

床板珊瑚形珊瑚的数值分类研究*

张松林

(南京地质矿产研究所)

内容提要 本文系统地论述了床板珊瑚形珊瑚的数值分类研究。实际工作是在传统“属”一级分类的基础上,对种一级的代表性化石标本进行的。作者根据数值分类的要求,在深入研究前人对各种的化石标本所作大量文字描述的基础上,提取有效的鉴定特征作为分类性状,然后对这些性状进行编码和标度,采用加权明氏距离估计任意两化石标本之间的相似性,最后根据每个属各种的化石标本间的相似性,利用系统聚类法实现对种的重新厘定。

关键词 床板珊瑚形珊瑚,数值分类

古生物化石的合理分类是古生物学研究工作的基础。传统的分类往往是根据专家对所研究类别业已形成的某些特定概念进行的。由于这些特定概念并非建立在标准条件之下,所以取得的分类结果必然存在因人而异的弊端。近年来,随着电子计算机的推广应用,数值分类方法由于其客观性和可重复性,已被古生物学工作者大量引用。我们在“中国床板珊瑚形珊瑚属种自动鉴定的微机处理系统”一课题的研究中,对床板珊瑚形珊瑚亚纲 476 个属中的 2000 多个中国种采用数值分类方法按属对种进行了必要的归并和重新厘定。排除了种一级传统分类中存在的某些混乱,获得了令人满意的结果。本文重点叙述所采用的数值分类方法。

数值分类学中使用的基本分类单位称作运算分类单位 (operational taxonomic unit) 简称 OTU,任何性质或级别的分类群集称为类元 (taxon)。数值分类学即借助于数值方法根据 OTU 的鉴定特征或分类性状 (taxonomic character) 将 OTU 归类成类元。数值分类的基本程序包括以下五个步骤: (1) 选择运算分类单位 OTU; (2) 选择一组描述 OTU 的性状; (3) 估计 OTU 之间的相似性; (4) 根据相似性划分类元; (5) 类元的解释和描述。

数值分类的大多数方法已设计运用到化石分类的任一级别,其基本单位 OTU 可以是一块标本、一个种,属或更高级的单位。在我们的实际工作中,OTU 是传统分类确定的一个种或亚种的代表性标本。它被作为普通化石标本来看待,数值分类的目的是对这些标本重新进行种的审定。不过下面的有关论述并不局限于这一情况,对属或更高级单位同样适用。

* 地质行业科学技术发展基金资助项目论文

一、分类学性状

无论传统分类学还是数值分类学,对化石分类的具体程序,均以分类学性状为基础。所谓性状,其含义是一项特性或“随机体种类而异的特征”^[3]。例如,表1中说明的是床板珊瑚形珊瑚亚纲 *Mesosolenia* 属诸种的性状集。其中,床板形态即为一性状,而水平、倾斜、下凹、弯曲则为性状的不同状态 (state)。“状态”一词在此统括定性及定量的再分。

表1 *Mesosolenia* 属各种的性状集

Tab.1 Characters of the Species of Genus *Mesosolenia*

个体横切面形状	个体直径 (目)	体壁厚度 (目)	床板形态				床板间距 (目)	隔壁刺长短			隔壁刺多少			联接构造类型			联接构造直径 (目)	联接构造列数	联接构造间距 (目)
			水	倾	下	弯		长	短	无	多	少	无	角	壁	联			
多角形			平	斜	凹	曲								孔	孔	接管			

数值分类如同传统分类一样,合理选取性状是影响分类效果的关键。选取分类学性状的依据主要与其在分类鉴定上的用途有关。某些性状变化较大,对属一级的分类无用,而在描述种或亚种时可以是极为重要的。例如表1中的个体直径。另一些性状在种与种之间没有反映差异,而在属或科的分类上是有价值的。例如,复体外形在 *Mesosolenia* 属的种与种之间没有反映差异,均为块状,但它在 *Mesosolenia* 属与 *Multisoleniidae* 科中其它属的区分上却是有用的。另外,分类学性状应该在生物的整体个体发生史中都存在且可以辨认,而不是仅在某一生长阶段中存在及可以辨认。

