ՀԱՅԿԱԿԱՆ ՍՍՌ ԳԻՏՈՒԹՅՈՒՆՆԵՐԻ ԱԿԱԴԵՄԻԱՅԻ ՏԵՂԵԿԱԳԻՐ
ИЗВЕСТИЯ АКАДЕМИИ НАУК АРМЯНСКОЙ ССР
Բիոլոգիական գիտ. XVII, № 12, 1964 Биологические науки

ПАВЕЛ ГАМБАРЯН

МАТЕМАТИЧЕСКИЙ МЕТОД ТАКСОНОМИИ

Существующие методы систематизации любых объектов используют их признаки. Биологическая систематика для классификации таксонов также пользуется их признаками. Но от любой другой классификации она отличается тем, что должна отражать эволюцию, происхождение и родственные связи таксонов, т. е. быть филогенетической.

Чем большее количество различных признаков будет использовано в филогенетических построениях, тем больше мы приблизимся от ряда форм к филогенетическому ряду; чем большее число общих признаков мы найдем у двух организмов, тем более вероятно их родство [8].

Использование большого числа признаков в таксономических исследованиях делает применение математических методов в систематике все более необходимым. Не случайно существует довольно много методов математической таксономии [3—7 и др.].

Методы таксономии сводятся к следующему: чем больше общей информации в признаках сравниваемых таксонов, тем эти таксоны ближе. Информация, общая для изучаемых таксонов, имеет 2 источника: количество общих признаков и количество информации в каждом признаке. В части работ [4, 6 и др.] используется лишь информация числа общих признаков, так как считается, что нет объективных критериев оценки признака, или количества информации в признаке. Но в систематике существует весьма объективный критерий ценности признака, связанный с его частотой: например, признак, общий всем видам рода—родовой, а свойственный лишь одному виду--видовой. Интерполируя, можно сказать, что признак, свойственный части видов рода, рангом ниже родового, но выше видового.

Е. С. Смирнов [3] предложил метод таксономического анализа с разной оценкой признака. В. М. Шмидт [10] популяризировал этот метод и предложил использовать его в ботанике. Суть метода такова: все признаки разбиваются на тезу и антитезу. Значимость признака оценивается делением числа таксонов без признака на число таксонов с признаком. Для расчета близости двух таксонов суммируют значимость всех признаков, по которым таксоны совпадают, вычитают число несовпадающих признаков и результат делят на число признаков. По методу Смирнова самую высокую оценку получает самый редкий признак, а теза и антитеза оцениваются различно. К несомненным достоинствам метода Смирнова надо отнести его относительную простоту и то, что признак рассматривается не сам по себе, а как полная система событий наличие его у части таксона и отсутствие у всей остальной части.

Если рассматривать все многообразие организмов, то наибольшую информацию о их близости даст действительно самый редкий признак, так как отсутствие признака в бесконечной системе не дополняет событие наличия признака у двух изучаемых таксонов. Но мы изучаем конечный таксон с ограниченным объемом, в котором признак рассматриваем как систему. Самый редкий признак в такой системе дает вовсе не максимальную информацию. Например: в роде 10 видов, из них 4 с желтыми цветками. Частота признака «желтые цветки» $P=4:10=0{,}4$ и частота антитезы, или «не желтые цветки» $q=1-P=6:10=0{,}6$. Рассчитаем информацию признака «желтые цветки».

Количество информации, приобретаемое при полном выяснении состояния некоторой физической системы, равно энтропии этой системы [2]. Энтропией системы называется сумма произведений вероятностей различных состояний системы на логарифмы этих вероятностей, взятая с обратным знаком [2].

