, {99, 100}}, {{{101, 102}, {103, 104}}, {105, {{106, 107}, {{108, 109}, {110, 111}}}}}}}}}}}, 112, {{{113, {114, 115}}, {116, 117}}, {{118, {{119, {120, 121}}, {{{{122, 123}, 124}, {125, 126}}, {{127, 128}, {{{129, {130, 131}}, {132, 133}, {134, 135, 136, 137}}, {138, 139}}}}}}, {{{140, {141, {142, 143}}}}, {144, 145}, {146, 147, 148}, {149, 150}}, {151, 152}}}}, {153, {154, {{155, {{156, 157}, {158, {{159, 160}, {{161, 162}, {163, 164}, 165}}}}}, {166, {{167, {168, {169, {{170, 171}, {{172, {173, 174}}, {175, {176, {177, 178}}}}}}}}, {{{{179, 180}, 181}, 182}, {{183, 184}, {185, 186}}}, {{{187, 188}, {{189, 190, {191, 192, 193}}, {194, {195, {{{196, 197}, {198, 199}}, {200, 201}}}}}}, {{202, {203, 204}, {205, 206}}, {{{207, 208}, 209}, {210, {211, 212}, {{{213, 214}, 215}, {{{216, 217}, {218, 219}}, 220}}, {{221, {{{222, 223}, 224}, 225}, {{{226, {227, {228, 229}, {230, 231}}}, 232}, 233}}, {{{{234, 235}, {236, 237}}, 238}, {239, {{{{240, 241}, 242}, {{{{243, 244}, 245}, {246, 247, 248}}, 249}}, {{250, 251}, {{{{252, 253}, 254}, 255}, 256}}}}}}}}}}}}}}}}}}}}}}