有若干类型的性状,其性质特殊,在数值分类中需要妥善处理:

1、逻辑相关性状:如果一性状为另一性状的逻辑结果,则应将该性状予以舍弃。例如,我们在表1中采用了隔壁刺长短,便不再采用隔壁刺有无。另外,还应避免运用构成逻辑结论的数学运算来构置新性状。例如,我们在表1中采用了个体直径,便不能再采用个体周长。

2、局部相关性状:一性状并非全部,而是局部取决于另一性状,这种现象有许多种情形存在。表1中我们选取的隔壁刺长短及隔壁刺多少即为一例。当隔壁刺不存在时,这两性状相关,实际上即出现了性状的重复。对这种情形,可通过下面谈到的性状编码及加权处理来解决由于性状重复而引起的某些混乱。

3、经验相关性状:如果两性状逻辑上无关,但经验上高度相关,是否应当作两个独立的性状?对这一问题,我们采用肯定的答案。因为这两项经验相关性状很可能并非受单一因素影响,分类时同时采用这两个性状不至于遗漏有关信息。

4、不变性状:一性状在全部已知 OTU 上都不变化。显然,这样的性状不会增加有

关 OTU 相似性的信息。但从鉴定的目的出发, 不变性状纳入分类计算可获取标准化的计算结果, 这一结果在进行未知 OTU 的鉴定时将是有用的。因为不能保证未知 OTU 的相应性状与已知 OTU 的性状相同。

二、性状的编码及标度

性状确定以后, 如何对性状进行编码及标度是数值分类的一个重要问题。这是因为数值分类一方面需要将化石原始资料转换成适合于计算机处理的形式, 另一方面又必须保留分类时所需要考虑的信息。编码指的是对性状作逻辑推断及区分, 而标度则是对性状作某种形式的数学变换。

性状可细分为二态性状、定量性状及多态性状, 其中多态性状还可再分为定序的和定性的。

二态性状通常指可作有或无、是或否回答的性状, 采用编码 1 或 0 表示, 通常将“有”或“是”记录为 1, “无”或“否”记录为 0。

定量性状是用连续量来度量的性状, 可用一数值表达。例如表 1 中的个体直径等。在实际工作中, 我们根据定量性状的特点, 分别取其下限值和上限值, 作为两个定量性状来考虑。这样做可更贴切地反映性状变化大小的信息, 比采用单一值例如平均值, 更符合实际。

定序多态性状由若干不同状态构成, 虽然这些状态没有具体的数值, 但能按某种次序进行排列, 其编码的办法是对每个状态按相应的次序确定一个有序的量化等级, 借以反映同一性状不同状态之间差异的信息。例如对表 1 中的隔壁刺长短及隔壁刺多少, 可按下面的方法进行编码 (表 2、表 3)。编码的原则是这样的: 通常, 隔壁刺长短的状态可为长、中、短、无, 而隔壁刺多少的状态则为多、中、少、无, 对隔壁刺长短的状态长、中、短编码为 3、2、1, 对隔壁刺多少的状态多、中、少也编码为 3、2、1 显然都能较为合理地反映同一性状不同状态之间的相对差异。当隔壁刺不存在时, 显然与具有隔壁刺有较大差异, 即使隔壁刺既短又少。但考虑到隔壁刺长短及隔壁刺多少的局部相关性, 我们对两性状的状态无均编码为 0, 这也比较合理。又鉴于 *Mesosolenia* 属各种在两性状上均无状态“中”, 我们便都不考虑这一状态。