Обозначим энтропию $H(X)$. Так как в нашей системе 2 возможных состояния—теза или антитеза, то $H(X)=-(P\log_2 P+q\log_2 q)$. В табл. 1 приведены значения $\S(P)=-(P\log_2 P)$ (табл. 1 взята из книги Е. С Вентцеля [2] с изменением). Находим значения $\S(P)$ для $P=0{,}4$ и $P=0{,}6$

Таблица 1

Значения $\S P=-P\log_2 P$. (0 целых всюду опущено)

P	0	1	2	3	4	5	6	7	8	9	q
0,0	0	0644	1128	1518	1858	2161	2435	2686	2915	3126	0,0
0,1	3322	3503	3671	3826	3971	4105	4230	4346	4453	4552	0,1
0,2	4644	4728	4806	4877	4941	5000	5053	5100	5142	5179	0,2
0,3	5211	5238	5260	5278	5292	5301	5306	5307	5305	5298	0,3
0,4	5288	5274	5256	5236	5210	5184	5153	5120	5083	5043	0,4
0,5	5000	4954	4906	4854	4800	4744	4685	4623	4558	4491	0,5
0,6	44 2	4350	4276	4199	4121	4040	3957	3871	3784	3694	0·6
0,7	3602	3508	3412	3314	3215	3113	3009	2903	2796	2687	0,7
0,8	2575	2462	2348	2231	2112	1992	1871	1748	1623	1496	0,8
0,9	1368	1238	1107	0974	0839	0703	0565	0426	0284	0144	0,9
1,0	0	1	2	3	4	5	6	7	8	9	q

и суммируем. $0{,}5288+0{,}4422=0{,}9710$. Если $P=0{,}2$ и $q=0{,}8$, то $H(X)=0{,}7219$. По Смирнову эти признаки получили бы оценку $6:4=1{,}5$ и $8:2=4$, т. е. редкий признак получает большую оценку, хоть энтропия его не велика. Нетрудно убедиться, что энтропия будет равной, возьмем ли мы два таксона с тезой или антитезой. Количество информации признака с его антитезой возьмем как оценку значимости признака. Сумму энтропий всех признаков, общих для двух сравниваемых таксонов, или общую информацию будем считать за показатель их близости.

Так как для законности сложения энтропий признаки должны быть независимыми, а степень зависимости признаков рассчитать пожалуй невозможно, то условимся объединять в I все признаки, встречающиеся только вместе. Например, если все виды в анализируемом таксоне с

Таблица 2

Распределение признаков, частоты и энтропия для групп признаков с одинаковой частотой

Признаки	1	2	3	4	5	6	7	8	9	10	P, q, 4Pq H(X)
Листья простые	+	+	+	+	+	—	—	—	—	—	P=0,5
Хромосом 6	+	+	+	—	—	+	+	—	—	—	q=0,5
Плод ягода	+	+	+	+	—	+	—	—	—	—	4Pq=1
											H(X)=1
Кора гладкая	+	—	—	+	—	+	—	+	—	—	P=0,4
Древесина с ядром	+	+	—	—	—	+	—	—	+	—	q=0,6
Почки черные	+	+	—	—	+	—	—	+	+	+	4Pq=0,96
Поры окаймленные	—	—	+	+	—	+	+	—	—	—	H(X)=0,971
Пыльца больше 60 μк	—	—	—	—	—	+	+	+	—	—	P=0,3
Растения болот	—	—	+	—	+	—	+	—	—	—	q=0,7
Листья опушены	+	+	—	—	—	+	—	—	—	—	4Pq=0,84
Кустарники	+	—	—	—	—	+	—	+	—	—	H(X)=0,881
Пыльники острые	+	+	+	—	—	—	—	—	—	—	
Растения Америки	—	—	—	—	—	—	—	+	+	+	
Цветки красные	—	—	+	—	+	—	—	—	—	—	P=0,2
Плоды висячие	—	—	—	+	—	—	—	+	—	—	q=0,8
Лепестки голые	—	—	—	+	+	+	—	—	—	—	4Pq=0,84
Плоды жгучие	—	—	—	—	—	+	—	—	—	+	H(X)=0,722
Вечнозеленое	—	—	—	—	—	+	+	—	—	—	

красными цветками имеют цельные листья, а все виды с цельными листьями красные цветки, то признаки «листья цельные» и «цветки красные» будем считать за один. В других случаях, где зависимость между признаками учесть не легко, можно рекомендовать просто подбирать признаки, предполагаемые независимыми. Для этого надо использовать признаки не только морфологические, но и биохимические, цитологические, эмбриологические, биоценотические и пр.

