

初探电子计算机在孢粉鉴定中的应用

陈民敏

(地质矿产部石油地质中心实验室)

据不完全统计,目前化石孢粉形态描述已发表了3000多个属,20000个以上化石种,还在以每周11个新孢粉型增加着,它们以不同的语种散布在200多种期刊上。孢粉工作者用在查找和阅读文献上的时间比用在研究化石孢粉粒的时间要多得多。因此,要使孢粉工作者从浩如瀚海的文献堆中解放出来,必须依靠电子计算机高速查找和惊人的记忆(贮存)本领了。美国孢粉学家G.O.W.Kremp就说过:电子计算机能够把孢粉学从分类的混乱中拯救出来!纵观目前国内外古生物学应用电子计算机的情况,大多是用来整理资料,而用电子计算机帮助鉴定化石孢粉粒方面,在国外有少数石油公司的孢粉工作者和数学工作者合作进行应用研究,取得了一定的效果。笔者近几年来对这个课题做了一些尝试,它包括三个部分内容:一、孢粉形态数字代码系统,二、孢粉鉴定的数学模拟,三、孢粉鉴定的电子计算机模拟技术。

一、孢粉形态数字代码系统简介

为了适应电子计算机以数字形式的量值运算的特点,就要把用语言描述的属种特征转变为数字代码。这里介绍的主要是依据(B.I.P.M)J.H. Germeraad & J. Muller 1970年设计的数字代码系统,这个代码系统的优点是避免目前孢粉学在分类上和名词术语上的不统一,用几何学的方法设计了一个P—A—I直角坐标系(图1),对孢粉粒的基本形态特征:极性、对称性、形状、口器、纹饰、大小等构造单元进行编码。

对于形状(包括孢粉粒形状,口器形状,基柱棒形状)的编码是根据对称性分为:不对称形状,两侧对称形状,(高于两侧对称的)高级对称形状;再根据外形轮廓的凸出,缢缩,顶、底的尖圆,边棱的尖圆等编成一行连续写的数字代码,每一位数字代表一项特征。

对于孢壁层的编码是用一组两位数字组成。第一位数字代表该层是第几层,在最里面的一层基柱棒层的基部设想有一个基准线(面),从基准线(面)开始向外数依次为第1层,第2层……,从基准线(面)开始向内数依次为第9层、第8层、……。第二位数字代表某一层是具基柱棒的构造层(代码“1”),还是均质层(代码“2”),还是空隙(代码“0”)。比如菊科花粉粒的孢壁编码如图2:

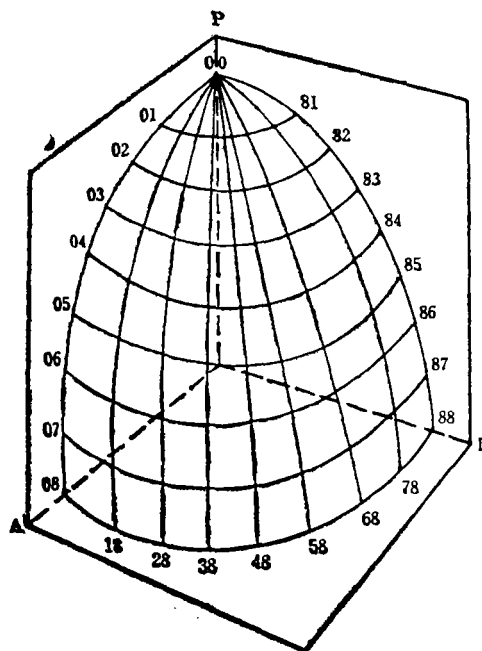


图1 P—A—I直角坐标系图 (据J.H. Germeraad & J. Muller, 1970年)