表 2 隔壁刺长短的编码

Tab.2 Coding to the Length of the Septal Spine

状 态	多	少	无
编 码	3	1	0

定性多态性状也由多个状态组成, 但这些状态一般不能按某种明确的次序进行排列, 其编码的办法通常是对每个状态各用一符号来表示。实际工作中, 直接利用这一方法进行编码存在一定困难, 因为对任一 OTU, 无序多态性状的具体表现在很多情况下不止一个状态。例如 *Mesosolenia* 属某一种的床板形态既可以是水平的同时又可以是下凹的。针对

这一情况,我们将无序多态性状转化为一列二态性状,然后根据二态性状进行编码。如表 1 中的床板形态,可分解成床板水平、床板倾斜、床板下凹、床板弯曲四个二态性状,对其中的每一个二态性状均可用是或否作出回答,这是很方便的。这种编码方法还为以后 OTU 之间相似性的估计带来很多便利。

性状编码完毕后,下一步的工作是性状的标度。我们首先注意这样的事实:*Mesosolenia* 属各种的个体直径下限的变化范围为 0.3 至 0.75mm,上限的变化范围为 0.55 至 1.2mm,而体壁厚度下限的变化范围为 0.03 至 0.07mm,上限的变化范围为 0.06 至 0.2mm (见表 4)。个体直径与体壁厚度在数值上的差别是相当明显的。如果直接利用这些性状数据进行种间相似性估计的计算,势必使这些性状在分类中所起的作用产生严重的不均衡现象,从而难以得出正确的分类结果,因此必须合理解决性状的标度问题。

鉴于床板珊瑚任一 OTU 的诸性状普遍具有量纲和数值差异较大的特点,我们认为,对性状数据采用极差变换比较合适。极差变换即将性状数据减去性状的最小值,其结果再除以性状最大值与最小值的差:

$$X' = (X - X_{\min}) / (X_{\max} - X_{\min})$$

性状数据经极差变换后,其变化范围在 0 至 1 之间,且为无量纲数据。这一变换消除了由于性状的大小及可变性不一可能对分类结果产生的不良影响。实际工作中,我们仅对定量性状及定序多态性状数据进行了极差变换,而对二态性状数据及按一系列二态性状进行编码的定性多态性状数据,则没有必要进行极差变换。

三、相似性估计

OTU 之间的相似性估计是数值分类学上最重要、最根本的一个环节。

为估计 OTU 之间的相似性,我们将经编码及标度处理后的性状数据排列成 $(X'_{ij})_{n \times t}$ 矩阵形式,其 t 列表示需根据相似性归类的 t 个 OTU, n 行则表示 n 项性状, X'_{ij} 表示第 j OTU 第 i 性状对应的数据。

两个 OTU 间的相似性一般借助相似性系数来估计,该系数表现为上述数据矩阵中两个 OTU 对应的两列元素间相似性的定量化值,其矩阵形式 $(S_{jk})_{t \times t}$ 称为相似性矩阵。这一矩阵是一 $t \times t$ 阶方阵, t 为 OTU 的数目,其中 S_{jk} 为第 j 个 OTU 与第 k 个 OTU 之间的相似性系数。矩阵对角线上的元素代表 OTU 与它自身的相似性。就大多数相似性系数而言,认为任一 OTU 与它自身是等同的,因而主对角线上的元素代表同一性。

实际工作中,在计算相似性系数时,需要将不同类型的性状数据组合在一起。显然,对于二态性状和定量性状数据采用距离系数加以组合是十分方便的。而当多态性状为定序时,也能利用距离系数,这时性状的比较可以反映不相符的程度,而非简单的匹配(相符)或错配(不相符)。当多态性状为定序时,两个 OTU 在多态性状上的匹配可以为是它们在相应的一系列二态性状上的匹配。这样,我们便采用如下形式的距离系数估计第 j 个 OTU 与第 k 个 OTU 之间的相似性:

$$D_{jk} = \frac{\sum_{i=1}^n W_{ijk} |X'_{ij} - X'_{ik}|}{\sum_{i=1}^n W_{ijk}}$$