Для удобства вычислений строим таблицу распределения признаков и антитез (для построения таблицы полезно использовать метод Балковского для политомического ключа. Бот. журнал, т. XLV, 1, 1960). Признаки с одинаковой частотой располагаем группами, так как они обладают одинаковой энтропией. Начертив на полоске бумаги для каждого таксона наличие всех признаков, проводим полоску по таблице и для каждого таксона подсчитываем, по скольким признакам и антитезам в каждой группе наши таксоны совпадают. Перемножив число совпадений в группе на энтропию, складываем и получаем количество информации, общей у сравниваемых таксонов (табл. 3).

По данным табл. 3 строим кольца связи, аналогичные сечениям корреляционных цилиндров [9, 11]. На кольце с номерами таксонов соединяем между собой №№ таксонов, информация связи которых равна или больше выбранного уровня кольца (рис. 1)

По данным колец строим дендрограмму, которая наглядно показывает всю информацию о связи таксонов. Дендрограмма строится так:

Таблица 3

Информация связи 10 таксонов и ошибки информации

	1	2	3	4	5	6	7	8	9	10
1	15,8 1,64	13,9 1,56	8,5 1,20	7,8 1,16	6,9 1,08	10,3 1,32	4,8 0,84	7,5 1,12	7,5 1,12	6,6 1,04
2			10,4 1,32	7,7 1,12	8,5 1,24	8,5 1,20	6,7 1,04	5,7 0,96	8,4 1,16	8,4 1,16
3				9,9 1,28	10,2 1,28	7,0 1,12	8,6 1,20	4,0 0,76	5,8 0,92	6,7 1,04
4					9,5 1,24	8,0 1,20	7,9 1,16	8,4 1,16	6,9 1,04	7,9 1,12
5						3,2 0,08	8,8 1,24	8,1 1,24	9,9 1,32	10,8 1,40
6							10,2 1,36	9,6 1,28	5,9 0,92	4,9 0,88
7								7,8 1,16	7,8 1,16	8,7 1,24
8									10,6 1,36	11,6 1,44
9									15,8 1,64	14,8 1,60
10										15,8 1,64

до того уровня, пока все таксоны на кольцах еще связаны между собой, идет общий ствол. Он ветвится на столько ветвей, сколько несвязанных групп появится, и ветви кончаются на том уровне, когда у таксона не остается связей (рис. 2).

Исследуя любой таксон, мы по существу исследуем лишь выборку из него, так как часть таксона вымерла, возможно не описана или просто нам недоступна. Из всего многообразия признаков, мы тоже исследуем выборку. Поэтому общими статистическими приемами определим стандартную ошибку наших выборочных показателей. Если ошибку энтропии H(X) обозначим S_x, а энтропию P - §P и q—§q, то ошибка H(X) будет

$$S_x = 4\sqrt{\frac{(\S P)^2\, nP - \frac{1}{n}(\S P\, nP)^2 + (\S q)\, nq - \frac{1}{n}(\S q\, nq)^2}{n(n-1)}},$$

где n объем таксона. Если информацию связи двух таксонов обозначим T, то $S_T = \sqrt{\sum_{i=1}^{m}(S_x)_i^2}$, где m число совпадающих признаков.

