古生物学报, 62(4): 516–530 (2023年12月)
Acta Palaeontologica Sinica, 62(4): 516–530 (December 2023)

·研究综述·

DOI: 10.19800/j.cnki.aps.2023035

浅谈大数据语境下的古生物学研究*

黄 冰**

中国科学院南京地质古生物研究所, 现代古生物学与地层学国家重点实验室, 南京 210008

提要 近半个世纪以来, 基于数据的古生物学研究日益占据重要位置。当下的科学研究进入大数据时代已得到公认, 虽然囿于总体上非实验性学科的特点, 古生物学数据产出速度有限, 目前尚难符合大数据的多数基本特征, 但大数据时代及相关理念显然对古生物学研究产生了积极效应, 比如近年来古生物学数据产出的多元化, 数学方法与模型的复杂化均与之密不可分。本文主要基于作者研究经验, 浅议了古生物学定量化研究历史三个阶段, 同时考虑古生物不同门类数据的相通性, 在大数据语境下将古生物学数据分为结构型、半结构型与非结构型, 并简介了基本研究方法。在讨论了定量古生物学与分析古生物学两大研究视角的异同后, 基于古生物学代表期刊最新论文的小样本抽样, 本文强调了分析古生物学的研究思路与统计模型较传统统计方法的优势。近年来古生物学展示了数据驱动型研究的特点, 未来可能需要同时重视模型驱动的研究视角, 将自上而下问题导向的模型设计与自下而上基于数据的收集分析相结合, 以保证古生物学数据研究的可持续发展。此外, 古生物学不是数据密集型科学, 未来将之与地学其他领域的数据有机整合, 会促成古生物学在学科交叉方向的深入。最后, 统计学领域最新的倡议同样需要古生物学者重视, 对统计模型的选择与对数据的解读需要考虑系统的复杂性与多解性, 统计显著性相关的因果规律识别尤其需要慎重。

关键词 大数据 定量古生物学 数据驱动 分析古生物学 模型驱动

中文引用 黄冰, 2023. 浅谈大数据语境下的古生物学研究. 古生物学报, 62(4): 516–530. DOI: 10.19800/j.cnki.aps.2023035

英文引用 Huang Bing, 2023. A brief discussion on Paleontology research in the context of Big Data. Acta Palaeontologica Sinica, 62(4): 516–530. DOI: 10.19800/j.cnki.aps.2023035

A brief discussion on paleontology research in the context of Big Data

HUANG Bing

State Key Laboratory of Palaeobiology and Stratigraphy, *Nanjing Institute of Geology and Palaeontology*, *Chinese Academy of Sciences*, *Nanjing* 210008, *China*

Abstract Over the past half-century, data-based research in paleontology has increasingly assumed a prominent role. It is widely acknowledged that contemporary scientific research has entered the era of Big Data. However, owing to the inherent characteristics of non-laboratory disciplines, the rate of production of paleontological data resources is limited, making it challenging to align with the fundamental characteristics associated with Big Data temporarily. Nevertheless, the era of Big Data and its associated concepts have clearly exerted a positive influence on paleontology. For instance, recent years have witnessed the diversification of data output in paleontology, along with the inherent complexity of mathematical methods and models, which are closely linked to this era. This article, primarily based on the author's research background, offers a concise overview of the three key stages in the history of quantitative paleontological

收稿日期: 2023-09-19; 改回日期: 2023-10-24; 录用日期: 2023-10-26

* 中国科学院战略性先导科技专项(B 类) (XDB26000000)资助。

** 通讯作者: 黄冰, 研究员, e-mail: bhuang@nigpas.ac.cn

research. Simultaneously considering the commonalities among paleontological data, it categorizes paleontological data within the context of Big Data as structural, semi-structured, or non-structured, while also providing an introduction to fundamental research methodologies. Following a discussion of the similarities and differences between the two major research perspectives of quantitative paleontology and analytical paleobiology, the article emphasizes the advantages of analytical paleobiology's research methodologies and statistical models over traditional statistical approaches. In recent years, paleontology has unmistakably displayed characteristics indicative of data-driven research. However, a model-driven research perspective may be necessary. The methodology combines top-down model design with bottom-up data collection and analysis could ensure the sustainability of paleontological data research. Furthermore, given that paleontology is not inherently a data-intensive discipline, its collaboration with data from other geoscientific fields, will in turn promote the interdisciplinary growth of paleontology. Finally, the latest developments in the field of statistics also warrant the attention of paleontologists. The selection of appropriate statistical models and the nuanced interpretation of data should account for the inherent complexity and potential multiple solutions within paleontological studies. Particular caution should be exercised when identifying causal relationships related to statistical significance.

Key words　Big Data, numerical paleontology, data driven, analytical paleobiology, model driven

1　引　言

自20世纪末以来，古生物学取得了一系列突破性认识，这些重要成果一方面是基于关键化石的发现，另一方面也离不开数据的分析与解读。毋庸置疑，古生物学已经从一个世纪或更早以前的描述性"学科"，过渡为与数据关系密切的"科学"。从数据集(Data set)到数据库(Data base)，古生物最终开始与大数据(Big Data)产生关联。虽然"大数据"概念提出并不算久远(Doctorow, 2008)，但它已经几乎渗透进各个学科和领域，并成为最为热门词汇之一，甚至主导行业方向。那么在大数据时代，古生物学又如何定位并响应这一浪潮呢？

"大数据"一词由于过于熟悉而有时模糊了它的涵义。关于大数据的特性，最早被提出高速(Velocity)，多样化(Variety)，以及大量(Volume)，被称为大数据的3V特征，之后又加入了价值(Value)，即4V特征(Mayer-Schönberger and Cukier, 2014)。虽然之后有新特征加入，如准确性(Veracity)，动态性(Vitality)，可视化(Visualization)等，但4V特征得到广泛认可。古生物学数据可以满足多样化及价值特征，但以大数据的命名特征即大量(Volume)来说，以TB(即1024 GB)级别以上的数据量，古生物学当下并不容易满足(例如PBDB或GBDB数据库仅是有限的GB级别)。而高速(Velocity)即数据流的吞吐速度，虽然古生物学数据资源近年来明显增长，但由于其非实验室科学本质，数据产出速度有限，该特征很难得以满足。因此严谨地讲，由于目前学科本身数据的限制，古生物学科暂时还不存在严格意义下的大数据。

值得期待的是，随着地学其他领域数据，特别是涉及深时(deep time)的相关大数据的发展(沉积学、地球化学、地质年代学甚至构造地质学等)，学科交叉的古生物学大数据研究大有可为。此外，大数据时代的新方法理念甚至研究范式，都对古生物学产生积极效应。古生物学自身限制并不意味着它要脱离大数据时代的"潮流"。该意义上，了解大数据语境下的古生物研究就十分必要。

本文主要基于作者个人研究过程中对古生物学定量化发展的初步理解，尝试探讨数据驱动与模型驱动的古生物学研究，并首次介绍分析古生物学(Analytical Paleobiology)的基本概念和研究方法，最终对大数据语境下的古生物学研究谈一些粗浅认识。作者对古脊椎动物与古植物等相关定量研究了解极为有限，但考虑到古生物学数据类型和方法的相通性，本文拟从极为有限的研究经验和视野入手，结合个人近年讲授的研究生课程资料，主要面向刚进入古生物学领域的研究生和青年科研人员谈一些体会，希望能够抛砖引玉，对读者有所启发，也供同行批评。