其中 W_{ijk} 为性状权。

我们认为,在对性状进行编码及标度后,还必须引入性状的加权,以区别各性状在分类中的不同重要程度。例如,表1中的个体直径与联接构造直径两性状,它们在种的分类中的重要程度显然是不同的。实际工作中,我们采用专家评议的方法对性状加权。从分类效果看,经专家评议确定的性状权基本上是合适的。

表1中各性状的权在表4中予以说明。这里需要特别指出的是,我们赋予个体直径的上、下限较大的权,是为了突出强调其在分类中的重要性。

上述距离系数实际为 Gower 一般相似性系数的一种具体转变形式。根据由上述距离系数构成的相似性矩阵,使得以点集形式在欧几里德空间中表示 t 个 OTU 成为可能(即相似性矩阵为半正定)。这样,我们便可形象地应用距离值 D 来反映各点间的亲疏程度,作为 OTU 分类的依据。显然,两 OTU 之间的距离值 D 愈小,则愈相似,这与习惯的距离概念一致。

四、聚类分析

数值分类一般以相似性矩阵为基础,类元的划分是通过揭示和概括这一矩阵的结构而进行的。从 $t \times t$ 的相似性矩阵出发,对 t 个 OTU 归类成类元的方法称为聚类分析。实际工作中常用的聚类方法是系统聚类法,即在聚类过程中,对所有的 OTU 按距离准则逐步聚类,类别由多到少,直到获取合适的类元为止。它的具体算法由以下四步进行迭代。

第一步:设共有 t 个 OTU,每个 OTU 自成一类,即建立 t 类: $G_1^{(0)}, G_2^{(0)}, \dots, G_t^{(0)}$ 。计算各类之间(即各 OTU 之间)的距离,可得 $t \times t$ 的距离矩阵 $D^{(0)}$,其上角标号(0)表示聚类运算处于起始状态。

第二步:如果在前一步聚类运算中已求得距离矩阵 $D^{(n)}$ (n 为逐步聚类合并的次数),则求 $D^{(n)}$ 的最小元素。如果它是 $G_i^{(n)}$ 与 $G_j^{(n)}$ 之间的距离,则将 $G_i^{(n)}$ 和 $G_j^{(n)}$ 两类进行合并,成为一类 $G_i^{(n+1)}$ ($i < j$),其余各类不变。因此建立新的分类 $G_1^{(n+1)}, G_2^{(n+1)}, \dots, G_{i-(n+1)}^{(n+1)}$ 。

第三步:计算合并后形成的新类元之间的距离,得矩阵 $D^{(n+1)}$,在计算 $G_i^{(n+1)}$ 与其它没有发生合并的 $G_1^{(n+1)}, G_2^{(n+1)}, \dots, G_{i-(n+1)}^{(n+1)}$ 之间的距离时,有多种计算方法,应用不同方法得出的分类结果不尽一致。实际工作中,我们采用了多种方法进行计算比较,最终选用了类平均距离法,具体说明如下:

设 H 和 K 为两个聚类, H 类中有 n_H 个 OTU, K 类中有 n_K 个 OTU, K 类由 I 和 J 两类合并而成, I 类中有 n_I 个 OTU, J 类中有 n_J 个 OTU,显然两类合并后的类中共有 $n_I + n_J$ 个 OTU,即 $n_K = n_I + n_J$ 。 H 类 K 类间的距离用两类间所有 OTU 之间距离的平均值来表示

$$D_{H,K} = \frac{1}{n_H n_K} \sum_{i \in H, j \in K} D_{i,j}$$

不难得出类平均距离 $D_{H,K}$ 的递推式为:

$$D_{H,K} = \frac{n_I}{n_I + n_J} D_{H,I} + \frac{n_J}{n_I + n_J} D_{H,J}$$

根据此式, 在计算新的距离矩阵时, 可直接利用前一次矩阵的计算结果。

第四步: 跳到第二步, 重复计算及合并, 可一直将全部 OTU 聚成一类为止。也可取距离阈值 D , 当 $D^{(n)}$ 的最小元素超过 D 时, 运算停止, 所得分类即聚类结果。