Расчет ошибки энтропии и самой энтропии можно упростить, за-

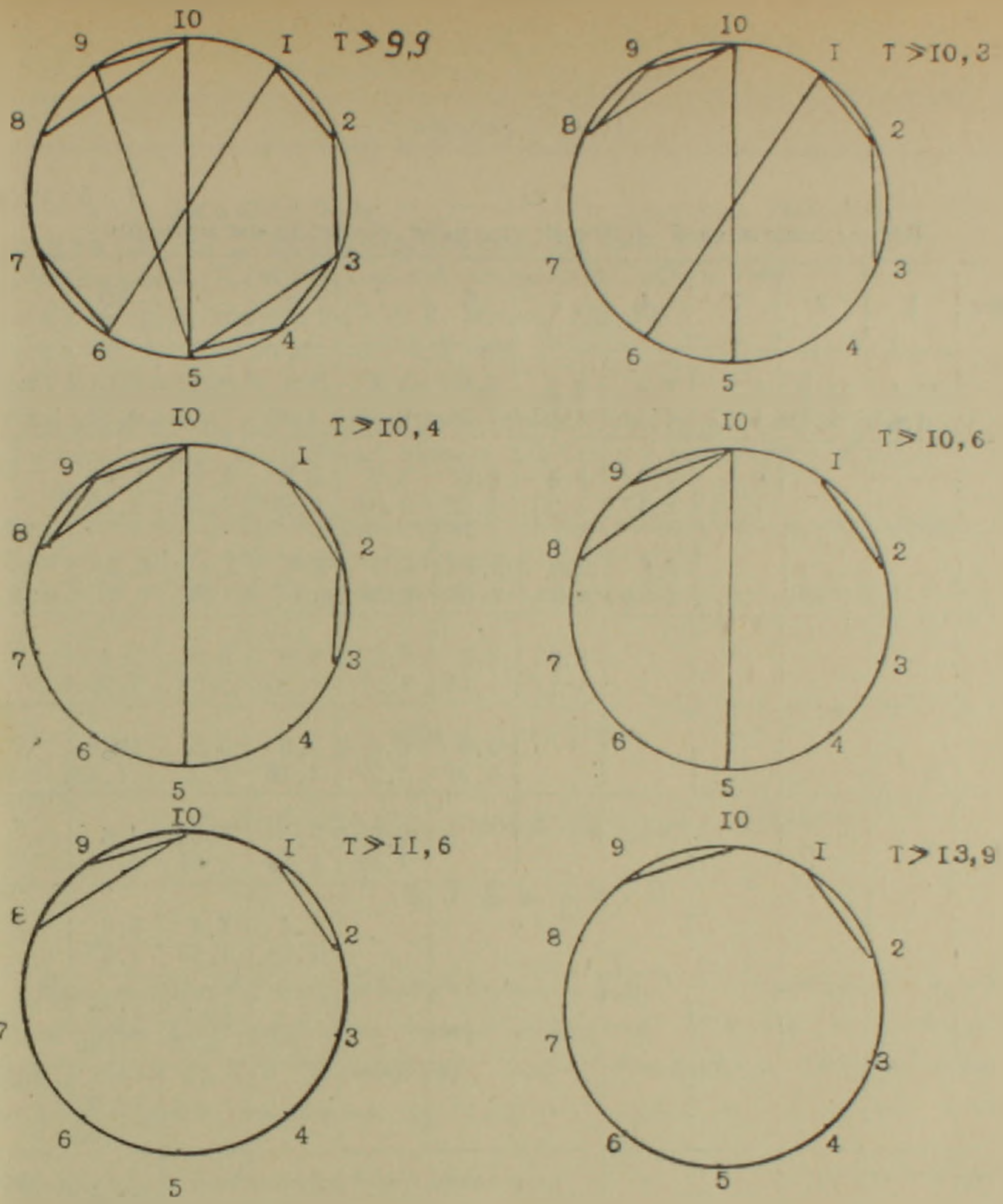
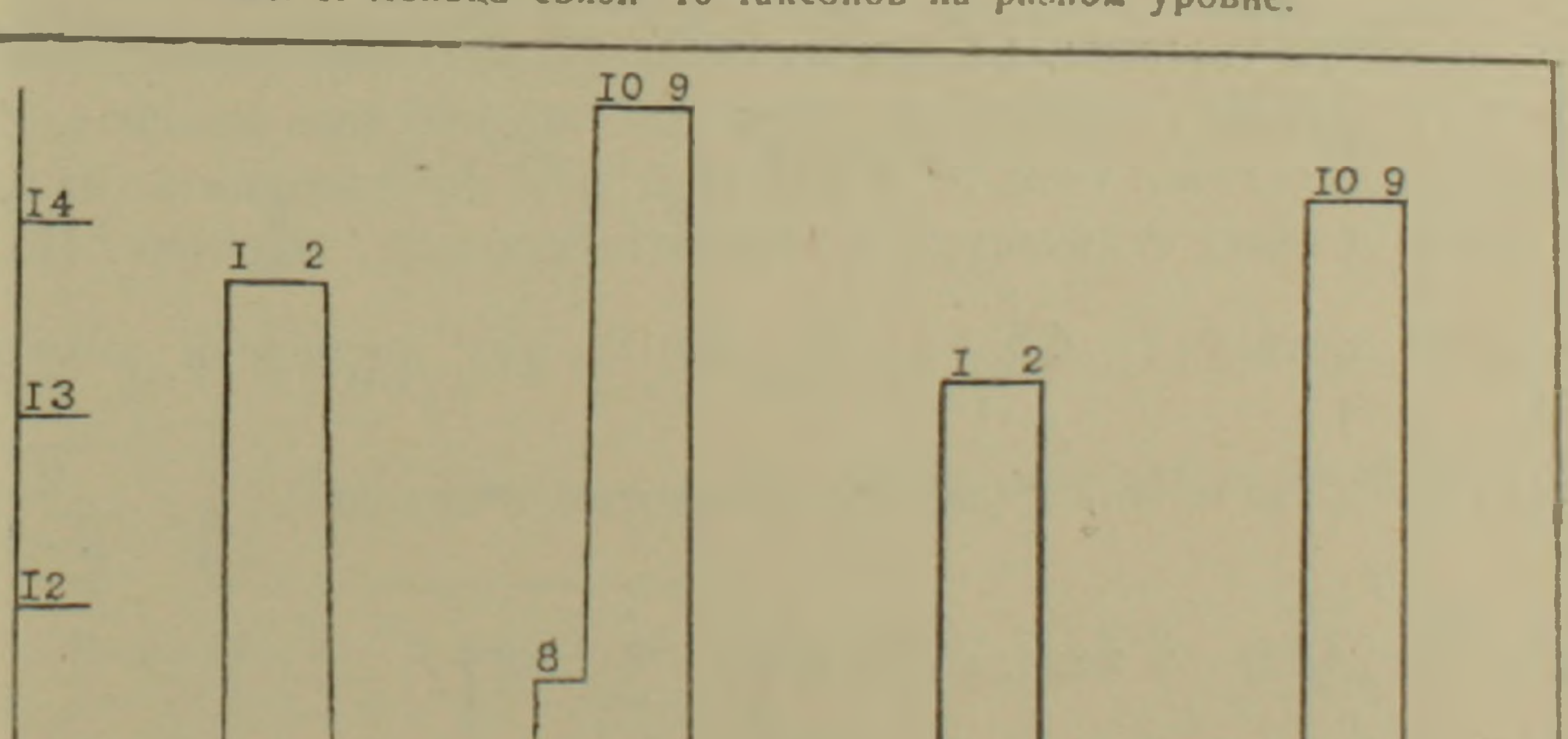