2　定量古生物学的发展历程

2.1　基础数据的描述性分析

古生物学早期文献中绝大多数几乎看不到数据，以描述性文字及图版为主。以腕足动物为例，

最早涉及到的研究(Duméril, 1806)及其后百年, 未能发现涉及数据的文献。作者了解到有腕足动物标本长宽数据的研究始于20世纪初(如Greene, 1908), 20世纪40年代才有多个形态参数的度量数据表(Bancroft, 1946), 以及最简单的最小二乘回归分析(Newell, 1949)。 直到20世纪80年代, 线性回归一直是腕足动物研究的主要定量方法(如Lenz, 1967; Grant, 1972; Jones and Smith, 1985等), 结合数据直方图(Histogram)的绘制(Worsley and Broadhust, 1975)也是较为常见的方式。甚至21世纪的系统分类研究中, 仍然离不开回归分析, 只是在算法上更为多样化, 如压缩主轴法(RMA), 以及鲁棒回归(黄冰, 2007)等。

由于作者专业所限, 对其他门类文献在小样本抽样中发现除了类似趋势外, 也有自身特色的研究, 如对菊石缝合线的定量化描述从很早就开始(如Furnish and Unklesbay, 1940), 甚至在分形几何正式提出后没多久就有相关的缝合线分形结构的研究(Long, 1985)。更为早期的, 近2个世纪之前研究腹足类壳的形态时就已涉及几何形态测量学(Moseley, 1838), 该开创性研究明确提出运用数学原理(mathematical principles), 最终发现腹足类壳的对数或等角螺旋(logarithmic or equiangular spiral)生长规律。相关研究被古生物学家用计算机重现已是其一个多世纪之后了(Raup, 1967)。

虽然最早的古生物多样性曲线是一项极具前瞻性的工作(Phillips, 1860), 但该曲线并没有确切的数据支持(无坐标轴数据), 只是描述性的趋势展示。类似的, 古生物学早期研究以分类为主, 定量化研究手段也多基于形态度量数据, 大多属于描述性统计(descriptive statistics), 主要涉及数据简单处理, 图表描述与基本分析, 目的是呈现数据的一般趋势。早期研究中数据量小而简, 除个别研究(如Raup, 1967; Gould and Katz, 1975), 均可通过手算或计算器辅助计算, 同时由于较其他基于实验数据的学科在统计分析要求上的滞后性, 一直以描述性分析占主导。

运用生物学方法研究古生物化石是一个很早的思路, 在20世纪70年代甚至已成为一种潮流, 也有不少专著发表(如Schopf, 1974; Boucot, 1975), 但这种思路真正作为一种研究理念, 离不开化石生物学(Paleobiology, 也常被译为现代古生物学、理论古生物学, 或古生物学)。《古生物学名词》中对该术语解释为, 视化石为地史时期的生物实体, 强调从生物学的角度, 运用生物学的原理和方法来研究化石, 并注重探讨生物与其生活环境关系的学科(方宗杰、杨群, 2009)。期刊*Paleobiology*创刊第一期就体现了该研究理念, 如基于计算机编程的托盘海绵形态学研究(Gould and Katz, 1975), 双壳类的流体力学实验及定量分析(Stanley, 1975)、形态演化速率(Schopf *et al*., 1975)和物种生存曲线(Raup, 1975)等。化石生物学或强调从生物学角度研究化石的古生物学的产生, 似乎可以视为古生物学定量化研究开始的重要里程碑。

2.2 计算机参与的统计分析

古生物学由于学科数据的特点和研究传统, 相对较晚使用计算机。最早通过编程进行的研究是对化石的壳(腹足类为主)进行三维建模分析(Raup, 1967), 这一研究以当时的计算机软硬件条件来说几乎达到了资源的最优配置。在之后的十余年内, 都很少有编程的相关研究, 直到化石生物学与*Paleobiology*期刊的出现。

古生物学多样性研究的开创性工作(Sepkoski *et al*., 1981)用相关性分析(correlation analysis)得出重要结论, 即显生宙海洋目级分类单元多样性数据存在共变模式, 从此通过对多样性数据相关的宏演化研究拉开序幕。Sepkoski (1984)最早识别出五次生物大灭绝事件, 是计算机参与统计分析在古生物学应用的早期经典。从此, 古生物学这门与地质历史相关的学科, 也如同历史被分为“History”和“histories”一般, 在定量化研究领域似乎可以被分为“Paleobiology”与“paleobiologies”两类, 前者研究多个门类的多样性变化, 如辐射、灭绝等宏大的规律, 后者聚焦单一门类, 对化石居群或组合进行形态或古生态相关研究, 两者大体上可对应宏演化与微演化。

20世纪90年代迎来了古生物学定量化研究的第一次高潮, 先驱工作激发了广大古生物学者的

兴趣。古生物学家不再是物理学家所诟病的“集邮者”, 数据分析与解读成为探索演化规律的重要依据。将今论古, 借鉴生物学已有方法是当时的常规手段(如Dodd and Stanton, 1990)。由于编程计算有一定门槛, 有学者设计出一些定量分析程序(如Tipper, 1991; Temple, 1992)专门供古生物学研究者使用。PAST (PAlaeontological STatistics)作为第一个较为全面的古生物学专用的免费统计软件也终于问世(Hammer *et al*., 2001)。PAST在功能设置上有较强的针对性, 除“常规统计”“多元统计”之外, 还设置了“多样性分析”“形态学分析”“地层学分析”及“时间序列分析”等专项功能菜单(详见黄冰等, 2013)。此外, 用户可向开发者提出要求和建议, 作者也曾为该软件回归分析提供算法并被采用(1.95版)。PAST最新一版是4.08 (2021年11月后暂未更新), 包括近200个子程序涉及上百种数理统计过程, 能满足绝大部分的古生物学定量分析需求。

21世纪初, 由于PAST普及的局限, 一些专业数据分析软件也是古生物学者的选择, 如Matlab, Origin, SPSS, 甚至Mathematica与SAS等。其中SPSS (Statistical Package for the Social Sciences)因其操作简单, Origin因其成图美观而得到相对较多使用, 而功能极为强大的Matlab与统计学最大商业软件SAS因编程复杂而受到较少关注。上述软件均价格不菲, 从而使用有限, 开源分析软件和平台的匮乏直到R语言的出现才得到缓解。

R是一种用于统计计算和数据分析的编程语言和环境(简介与案例见黄冰, 2015)。它与SAS有渊源, 但R是免费开源的, 由统计学家和计算机科学家共同开发和维护。古生物学者较早就开始使用R (如Hunt, 2006; Novack-Gottshall, 2007等)。虽然R有一定使用门槛, 但其集成开发环境(IDE) RStudio友好, 特别是R生态系统中有大量的扩展包(packages), 使R语言并不难掌握。R的扩展包极为多元, 如“ggplot2”用于创建高质量的数据可视化, “stats”用于基本统计, “caret”用于机器学习, “igraph”用于网络分析等等。有些包也开始高度集成化, 如“tidyverse”包括80个左右的小程序包, 可满足多数统计分析研究。此外, 古生物学者自己也可开发相关的包, 如最早的PaleoTS包(Hunt, 2006), 以及较近发布的包DivDyn (分析PBDB数据, Kocsis *et al*., 2019), 和越来越多的如PlaeoDB, Palaeoverse, Paleomorph等古生物学相关包。有了它们, 古生物学家的编程工作量得到极大压缩。可以说, R语言的使用开启了古生物学定量化研究的新时代。