实际工作中, 我们首先将全部 OTU 聚成一类, 然后根据逐步聚类的结果, 经专家分析, 选取一个合理的距离阈值 D , 当逐步骤类过程中两类之间的距离超过 D 时, 便不再进行合并。由此获得 OTU 的最终分类, 同时也确定了合适的类元。在确定距离阈值 D 时, 除深入分析 OTU 之间的相似性及差异外, 还兼顾到 OTU 的地理、地层时代等方面的因素。

五、数值分类实例及命名问题

表 4 列出了 *Mesosolenia* 属原定 7 个种共 8 个标本的性状编码数据、性状权及数值分类结果。原来的 7 个种现被归并成 4 个种。

分类的改变不可避免地会导致命名的变化。数值分类结果对原有命名的影响通过对原有分类的订正来体现: 如果原定的几个种的标本通过数值分类被归并成一个种, 则该种由原定的几个种的特征来共同体现, 原有各种的描述同等重要。对该种的定名遵循古动物命名的优先律原则, 其余的几个原定种被认为是该种种名持有者的同义名种。如果原先被确定为一个种的两个标本通过数值分类被分定成两个种, 则其中的非正型标本将重新定名。

根据数值分类结果, 表 4 中原定的 2、4 及 5 号种的标本被归并成一个种, 它们的各项性状均基本一致。据优先律, 4 号种 *M. multisoleniformis* Lin (1975) 及 5 号种 *M. multispinosa* Lin (1975) 应为 2 号种 *M. biformis* C.M. Yu (1962) 的同义名种。表 4 中原定的 1 号及 6 号种的标本亦被归并, 6 号种 *M. ? ningchiangensis* Yang (1975) 应为 1 号种 *M. festivus* Tchernychev (1951) 的同义名种。表 4 中定的 7 号及 8 号同一种两标本的归并说明金淳泰 1978 年鉴定的中国标本 *M. prima* Sokolov et Tecakov (1963) 是可靠的。

表4 *Mesosolenia* 属原定各种性状编码数据及数值分类结果Tab.4 Data of the Coded Characters of the Classical Species of Genus *Mesosolenia* and the Result of Numerical Taxonomy

现定种号				1		2			3		4
原定种号				1	6	2	4	5	7	8	3
性状及 性状权	原 定 种 名			<i>M. festinus</i> (Tcherny- chev, 1951) **	<i>M. ?</i> <i>ningchiange</i> <i>nsis</i> (Yang, 1975) *	<i>M. bifurcata</i> C. M. Yu, 1962 *	<i>M. multispinosa</i> <i>seniformis</i> Lin in Li et al., 1975 *	<i>M. multispinosa</i> Lin in Liet al., 1975 *	<i>M. prima</i> Sokolov et Tecakov, 1963 *	<i>M. prima</i> Sokolov et Tecakov, 1963	<i>M. hubeiensis</i> Wu in Jia et al., 1977 *
个 体	切面	多角形	0.25	1	1	1	0	0	0	1	1
		角圆形	0.25	1	0	1	1	1	1	1	0
	直径	下限(mm)	1.00	0.40	0.40	0.45	0.45	0.30	0.70	0.75	0.50
		上限(mm)	1.00	0.60	0.55	0.75	0.60	0.60	1.00	1.20	1.20
体 壁 厚 度	下限(mm)		0.50	0.06	0.03	0.03	0.07	0.07	0.07	0.05	0.04
	上限(mm)		0.50	0.06	0.05	0.06	0.10	0.10	0.07	0.10	0.07
床 板	形态	水平	0.25	1	1	1	1	1	1	1	1
		倾斜	0.25	0	1	0	0	0	1	0	0
		下凹	0.25	0	0	1	1	1	0	1	1
		弯曲	0.25	1	0	0	0	0	1	0	0
	间距	下限(mm)	0.50	0.20	0.20	0.28	0.40	0.30	0.10	0.15	0.30
		上限(mm)	0.50	0.30	0.30	0.42	0.50	0.40	0.30	0.50	0.40
隔 壁 刺	长短		0.50	1	1	1	1	1	1	1	0
	多少		0.50	3	3	1	1	1	1	1	0
联 接 构 造	类型	角孔	0.33	1	1	1	1	1	1	1	0
		壁孔	0.33	1	1	1	1	1	1	1	1
		联接管	0.33	0	0	0	0	0	0	0	1
	直径	下限(mm)	0.25	0.15	0.10	0.13	0.20	0.12	0.22	0.25	0.18
		上限(mm)	0.25	0.24	0.10	0.22	0.40	0.20	0.22	0.40	0.20