Рис. 1. Кольца связи 10 таксонов на разном уровне.



ис. 2. Дендрограмма связи таксонов по уровню Т полным (а) и сокращенным (б) методам расчета.

Таблица 1

Расчет показателей связи 10 таксонов упрощенным методом.

Таксоны	1	2	3	4	5	6	7	8	9	10	
1	15,1 1,61	13,2 1,58	8,1 1,22	7,4 1,17	6,5 1,10	9,8 1,34	4,4 0,86	7,1 1,13	7,1 1,13	6,2 0,93	T S_T
2			9,8 1,35	7,3 1,17	8,3 1,25	8,0 1,22	6,2 1,04	5,3 0,95	8,0 1,20	8,0 1,20	
3				9,6 1,34	9,6 1,34	6,7 1,13	8,2 1,28	3,7 0,67	5,4 0,95	6,3 1,07	
4					9,1 1,29	7,7 1,13	7,5 1,07	8,0 1,20	6,5 1,09	7,4 1,06	
5						3,0 0,71	8,4 1,27	7,8 1,16	9,5 1,37	10,4 1,49	
6							9,6 1,33	9,1 1,29	5,6 1,01	4,6 0,90	
7							7,4 1,18	7,4 1,18	8,4 1,27		
8								10,2 1,31	11,2 1,40		
9									14,2 1,63		
10									15,1 1,61		