正是由于计算机的参与, 古生物学工作者开始真正与数据打交道。严谨的统计学分析使古生物学成为一门有更为客观指标的科学, 其研究范式也随之转变。古生物学数据甚至开始与现代生物学数据整合(如构建演化树, 探讨古生物与现代生物的系统关系)。通过计算机, 面向化石的研究产生了更多的数据, 多聚焦化石形态(详见第3节)及相关功能(包括如古脊椎动物如何运动与器官的功能等, 如Gai *et al*., 2022), 或特殊特征(用小波分析研究菊石的缝合线, 如Ubukata *et al*., 2014; 或通过植物叶片的气孔推测古大气CO2浓度, 如Wang *et al*., 2014等)。化石出现信息数据也衍生出了更多其他类型的数据。相关数据的复杂化和多元化开始要求新的统计方法与数学模型, 这一趋势在大算力时代愈演愈烈, 也同时成就了一些新兴方向。

2.3　大算力时代的古生物学建模分析与AI的介入

虽然目前已进入大数据时代, 古生物学相关数据也在21世纪初得到迅猛增长, 但促进古生物学研究的因素可能并不只是大数据本身, 更多的是与之相关的大算力及其带来的新分析方法。大算力依赖计算机硬件发展, 摩尔定律清晰揭示了这一过程。当下计算机硬件已能满足绝大多数分析, 个人高性能计算机与超算标志着大算力时代的到来。

大算力需求与古生物学数据特点相关。由于化石纪录不完备性, 古生物学数据对分析过程也有相应要求。以古生物多样性数据为例, 其不均衡性要求稀疏化(见黄冰, 2012), 对其估计需要外插值(如Huang *et al*., 2014), 这些均涉及重采样。动辄上千次或更多重采样只是基本要求, 虽然一般没有问题, 但对于较大数据(如显生宙多门类出现信息数据), 个人计算机就相形见绌了。此外, 传统的分支系统学研究使用PAUP软件, 树的生成有时需数天乃至数周, 新算法与软件的出现(如MrBayes或

TNT), 计算时间显著缩短, 但随着数据量与新模型复杂度的增加, 未来对算力的要求将进一步提升。近期一些新研究使用算法(如模拟退火算法)需要大量迭代计算, 导致收敛速度很慢, 甚至必须使用超算才能满足(如Fan *et al*., 2020)。

古生物学数据一般较为简单, 结构型与半结构型居多(详见第3节), 传统定量分析手段对已有数据的解读可能已近耗尽, 如PBDB相关研究论文增加速度开始有所下降。此外, 由于化石保存与发现的条件导致古生物学数据高度不确定性, 相关学者逐渐认识到一般统计学方法可能无法应对。上述情况促进了古生物学定量化研究进一步发展, 最直接的便是研究手段升级, 对数据建模, 结合统计模型而不是简单统计方法对小数据进行分析(或再分析), 以得出更为深入的结论。越来越多对古生物学定量研究关注的同行认为未来发展离不开模型(陈中强, 个人交流, 2023), 其中显然包括统计模型。

国际上统计模型的应用研究起步较早, 除了早期的常规模型外, 较近的如对奥陶纪末大灭绝后腕足动物的研究(Finnegan *et al*., 2016), 就同时采用了通用梯度回归模型(generalized boosted regression models, GBMs)与决策树模型(Classification tree model)。应用模型的趋势开始常态化, 甚至一些传统的分析方法也被基于模型的更复杂也更为准确的统计过程所开始取代。比如使用基于模型的聚类分析(与传统聚类在原理和实现上迥异)中的高斯混合模型(Gaussian mixture models)来研究埃迪卡拉动物群中的某个种的形态及分类(Evans *et al*., 2023), 或用多个线性模型来估计舌羊齿的叶脉密度(Esperança Júnior *et al*., 2023)等。更为具体的如MBD模型(multivariate birth-death model)来评估生物与非生物因素对多样性的影响(Guo *et al*., 2023)。古脊椎动物学与古昆虫学在性状数据与分支系统学研究应用最为成熟, 近年来也出现了不少新模型研究实例(如Zhang *et al*., 2023)。

近年来古生物数据产出向多元化和高维度发展, 特别是非结构型数据(如图像)。数据驱动型的古生物学研究成果仍然稳步增长, 研究内容与方法也越来越多样化。如形态学数据, 从最初的几何形态测量数据, 到界点(landmark)或半界点(semi-landmark)数据, 以及轮廓数据和三维界点数据直到细致的性状数据, 已有大量不胜枚举的研究。图像的人工智能(AI)识别与化石自动鉴定, 也已在某些门类取得了较好的效果(如Liu and Song, 2020等)。这类研究也促使国内外同行开始着手建立化石图片数据库(如http://www.fossilmuseum.net/等)。

此外, 大算力时代对交叉学科数据的整合也是必然趋势。古生物学如何在地学数据体系内寻找定位是未来新热点。硬件条件已满足, 重要的是如何建立知识图谱, 从而把数据用逻辑关系整合, 这也是近年开展的深时数字地球DDE大计划的理念之一(Wang *et al*., 2021)。事实上, 纳入地学数据体系也是古生物学者所期待的。而面向具体领域科学问题的科研人员, 最需要关注的是如何用数学模型探索古生物学问题, 这就涉及到古生物学数据的基本分类与研究方法。

3 古生物学数据与研究方法

3.1 大数据框架下的数据类型与研究方法

目前古生物学数据暂不符合大数据的基本特征, 但考虑到未来古生物学需要融入地学大数据框架, 我们仍可以按相关规则对古生物数据分类。大数据有多种分类标准, 最为通用的即涉及数据存储和处理时, 可以根据其组织和格式分为三种类型: 结构化数据、半结构化数据和非结构化数据。古生物学根据这该定义可大致代入, 分别对它们及相关统计分析方法简要介绍如下。

3.1.1 结构化数据与传统相关研究方法

结构化数据是按照预定义的方式进行组织的数据, 通常存储在关系型数据库中。结构化数据容易查询、分析和汇总。古生物学涉及到的出现信息(occurrence)、丰度(abundance)以及多样性(richness)等均为结构化数据, 它们跨门类可统一且关联性强, 相对而言也较容易分析。