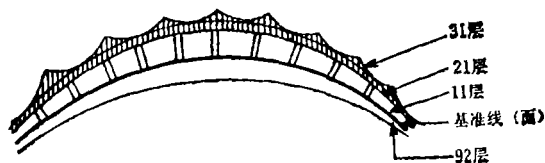


图2 菊科花粉粒孢壁层编码图

对于每一层面上的形态单元的编码是根据该形态单元的中心在P—A—I直角坐标系上的坐标位置来确定。一般说来,在每一层面上都可以划分出“口器区”,“过渡区”“剩余区”等单元,分别从它们的位置和排列形式编码。

对于每一个区中的次一级形态单元,主要指的是纹饰图案特点,其编码是根据排列形式(散乱状、线状、分叉状和网状等),条理化程度,系列性,图案中的次级单元数,对称形,随机性等编成一行连续写的数字,每一位数字代表一项特征。

这些代码称之为原码(Primary code)。形状、区、纹饰的原码往往长达11—12位数字,使用不便。因此,对应着它们的每一个原码都有一个辅码(Secondary code),辅码一般只有4位数字。电子计算机里贮存的是原码,孢粉工作者实际使用的是辅码,原码和辅码的变换是由电子计算机自动完成的(图3)。

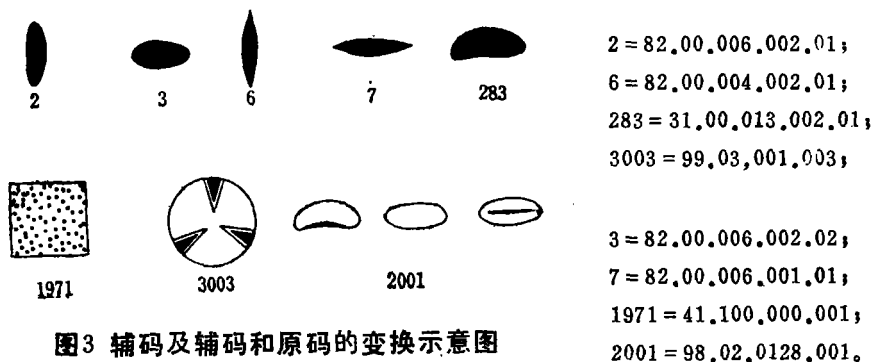


图3 辅码及辅码和原码的变换示意图
(据J.H. Germeraad & J. Muller 1970)

二、孢粉鉴定的数学模拟

模型,是一个理想的模拟物。它集中了所研究对象的全部有关特性。比如,研究一个古生物属种,往往在一块标本上不具有它的全部特征,但通过对它的正型、副型、辅型等许多标本的研究,对这一古生物属种的特征就能比较全面的了解。据此,就可以建立一个理想的模型,它集中了这一属种的全部特征。

数学模型,把模型的全部特征用数学语言来描述,可以看成是一个有序集合或特征向量,每一项特征就是这个集合(向量)中的一个元素。假设用 A 代表属种的全部特征集合, $a_1, a_2, \dots, a_i, \dots, a_n$ 代表每一项特征,用数学式表示:

$$A = \{ a_1, a_2, \dots, a_i, \dots, a_n \}$$

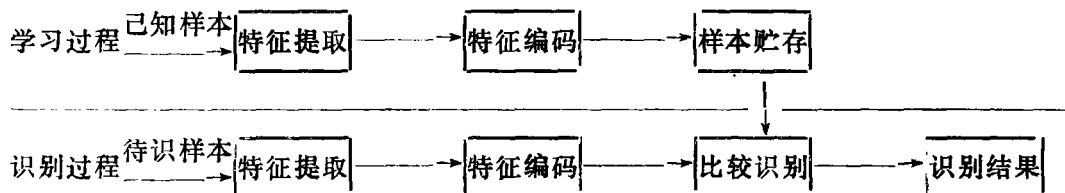
简写为 $A = \{ a_i \mid a_i \in A \}$

或是 n 维特征向量:

$$A = (a_1, a_2, \dots, a_i, \dots, a_n)$$

式中的每一项元素(特征)都是有序的,确定的,不重复的。

古生物鉴定用数学模拟就是“模式识别”(Pattern recognition)问题,其基本原理是“模板匹配原理”。模板就是用以比较的标准,就是数学模型。每一个已知属种都可以构成一个模板,把这些模板事先贮存在电子计算机中,叫做计算机的训练和学习过程,当一个待定属种的特征向量输入计算机时,就和原先存放在计算机中的模板相比较,看它与哪一个已知属(种)的模板相同或最相似。这叫做计算机的识别过程。这一识别系统用框图表如下:



但是,在实际鉴定工作中往往遇到待定属种与几个模板都相似,只是相似程度有所不同。不是“非此即彼”,而是“亦此亦彼”。用经典数学的“0”(否定),“1”(肯定)两种形式是不能表达的。如何来表达这种差异程度呢?用“模糊数学”的隶属(从)属度来刻画最为合适。

假设有 n 个模板,它是论域 U 上的 n 个模糊子集: $A_1, A_2, \dots, A_i, \dots, A_n$ 。现在 U 上有一具体的待识样本 u_0 ,问它属于哪一个模板?按最大隶属度原则:

$$\theta = \max \{ \mu_{A_1}(u_0), \mu_{A_2}(u_0), \dots, \mu_{A_i}(u_0), \dots, \mu_{A_n}(u_0) \}$$

如果 $\mu_{A_1}(u_0) = \theta$

则认为 u_0 隶属于 A_1 。

上式的意思是分别计算出被识别对象 u_0 对于模板 $A_1, A_2, \dots, A_i, \dots, A_n$ 的隶属度,值最大的,就认为 u_0 隶属于它。

这个方法是很简单、直接的。现在的问题是确定合适的隶属函数。笔者使用逻辑代数这一数学工具,将模板的特征向量 $A_i = (a_i^1, a_i^2, \dots, a_i^1 \dots a_i^m)$ 的参数与待识样本的特征向量 $u_0 = (u_0^1, u_0^2, \dots, u_0^j, \dots, u_0^m)$ 的参数逐位进行模2加,记为 $\oplus$, 根据二进制加法法则,当两个特征参数相同时,它们的和为0,当两个特征参数相异时,它们的和为1。即:

$$A_i^j \oplus u_0^j = \begin{cases} 0 & A_i^j \text{ 与 } u_0^j \text{ 相同时} \\ 1 & A_i^j \text{ 与 } u_0^j \text{ 相异时} \end{cases}$$

在两个特征向量的参数逐位进行模2加的结果中,1的个数就是不同的特征数,记为:

$$D_i = \sum_{j=1}^m A_i^j \oplus u_0^j$$

在与所有模板比较的结果,其中 D_i 最小,就是相同特征最多的,其最大隶属度可表示为:

$$\theta = \max(m - D_i)$$

即可确定待识样本 U_0 隶属于模板 A_i 。

三、电子计算机模拟技术

假设已知鳞毛蕨属孢的两个种(*Dryopteris desvauxii*和*Dryopteris chinensis*)和一个待定种(*Dryopteris sp.*)的相应的种征,用文件名PSMD把这两个已知种的特征代码以关键字为DD和DC的两个记录(当然可以有多个记录)贮存在磁盘上作为模式识别的模板,它们是:

DD = (283, 283, 3, 46, 27, 30, 3, 1, 99999, 2, 92, 1, 2001, 2, 4007, 25, 24, 299, 7, 1, 82, 1, 2001, 2, 400, 7, 25, 24, 299, 7, 1,)

DC = (283, 283, 3, 60, 30, 33, 4, 1, 99999, 2, 92, 1, 2001, 2, 400, 7, 25, 24, 299, 7, 1, 82, 1, 2001, 2, 400, 7, 25, 24, 299, 7, 1,)

待定种 *Dryopteris sp.*的特征代码是:

DX = (283, 283, 3, 60, 30, 33, 4, 1, 99999, 2, 92, 1, 2001, 2, 400, 7, 25, 24, 299, 7, 1, 82, 1, 2001, 2, 400, 7, 25, 24, 299, 7, 1,)

将其读入电子计算机,根据“模板匹配”原理,用模2加法,求出这个待定种与所有已知种的隶属度,值最大者,即隶属于该已知种。这一过程用电子计算机模拟的程序框图如图4所示。

框图中:

DX(N)代表*Dryopteris sp.*,括号中的N代表它的全部特征。

DD(I)代表*Dryopteris desvauxii*,括号中的I代表它的每一项特征。

DC(I)代表*Dryopteris chinensis*,括号中的I代表它的每一项特征。

XD代表*Dryopteris sp.*和*Dryopteris desvauxii*的相同特征数。

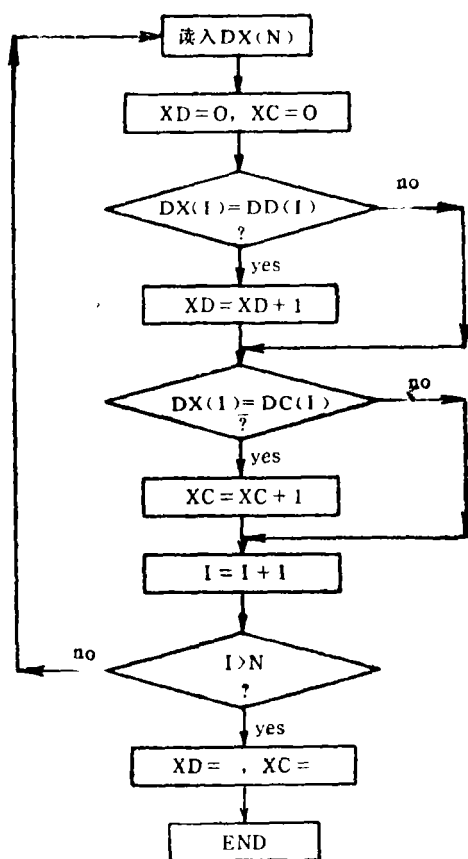


图4 电子计算机模拟程序框图

XC代表*Dryopteris* sp.和*Dryopteris chinensis*的相同特征数。

计算结果表明*Dryopteris* sp*和*Dryopteris desvauxii*只有28项特征相同,而和*Dryopteris chinensis*有32项特征相同。用户就可以知道这一待定种更相似于*Dryopteris chinensis*。

如果用数据库管理系统,其存储量将大大增加,数据的共享性将更高,使用也更方便。

笔者在工作中得到中国建筑科学研究院抗震研究所李中锡同志和我室丁太木,朱春才,罗秋霞、周锦南等同志的帮助,深表感谢。

(收稿日期:1985年1月18日)

参 考 文 献

- [1] 中国科学院北京植物研究所古植物室, 中国蕨类植物孢子形态, 科学出版社, 1976年。
- [2] 楼世博等, 模糊数学, 科学出版社, 1983年。
- [3] 施鹏飞, 模式识别, 自然杂志, 5卷1期, 1982年。
- [4] J.H.Germeraad and J.Mueller, 1970. A Computer—based Numerical Coding System for the Description of Pollen Grain and Spores. Review of Palaeobotany and Palynology.No.10,p.175—202

AN APPROACH TO COMPUTER IDENTIFICATION OF SPORO—POLLEN

Ghen Minmin

(Central Laboratory of Petroleum Geology,
Ministry of Geology and Mineral Resources)

Abstract

As an experimental report, this paper consists of three parts:

1. Numeric coding system for the description of sporo-pollen.
2. Mathematical simulation for identifying the Genus (species) of sporo-pollen.
3. The technique of computer processing for identifying the Genus (species) of sporo-pollen.

Finally, an example of two recent spores is taken to elucidate the creation of a Genus (species) file of sporo-pollen, as well as the identification simulation of computer