ранее составив таблицы, но эти таблицы будут громоздкими. Вместо этого я предлагаю упрощенный метод таксономического анализа. Если перемножить частоты тезы P и антитезы q и Pq умножить на 4, то 4Pq незначительно отличается от соответствующей энтропии. По упрощенному методу $T = 4\sum_{i=1}^{m} Pq_i$. Так как Pq это дисперсия альтернативного распределения, а ошибка дисперсии S^2 равна $S^2\sqrt{\frac{2}{n-1}}$, то для упрощенного метода $S_T = 4\sqrt{\sum_{i=1}^{m}\left(Pq\sqrt{\frac{2}{n-1}}\right)^2}$.

Рассчитанные значения для того же таксона по упрощенному методу даны в табл. 4. Как нетрудно убедиться, результаты упрощенного метода незначительно отличаются, только количество информации несколько меньше. При не очень ответственных исследованиях или в качестве прикидки конечно, лучше пользоваться упрощенным методом.

Биологический факультет
Ереванского государственного университета

Поступило 29.II 1964 г.

ЛИТЕРАТУРА

1. Ван дер Варден Б. Л. Математическая статистика. ИЛ., 1961.
2. Вентцель Е. С. Теория вероятностей. М., 1962.
3. Смирнов Е. С. Журнал общей биологии, т. XXI, 2, 1960.
4. Sneath P. H. and Sokal R. R. Nature, 193, 1962.
5. Sokal R. R. Evolution, XIII; 420—423, 1959.
6. Sokal R. R. Taxon, vol. 12, 5, 1963.
7. Тамамшян С. Г. 2-ое Московское совещание по филогении (тезисы докл.), 1964
8. Тахтаджян А. Л. Бюлл. Москов. общества испытателей природы, т. L., 11(5), 1947.
9. Терентьев П. В. Сб. Применение математических методов в биологии, I. Л., 1960
10. Шмидт В. М. Ботанический журнал, 11, т. 47, 1962
11. Шмидт В. М. Сб. Применение математических методов в биологии. II. Л., 1963

ՊԱՀԵԼ. ՂԱՄԲԱՐՅԱՆ

ՏԱՔՍՈՆՈՄԻԱՅԻ ՄԱԹԵՄԱՏԻԿԱԿԱՆ ՄԵԹՈԴԸ

Ա մ փ ո փ ո ւ մ

Առաջարկվում է տաքսոններն իրար այնքան մոտ համարել, որքան ավելի ինֆորմացիա կա համընկնող հատկանիշներում։ Եթե p-ն հատկանիշի հաճախակիությունն է, q-ն անտիթեզի, ապա հատկանիշի ինֆորմացիան կամ նրա էնտրոպիան հավասար է $H(x) = -(p\log_2 p + q\log_2 q)$։ Համընկնող հատկանիշների էնտրոպիանների գումարը՝ $T = \sum_{i=1}^{m} H(x)_i$-երկու տաքսոնների մոտության ցուցանիշն է։ Ամեն մի հատկանիշի էնտրոպիայի սխալը՝

$$Sx = 4\sqrt{\frac{(\S p)^2 np - \frac{1}{n}(\S p \cdot np)^2 + (\S q)^2 nq - \frac{1}{n}(\S q nq)^2}{n(n-1)}},$$

որտեղ $\S p = -p\log_2 p$ և n-ը տաքսոնի ծավալն է։ Սխալը T-ին. կամ $S_T = \sqrt{\sum_{i=1}^{m}(Sx)_i^2}$, որտեղ m-ը համընկնող հատկանիշների թիվն է։

Առաջարկվում է հաշվարկի պարզեցված մեթոդ։ Հատկանիշի էնտրոպիան $\frac{m}{n}$ պետք փոխարինվում է $4pq$. $T = 4\sum_{i=1} pq_i$ և $S_T = 4\sqrt{\sum_{i=1}\left(pq\sqrt{\frac{2}{n-1}}\right)^2}$ արտահայտությամբ։

Հաշվարկն երկու մեթոդներն էլ պատկերացված են 10 միավորներից և 18 հատկանիշներից բաղկացած տաքսոնի օրինակի վրա։