广义上出现信息数据包括丰度信息(但经常未记录), 也可据此得出生物多样性数据。单纯的出

现信息数据多用来进行古生物地理分析, 或动物群的比较与分组等研究。这类需要寻找数据间关联规律的常规研究方法如聚类分类(CA),非度量性标度变换(NMDS)等。上述方法需注意相似性测度(similarity measure)的选择(参考戎嘉余等, 1995; 黄冰, 2011; Shi, 1993)。近来也有用网络分析(network analysis, NA)进行相关研究(如Sidor *et al*., 2013), 国内学者也逐渐开始应用该方法(如Huang *et al*., 2016; Fang *et al*., 2019; Xu *et al*., 2022等), 但需要重视分组的客观性依赖于社区检测(Community Detection)算法的选择。推荐使用R语言与相关程序包(如igraph)进行网络分析(见 王骞、黄冰, 2020), 社区检测算法包括Fast Greedy, Edge betweenness, Spinglass, Infomap等可供选择。除上述外, 其他一些多元统计方法也可以应用于相关数据, 通过降维来得出变量之间的关系, 如主成分分析(PCA)或应对分析(correspondence analysis, COA)等等。此外, 就古生物地理研究而言, 地理信息系统(GIS)与现代生物地理学研究方法的介入, 使结果的可视化及与环境信息的整合度得到很大提升(如Zhang *et al*., 2023), 因而也是未来重要方法之一。

丰度数据一直鲜受关注, 它对采样的方式与力度依赖性强是主要原因。事实上丰度数据在现代生物学中应用广泛, 比如物种丰度模型(species-abundance model)的研究。物种丰度模型与群落稳定性、资源配置以及演化过程有着密切关系, 它能反映群落中最基本的结构(Watkins and Wilson, 1994)。古生物丰度数据研究最有代表性的成果之一是对澄江生物群的相关研究(Zhao *et al*., 2014)。作者也曾用R语言与Vegan程序包(Oksanen *et al*., 2010)尝试对奥陶纪末大灭绝后的腕足动物群落进行相关研究, 虽有初步成果, 但分辨率相对较低(黄冰, 2015)。

古生物多样性研究有近半个世纪历史, 特别是长时间尺度多门类生物多样性规律成果颇丰(如Alroy *et al*., 2008等)。近年来, 显生宙多样性曲线绘制在超算的帮助下于分辨率上得到了极大提升(Fan *et al*., 2020; Deng *et al*., 2021), 未来对多样性更为细致准确的描述值得期待。古生物学多样性数据因研究程度, 保存条件等造成的不均衡, 过去通过稀疏化可部分解决, 但会造成数据的大量损失。借鉴现代生态学相关研究(Chao, 2005), 将多样性估计及外插值方法引入古生物研究中(Huang *et al*., 2014)逐渐开始被接受。此外, 多样性与环境信息的共变分析是一个重要方向, 比如在使用多样性估计方法的同时, 探讨纬度与多样性的共变模式(Song *et al*., 2020)。事实上, 多样性相关的重要宏演化事件一直都是研究热点, 用新的统计模型解决经典争议问题也有重要成果产出(如Guo *et al*., 2023)。多样性研究需要注意数据厘定, 虽然该工作非常耗费时间与人力(如: 宋海军, 个人交流, 2021), 但这一过程并不能忽视。

由于结构化数据具有跨门类可比较的特征, 因而也是探索宏演化规律最为重要的数据类型之一。相关研究已经较为成熟, 很多新方法也取得了重要的成果。采用统计模型的研究也在多方面得出新认识(见本文第4节)。与之相比, 更为复杂的半结构化数据大多被限制在不同门类之中, 相关研究近年来已受到重视并快速发展。

3.1.2　以形态学数据为代表的半结构化数据类型

半结构化数据在某种程度上有组织, 但不像结构化数据那样严格地遵循固定模式。这种数据的字段可能因数据的特定部分而异。大数据里典型的半结构化数据以“个人简历”为例, 该类数据在复杂程度与字段上差异很大。类似的, 在古生物学中, 由于不同门类属性迥异, 形态相关数据(如形态测量数据、界标点数据以及性状数据)跨门类难以整合, 因而都可以归入半结构化数据。由于古生物学分类研究建立在对形态与性状基础上, 这类数据是研究系统分类的关键, 可以说形态数据是古生物学数据中最为基础重要的类型之一。此外, 一些与古生物有关的跨学科数据(包括生物力学, 流体力学甚至地球化学等实验数据), 也可以作为关联字段在这一类型数据中体现。

就形态数据研究方法而言, 从几何形态测量到界标点分析的研究几乎在各生物门类中都有体现, 在早期研究中一些简单的软件均能实现相关研究(如用PAST做界标点的薄板样条分析, 见

Huang and Harper, 2013)。而R语言中也提供相关软件包，甚至有古生物学专用的(如Paleomorph包)。事实上，当用TpsDig2工具箱(Rohlf, 2006)将界标点数字化后，即可采用各种统计方法分析，如协方差分析、判别分析(Huang and Rong, 2011)或主成分分析(Zhang *et al*., 2021)，甚至傅里叶分析(史宇坤, 2017)等。当然，用R语言编程进行形态空间分析(如Guo *et al*., 2020; Wang and Huang, 2022等)是最为灵活的。在形态学分析方面，古脊椎动物与古人类学领域方法先进。比如，21世纪初古人类学研究中就使用三维界标点研究，甚至动态变化过程均可以展示，在当时甚至现在均令人惊叹。

性状信息是最完备的形态特征数据，古脊椎动物学研究积累极为丰富的性状数据，并在很早就开始关注分支系统学研究(见：周明镇等，1983)，近年发表的大量重要论文几乎都离不开相关方法。甚至开始结合其他类型数据进行更为广泛的分析，如分支系统古生物地理学研究等。古脊椎动物学对分支系统学研究的早期成果，可能也在一定程度上推动了古生物学其他门类(如古昆虫与古植物)开展相关研究。从古脊椎动物发育研究培训班的受益，也帮助了作者对腕足动物演化的研究(如Huang *et al*., 2023; Chen *et al*., 2023)。

形态空间演化的研究，在数据层面上可视为分支系统学研究的降阶版本，虽然两者研究目的不同。考虑到分支系统学的严谨与对数据某种程度的苛求(如要求全部性状的编码)，形态学空间研究就显得相对灵活很多。相关研究可以探索特定门类的特定形态在某段地史时期的演化规律(如Zhao *et al*., 2021)。作者认为，考虑到基因是自然选择的最小单位，而古生物学研究面对的则大多是基因表达的性状，通过形态空间研究则能探索环境变化对生物演化的直接作用。此外，考虑到不少分类单元系统分类意见的不统一，对形态空间的编码则能避开该矛盾，直接研究环境对生物影响的结果，即形态特征的改变。该类研究或许是未来演化探索的重要方向之一。

3.1.3 非结构化数据与化石的自动鉴定研究

结构化与半结构化数据是当前古生物学研究的主体，但大数据时代最丰富也是最符合时代特征是的非结构化数据。非结构化数据没有固定模式，很难对其内容进行检索，因而也更难以处理。这类数据主要包括图像、视频等。古生物学数据与相关成果，囿于发表载体，目前多集中图片，电子期刊也开始支持视频，但多限于在线的支持附件，而暂未内嵌入论文。

伴随计算机视觉及相关方法的发展，一些研究开始聚焦于化石自动鉴定，并对于依赖轮廓鉴定，且化石图像资源丰富的化石门类取得了较好结果，如对笔石标本图像建立智能识别模型(Xu *et al*., 2023)等。由于这一领域刚起步不久，所以对构造相对复杂的生物门类(如腕足动物，其描述构造的术语就超过 2000 个)，准确率仍亟待提高。