注: 表中*表示正型种, **表示属型种。

结束语

数值分类学在古生物学中应用的最大困难是将化石形态资料转换为机器可以识别的数据形式。只要这一困难存在,人的因素在分类中的影响仍将是重要的。正确解决这一问题的途径是,在保证研究成果的科学性的前提下,选择最适合于计算机工作的数据形式。

在数值分类学研究成果的应用中,古生物学者更关心的是分类的稳定性。这可从以下两个方面进行查证。

1.以新性状形式积累更多的信息,如果原先已根据众多的性状作出相似性估计,则增加更多新的性状,OTU 间的相似性应变化不大;

2.在后继的研究中,如果增加新的 OTU,应用原先的相似性水平及标准(阈值),可能会得出新的不同类元,但业已确定的类元不应有较大改变。

由于受各方面因素的限制,本文尚未考虑的问题还很多,诸如分类学性状必需的数目,相似性的显著性检验、聚类结果的数学评价等。另外,当一已知 OTU 的某一性状因无法观察或无法测量而缺失时,如何对这一 OTU 与其它 OTU 之间进行比较等问题也未作深入研究,只是简单地采用要么对该性状弃之不用,要么不考虑这一 OTU 的办法,这显然不是十分完善的。

当前,化石数值分类工作,在较大程度上还处于探索性阶段,本文采用的技术方法是新颖的,其效果究竟如何,还有待于实践的进一步检验。文中不当之处,敬请读者批评指正。

主要参考文献

- [1] 林宝玉,池永一等,1988,床板珊瑚形珊瑚,地质出版社。
- [2] [美] D.M.劳普, S.M.斯坦利著,武汉地质学院古生物教研室译,1978,古生物学原理,地质出版社。
- [3] [英] P.史尼斯, [美] R.索卡尔著,赵钱桥译,1984,数值分类学,科学出版社。
- [4] 蔡元龙编,1986,模式识别,西北电讯工程学院出版社。
- [5] 严幼因等,1988, *Favosites* 属的 Q 型聚类分析,古生物学报,第 27 卷第 4 期,498-513 页。

STUDY OF THE NUMERICAL TAXONOMY OF TABULATE MORPH CORALS*

Zhang Song-lin

(IGMR, Nanjing)

Abstract

This paper discusses the numerical taxonomy of tabulate morph corals systematically. The work was carried out on the representative fossil specimens of the species of each classical genus. According to the requirements of numerical taxonomy, the effective identifying characteristics were selected as taxonomic characters based on the detailed study of the description of the fossil specimens. The characters were coded and scaled, then the similarity between every two specimens was estimated by the Weighted Minkowski Distance, finally the species were reset by the Hierarchical Clustering Method according to the similarities.

Key Words Tabulate Morph Corals, Numerical Taxonomy

* The subsidized project paper of the fund about the development of science and technology of geological profession.