化石自动鉴定是一个诱人的研究方向，然而其难度却远超我们想象。目前相关方法已较成熟，比如卷积神经网络(Convolutional Neural Networks, CNN)是图像识别的核心方法，而迁移学习(Transfer Learning)，深度强化学习(Deep Reinforcement Learning)，以及深度神经网络(Deep Neural Networks, DNN)等手段为该研究在方法上几乎铺平了道路。化石自动鉴定当前最大的问题主要来源于有效的训练数据不足。

在作者早期的尝试中，化石自动鉴定有时展示出较高的准确率，但这可能是过拟合的假象。当训练数据量相对较少时，模型可能会在有限的数据上过于自信，而无法很好地泛化到更一般的情况。另外一个更难避免的问题是训练数据的不平衡，如果不同类别的样本数量严重不平衡(古生物学数据的常态)，模型可能会偏向于预测数量较多的类别。上述问题根本在于古生物学用于有效训练的数据不足。此外，相关研究在测试有效性的方式上也不宜采取传统的交叉验证(cross validation)，会对模型造成污染，只能将一部分原始数据用来验证，因而数据缺失的问题就更为明显。该领域未来的发展，取决于数据的积累程度。

古生物化石的三维重建也是近年来的新兴领域，与三维界标点不同，它一般利用断层扫描设备(CT或Micro-CT)获取高分辨率的三维数据。较近期的相关研究如鱼类牙齿的三维重建识别出了最古

老的有颌类牙齿(Andreev *et al*., 2022)等。古生物学视频数据目前多集中于各类科研汇报幻灯片中, 如全角度展示三维重建等。事实上, 很多古生物数据是天然的时间序列, 作者曾用区域变形技术(Field morphing)模拟了腕足动物个体发育过程, 由于某期刊限制, 只发表了视频关键帧(Huang and Harper, 2013)。随着期刊电子化的发展, 未来更多视频将会整合进论文中, 古生物化石的动态还原也会为古生物学数据增加新的内容。

3.2　定量古生物学 VS. 分析古生物学

有数据就会有对应的研究方法, 前文所述, 古生物学定量化研究早已开展。20世纪末, 国外有些学者就已考虑将定量分析方法加入古生物学专业研究生甚至本科生的教学中。很快, 哥本哈根大学就开设了定量古生物学(numerical palaeontology)课程, 所用的教材(Hammer and Harper, 2006)从基本统计方法开始教学, 介绍了多元数据分析、形态学分析、分支系统学分析以及古生物地理、古生态等分析方法, 甚至还涉及到了定量生物地层和时间序列分析(包括小波分析等), 多数的案例均可用PAST软件实现。定量古生物学课程的理念即教授基于数据的古生物学研究。

事实上, 远在该门课程之前, “分析古生物学”(analytical paleobiology)被作为美国古生物学会年度课程提出来(Signor and Gilinsky, 1991)。该方向介绍中提出分析古生物学的第一要义就是像其他自然科学研究方法一样, 对古生物学假说(paleontological hypotheses)进行敏感性和严格测试(sensitive and rigorous tests), 从而探讨古生物化石记录蕴含的模式。该文有一节都是在强调如何建立零假设(null hypothesis)模型, 并用统计学方法给出结论。同期短课程中还涉及宏演化的居群生物学模型(Sepkoski, 1991)以及随机模型(Signor, 1991)等介绍。

美国古生物学会自1978年起就开始古生物学年度的短课程, 同年美国国家博物馆也出版《古生物学中的随机模型入门》(Raup and Schopf, 1978)。分析古生物学短训班自提出后几乎隔几年就开一次, 最新版的学习手册是2019年由美国加州大学伯克利分校的一个项目所支持, 内容围绕古生物学数据进行基于模型的统计分析。该手册仅包括一些纲要和推荐读物以及示例代码, 所以并没有正式发表而仅有网页版(https://psmits.github.io/paleo_book/index.html)。

那么定量古生物学与分析古生物学究竟有什么区别呢? 需要强调, 两者都不能算是学科, 而只是从方法论角度讲的两种不同的研究思路。有学者认为后者更强调对结果的解释(Harper, 个人交流, 2023), 但作者认为, 从其产生根源来看, 分析古生物学更强调模型的特点更接近科学研究的本质, 即提出问题建立模型以及验证并解释模型。分析古生物学的学习手册第一章介绍基本数据管理之后, 就开始讨论贝叶斯数据分析、线性回归模型、逻辑回归模型(Logistic regression)以及其他的广义线性模型(generalized linear model, GLM), 最后以模型比较与时间序列数据分析结尾。

前文已经提到, 目前古生物的定量研究已经开始使用各种统计模型解决问题, 事实上, 国际同行很早就开始使用统计模型, 只是由于数据与算力不够, 早先统计模型选择相对有限, 所以才发展相对较为缓慢。简单来说, 传统统计方法更适合于简单的数据分析和基本统计推断, 而统计模型更适用于复杂的数据建模和更深入的数据分析。在大数据语境下, 古生物学数据的研究显然更依赖于后者。

4　大数据语境下的分析古生物学

4.1　来自*Palaeontology*与*Paleobiology*的小样本抽样

在介绍分析古生物学之前, 有必要了解当下古生物学(Paleontology)与化石生物学(Paleobiology)研究涉及到哪些主要研究方法。笔者选择代表性的古生物学专业期刊*Palaeontology*与*Paleobiology*, 对其2022—2023年发表的各30篇文献, 做了小样本抽样调查, 以求管窥现今相关研究的大致情况, 所得结果仅供参考。

在共计60篇文章中, 主要涉及数据的研究共计55篇(26∶29, 前者*Palaeontology*后者*Paleobiology*, 下同)。其中与形态性状相关的研究最多, 共26篇(17∶9), 这26篇中涉及分支系统学

研究共10篇(9∶1), 涉及形态空间或生态形态空间的共8篇(4∶4, 多数用到界标点方法), 另外就是关于体型、三维解剖重建或者用数学方法估计体重(Hart *et al*., 2022)等研究。

值得重视的是, 60篇论文中建立统计模型或数学模型的共计15篇(4∶11), 其中包括积分投影模型(Brombacher *et al*., 2023), 高斯混合模型(Evens *et al*., 2023; Pauly and Holmes, 2023)等现代生物学与生态学常用方法, 此外还有在层次贝叶斯框架(Hierarchical Bayesian framework)下自建数学模型(Reitan *et al*., 2023), 或基于流体力学实验的数学模型(Darroch, 2022), 甚至有应用热力学定律研究耗散结构模型与演化的关系(de Castro and McShea, 2022)的研究等。此外, 生物多样性研究仍有7篇(3∶4), 而新技术相关研究也有不少, 包括古脊椎动物牙齿的模拟与仿真(Yang *et al*., 2022), 用机器学习研究生态选择性(Foster, 2022), 以及用神经网络方法自动鉴定(Liu *et al*., 2023)等等。

明显发现, 如此小样本就能反映出当下古生物研究类型之丰富, 使用手段之多元。两个期刊的对比表明, *Palaeontology*涉及大量形态学相关研究, 强调分支系统树的建立, 而*Paleobiology*更多借鉴现代生物学的方法, 更强调数学模型或统计模型。分析古生物学的肇始与后者创刊关系密切也就不难理解了。

需要强调上述调查样本过小, 同时值得注意*Palaeontology*期刊传统上多以化石系统分类与描述为主, 近年才开始不接受仅基于标本描述的文稿而倾向于有数据分析的研究。此外, 一些发表在高影响力期刊更为多元, 采用更为复杂的方法与模型的研究更受到关注, 比如生态系统与食物网模型(Huang Yuan-geng *et al*., 2021), 更先进的多样性演化模型(Guo *et al*., 2023)等等。这里从古生物学代表性专业期刊入手, 旨在初步了解未来古生物学的可能方向, 供向面向数据研究的古生物学者, 特别是青年古生物工作者参考。

4.2 分析古生物学与统计模型

分析古生物学强调的第一个概念就是假设检验，而显著性检验, 就是一种特定的假设检验。通过显著性检验可以评估某个变量对研究结果的影响程度。在科研论文中, 涉及到“显著地(significantly)”一词都需要提供显著性检验。早期的*t*检验等参数检验方法已经让位于非参数检验, 如 Mann-Whitney 检验和 K-S (Kolmogorov-Smirnov)检验等(见Huang *et al*., 2010)。显著性检验是为了支持结果的可信度, 而科学研究往往以一个或多个假设为基础, 科学家通过假设检验来验证它们是否被数据支持, 古生物学研究也不例外。

假设检验与统计模型关系密切, 因为假设检验常常在统计模型的框架下进行。统计模型和传统统计方法之间的主要区别在于复杂性和应用范围。传统统计与定量分析方法通常涉及基本的统计概念和技术, 如常见的CA、PCA和NMDS等, 用于分析数据的基本特征和关系。而统计模型是一种用于描述数据生成过程的数学框架, 涉及多个变量与不同的概率分布, 可以更深入理解数据的生成机制。统计模型可以用来预测、精准分类等更复杂的数据分析任务。

大数据语境下的古生物研究已经不能被一般统计方法所满足, 它要求统计模型的参与。古生物学涉及最重要的问题就是“分类”, 大多数的古生物学问题都可以纳入分类的范畴。从物种的系统分类, 形态空间的分类, 到宏演化模式的识别等。而本质上回归和分类的问题是相同的, 因变量是数量变量时, 则为回归建模, 而因变量为分类变量(定性变量)时, 则为分类建模(吴喜之、张敏, 2020)。以回归来说, 包括广义线性模型、逻辑回归模型以及决策树模型和人工神经网络回归等模型; 就分类而言, 也有对应的分类模型(如决策树分类、随机森林分类等)。这些模型在现代生物学中广泛使用, 也已经有在古生物学应用的案例(如Finnegan *et al*., 2016等)。

那么统计模型较传统统计方法的优势主要在哪里呢? 以在60篇论文小抽样就有2篇使用的高斯混合模型(Gaussian Mixture Models, GMM)为例, 用GMM进行聚类在如下三方面均优于传统聚类分析: (1) GMM是一种基于概率分布的聚类方法, 而传统聚类无法提供数据点属于不同聚类的概率信

息; (2) GMM通常采用信息准则(如贝叶斯信息准则Bayesian Information Criterion, BIC, 或赤池信息准则Akaike information criterion, AIC)或交叉验证等方法来自动确定簇的数量, 而传统聚类需要加入主观判断; (3) GMM能够对数据不确定性良好建模, 而传统聚类对噪声较为敏感, 难以明确区分噪声点和有效簇。可以看出, GMM聚类在处理复杂数据、不确定性方面具有优势, 显然古生物学的很多数据在不确定性特点上更适合用GMM聚类。

此外, 近年来分析古生物学对贝叶斯统计学开始越发重视。统计学公认分为两大学派, 即频率统计学(又被称为古典学派)与贝叶斯统计学, 两者均有其优势和应用领域。频率统计学在传统统计分析中应用广泛, 而贝叶斯统计能够整合先验知识, 随着模型复杂性的增大, 贝叶斯模型的优势趋于明显, 其可以将很多“不确定性”因素整合进模型, 因而尤其适合古生物学数据的特点。未来贝叶斯统计学在古生物学中将发挥重要作用, 国内已有学者开始使用相关方法取得重要成果(Guo *et al*., 2023)。

综上, 大数据语境下的分析古生物学在统计模型的帮助下较传统统计方法能够得出更为客观的结论, 而古生物学数据日益多元化也能促使新的数学模型被建立使用。只有通过观念与方法上的革新, 古生物学在数据产出速度有限的情况下才能向纵深发展。

5　数据驱动与模型驱动的选择

5. 1　数据驱动与自下而上(bottom-up)的思路

伴随着大数据的出现, “数据驱动”一词也开始深入各个领域。数据驱动是一种方法论, 它强调通过收集、分析和利用数据来做出决策、优化过程和实现研究目标。一般而言, 在学科定位上, 数据驱动往往与数据密集型科学紧密相联。一些学科很早就是数据密集型科学, 比如从基础学科的天文学、物理学, 到应用学科的气象学, 环境科学甚至与临床医学等等。

地球科学显然也是数据密集型学科, 并且已有数据驱动型研究的介绍(Wang *et al*., 2021), 那么古生物学科呢? 就其发展历史看, 古生物学家自己都认为古生物学与物理学等学科有较大的差异, 属于不那么“坚固”的学科(Gould, 2011)。虽然古生物学可以涉及一些实验, 但它并不是通俗意义上实验为主的科学。古生物学进行基于数据的研究不过几十年, 它目前仍不是数据密集型科学。

数据驱动型的古生物研究似乎与自下而上(bottom-up)的研究思路相吻合。即基于数据的归纳总结来构建一般性原则。它侧重于从底层开始, 收集数据, 根据数据及其分析逐步建立对整体的理解。典型的数据驱动型研究强调大量的数据集本身能够提供新知识来源, 而不需要为科学现象建模, 这是古生物学目前很难达到的。但当前古生物学研究的确展示出数据驱动的特点, 如有目的地获得数据, 并基于数据进行分析, 甚至根据数据“定制”相关研究。这是古生物学研究范式转换过程中出现的现象。

未来向数据驱动型研究发展的古生物学需要解决的首要问题在于数据本身, 也即大数据中的速度(Velocity)与大量(Volume)两个维度。向数据密集型科学借鉴经验(如增加更多的实验数据), 或与其他学科数据进行整合以保持一定规模的数据流是可行的解决方案。

5. 2　模型驱动与自上而下(top-down)的设计

模型驱动方法强调使用数学模型或计算模型来表示系统、现象或问题, 这些模型可以基于理论或经验构建。模型驱动方法侧重于理论推导和建模, 以便理解系统内部机制, 并使用模型来指导问题解决。如前文所言, 越来越多的古生物学研究开始基于统计模型或更一般的数学模型进行研究。这些研究大都是面向特定的科学问题构建一个理论模型, 进而根据模型来去寻找数据, 建立模型, 进而分析评估模型, 最终得出问题的最优解。

从某种意义上讲, 模型驱动的研究, 更像是一种自上而下(top-down)的设计, 即首先确定科学问题, 从高层次、整体的视角开始, 将问题分解为更具体的组成部分, 以理解整体系统或现象的机制。比如研究者可能制定一个演化模型, 然后分析各个因素如环境、物种、食物网等对整体生态系统的影响。

当前古生物学数据的规模仍相对较小, 相关

研究更多聚焦于一个相对明确的问题(而非多个有关联的问题), 涉及到的参数有限, 所以并不涉及数据挖掘与数据内隐规律的启发式发现。此外, 模型驱动更接近科学研究的一般理念, 即从科学问题着手, 选择科学模型, 进而收集数据并分析验证。可以认为近年来的涉及分析古生物学的研究范式更像是一种基于模型的计算科学(见下文)。

5.3 数据与模型相结合的古生物学研究范式

《第四范式: 数据密集型科学发现》一书(Hey, 2009)中提出了科学研究范式发展的四个阶段, 即实验科学、理论科学、计算科学和数据密集型科学。作者认为古生物学当前仍处于第二与第三范式。只有在第四范式中, 即数据密集型科学中数据驱动的研究才占据主导地位。数据驱动和自下而上(Bottom-Up)与模型驱动和自上而下(Top-Down)是两种研究思路或方法, 它们在问题解决和研究中有不同的侧重点, 但也可以互相关联和补充。

模型驱动强调使用数学模型或计算模型来表示古生物学相关现象或演化问题, 数据驱动强调使用实际数据来指导相关问题的解决。很多研究可以将二者结合在一起, 使用模型来指导研究, 并利用数据来交叉验证模型的结果。一般而言, 偏向模型驱动还是数据驱动方法通常取决于问题的性质和需求。古生物学问题如果有坚实的理论知识和模型可以准确描述, 那么模型驱动方法可能更合适。有些时候如果问题复杂或数据丰富, 数据驱动方法可能更有优势。思路上, 自上而下可以提供整体理论框架, 而自下而上则可以用来验证该理论并提供具体的案例支持。

虽然目前很多学科都在从模型驱动向数据驱动发展, 甚至统计学自身也是有同样的趋势, 强调数据而弱化模型的人为性(吴喜之、张敏, 2020)。但古生物学由于目前所处阶段的特点, 将数据驱动与模型驱动相结合的研究方式, 可能更有助于理解和解决古生物学相关问题。

6 展 望

近年统计学界有一个大事件, *Nature*期刊发表了以3名统计学家为首, 800多名学者签名的评论(Amrhein *et al*., 2019), 文章标题是《科学家们起来反对统计显著性》。新研究表明, 数据科学模型应该满足三个原则, 即可预测性、可计算性与稳定性(Yu and Kumbier, 2020), 不再提显著性。“显著性”甚至是很多统计学家的信仰, 而且很多研究离开统计显著性检验几乎不可想象。但科学家们发现, 问题并不完全在显著性本身, 更多是人为因素导致。很多显著性很强的结论并不能解释真实世界, 有时候古生物学数据呈现出的显著性结论, 可能就是类似的陷阱。统计分析意义在于解决问题, 而不是给主观的设定画一幅好的装饰画, 好看的画也许并不能支持结论, 需要理性对待数据及分析结果。

地球是一个复杂非线性多重耦合系统(谢树成等, 2006), 对复杂系统的行为进行解释远比我们想象的要困难得多。单纯的因果关系几乎不存在, 取而代之的是复杂的反馈网络。古生物学研究也开始使用统计学模型或数学模型对更多元的数据进行分析。近年人工智能的发展开始助力各学科研究, 比如伴随而来的概率图模型(Probabilistic Graphical Models, PGM)值得关注, 它基于图论与概率论的强大数学理论基础, 广泛应用于不确定性推理任务(Sucar, 2021)。类似的模型方法已日趋成熟, 开始在很多学科崭露头角, 未来也有可能应用于古生物学研究中。

经由模型与数据两方面的驱动, 未来在大数据语境下的古生物学研究将会更为多元化, 科学问题也将在深度和广度上得到发展。国内古生物专业的研究生与青年学者对编程、统计方法与新模型的掌握速度, 以及不囿于思维定式与传统研究范式的新思路, 将使未来我国古生物学数据相关研究值得期待。此外, 需要强调门类古生物研究在未来还将持续发挥重要作用, 它是理解生命与地球环境协同演化的第一手资料的重要来源。

致谢 感谢中国科学院南京地质古生物研究所王博研究员与评审徐洪河研究员对文稿提出宝贵意见。

参考文献 (References)

方宗杰, 杨群, 2009. 总论. 见: 古生物学名词审定委员会(编), 古生物学名词(第二版). 北京: 科学出版社. 1–71.

黄冰, 2007. 系统古生物研究中的统计新方法初探——以浙赣边区志留纪初期雕正形贝属为例. 古生物学报, 46: 278–292. DOI: 10.3969/j.issn.0001-6616.2007.03.002

黄冰, 2011. 简论相似性测度的选择——以奥陶纪末大灭绝后全球腕足动物古地理为例. 古生物学报, 50: 304–320. DOI: 10.19800/j.cnki.aps.2011.03.003

黄冰, 2012. 浅谈稀疏标准化方法(Rarefaction)及其在群落多样性研究中的应用. 古生物学报, 51: 200–208. DOI: 10.19800/j.cnki.aps.2012.02.005

黄冰, Harper D A T, Hammer Ø, 2013. 定量古生物学软件PAST及其常用功能. 古生物学报, 52: 161–181.

黄冰, 2015. 志留纪华夏正形贝动物群丰度模型研究及其意义——兼浅介R语言的古生态学应用. 古生物学报, 54: 472–480. DOI: 10.19800/j.cnki.aps.2015.04.006

戎嘉余, 李荣玉, 尼·库尔科夫, 1995. 亚洲志留纪Llandovery世腕足类生物地理分析——兼对亲缘关系指数公式的推荐. 古生物学报, 34: 428–453. DOI: 10.19800/j.cnki.aps.1995.04.003

史宇坤, 2017. 形态测量学(Morphometrics)常用方法及其在微体古生物学中的应用. 微体古生物学报, 34: 179–191. DOI: 10.16087/j.cnki.1000-0674.2017.02.006

王骞, 黄冰, 2020. 浅谈网络分析法及其在古生物学中的应用. 古生物学报, 59: 380–392. DOI: 10.19800/j.cnki.aps.2020.014

吴喜之, 张敏, 2020. 应用回归及分类——基于R与Python的实现. 2版. 北京: 中国人民大学出版社. 1–332.

谢树成, 龚一鸣, 童金南, 史晓颖, 赖旭龙, 陈中强, 冯庆来, 王红梅, 杜远生, 王永标, 颜佳新, 张克信, 殷鸿福, 2006. 从古生物学到地球生物学的跨越. 科学通报, 51: 2327–2336.

周明镇, 张弥曼, 于小波 等译, 1983. 分支系统学译文集. 北京: 科学出版社. 1–209.

Alroy J, Aberhan M, Bottjer D J, Foote M, Fürsich F T, Harries P J, Hendy A J W, Holland S M, Ivany L C, Kiessling W, Kosnik M A, Marshall C R, McGowan A J, Miller A I, Olszewski T D, Patzkowsky M E, Peters S E, Villier L, Wagner P J, Bonuso N, Borkow P S, Brenneis B, Clapham M E, Fall L M, Ferguson C A, Hanson V L, Krug A Z, Layou K M, Leckey E H, Nürnberg S, Powers C M, Sessa J A, Simpson C, Tomasovych A, Visaggi C C, 2008. Phanerozoic trends in the global diversity of marine invertebrates. Science, 321: 97–100. DOI: 10.1126/science. 1156963

Amrhein V, Greenland S, McShane B, 2019. Scientists rise up against statistical significance. Nature, 567: 305–307. DOI: 10.1038/d41586-019-00857-9

Bancroft B B, 1945. The brachiopod zonal indices of the Stages Costonian to Onnian in Britain. Journal of Paleontology, 19: 181–252.

Boucot A J, 1975. Evolution and extinction rate controls. New York: Elsevier Scientific Publishing Company. 1–427.

Brombacher A, Schmidt D N, Ezard T H G, 2023. Developmental plasticity in deep time: a window to population ecological inference. Paleobiology, 49: 259–270. DOI: 10.1017/pab.2022.26

de Castro C, McShea D W, 2022. Applying the Prigogine view of dissipative systems to the major transitions in evolution. Paleobiology, 48: 711–728. DOI: 10.1017/pab.2022.7

Chao A, 2005. Species richness estimation. *In*: Balakrishnan N, Read C B, Vidakovic B (eds.), New York: Encyclopedia of Statistical Sciences. 7909–7916.

Chen Di, Huang Bing, Candela Y, 2023. Evolutionary trends in trimerellid brachiopods. Palaeogeography, Palaeoclimatology, Palaeoecology, 617: 111472. DOI: 10.1016/j.palaeo.2023.111472

Darroch S A F, Gibson B M, Syversen M, Rahman I A, Racicot R A, Dunn F S, Gutarra S, Schindler E, Wehrmann A, Laflamme M, 2022. The life and times of Pteridinium simplex. Paleobiology, 48: 527–556. DOI: 10.1017/pab.2022.2

Deng Yi-ying, Fan Jun-xuan, Zhang Shu-han, Fang Xiang, Chen Zhong-yang, Shi Yu-kun, Wang Hai-wen, Wang Xin-bing, Yang Jiao, Hou Xu-dong, Wang Yue, Zhang Yuan-dong, Chen Qing, Yang Ai-hua, Fan Ru, Dong Shao-chun, Xu Hui-qing, Shen Shu-zhong, 2021. Timing and patterns of the Great Ordovician Biodiversification Event and Late Ordovician mass extinction: Perspectives from South China. Earth-Science Reviews, 220: 103743. DOI: 10.1016/j.earscirev.2021.103743

Doctorow C, 2008. Big data: welcome to the petacentre. Nature, 455: 16–21. DOI: 10.1038/455016a

Dodd J R, Stanton R J J, 1990. Paleoecology, Concepts and Applications (2nd Edition). New York: John Wiley & Sons, Inc. 1–528.

Duméril C, Kellogg R, Zoologie analytique, ou, Méthode naturelle de classification des animaux: rendue plus facile a l'aide de tableaux synoptiques / par A.M. Constant Duméril. Paris: Allais, libraire, 1806. DOI: 10.5962/bhl.title.44835

Esperan ç a, Júnior M G F, Cybis G B, Iannuzzi R, 2023. An efficient method for estimating vein density of Glossopteris and its application. Palaeontology, 66: e12640. DOI: 10.1111/pala.12640

Evans S D, Hunt G, Gehling J G, Sperling E A, Droser M L, 2023. Species of Dickinsonia Sprigg from the Ediacaran of South Australia. Palaeontology, 66: e12635. DOI: 10.1111/pala.12635

Fan Jun-xuan, Shen Shu-zhong, Erwin D H, Sadler P M, MacLeod N, Cheng Qiu-ming, Hou Xu-dong, Yang Jiao, Wang Xiang-dong, Wang Yue, Zhang Hua, Chen Xu, Li Guo-xiang, Zhang Yi-chun, Shi Yu-kun, Yuan Dong-xun, Chen Qing, Zhang Lin-na, Li Chao, Zhao Ying-ying, 2020. A high-resolution summary of Cambrian to Early Triassic marine invertebrate biodiversity. Science, 367: 272–277. DOI:10.1126/science.aax4953

Fang Xiang, Burrett C, Li Wen-jie, Zhang Yun-bai, Zhang Yuan-dong, Chen Ting-en, Wu Xue-jin, 2019. Dynamic variation of Middle to Late Ordovician cephalopod provincialism in the northeastern peri-Gondwana region and its implications. Palaeogeography, Palaeoclimatology, Palaeoecology, 521: 127–137. DOI: 10.1016/j.palaeo.2019.02.015

Fang Zong-jie, Yang Qun, 2009. Overview. *In:* Review Committee of Chinese terms in Palaeontology (ed.), Chinese terms in Palaeontology. Beijing: Science Press. 1–71 (in Chinese).

Finnegan S, Rasmussen C M Ø, Harper D A T, 2016. Biogeographic and bathymetric determinants of brachiopod extinction and survival during the Late Ordovician mass extinction. Proceedings of the Royal Society B-Biological Sciences. 283: 20160007.

DOI: 10.1098/rspb.2016.0007

Foster W J, Ayzel G, Münchmeyer J, Rettelbach T, Kitzmann N H, Isson T T, Mutti M, Aberhan M, 2022. Machine learning identifies ecological selectivity patterns across the end-Permian mass extinction. Paleobiology, 48: 357–371. DOI: 10.1017/pab.2022.1

Furnish W M, Unklesbay A G, 1940. Diagrammatic representation of ammonoid sutures. Journal of Paleontology, 14: 598–602.

Gai Zhi-kun, Li Qiang, Ferrón H G, Keating J N, Wang Jun-qing, Donoghue P C J, Zhu Min, 2022. Galeaspid anatomy and the origin of vertebrate paired appendages. Nature, 609: 959–963. DOI: 10.1038/s41586-022-04897-6

Gould S J, Katz M, 1975. Disruption of ideal geometry in the growth of receptaculitids: a natural experiment in theoretical morphology. Paleobiology, 1: 1–20. DOI: 10.1017/S0094837300002153

Gould S J, 2011. The Hedgehog, the Fox, and the Magister's Pox: Mending the Gap between Science and the Humanities. Boston: Harvard University Press. 1–288.

Grant R E, 1972. The Lophophore and Feeding Mechanism of the Productidina (Brachiopoda) (Permian). Journal of Paleontology, 46: 213–248.

Greene F C, 1908. The Development of a Carboniferous Brachiopod, Chonetes granulifer Owen. Journal of Geology, 16: 654–663.

Guo Zhen, Chen Zhong-qiang, Harper D A T, 2020. Phylogenetic and ecomorphologic diversifications of spiriferinid brachiopods after the end-Permian extinction. Paleobiology, 46: 495–510. DOI: 10.1017/pab.2020.34

Guo Zhen, Flannery-Sutherland J T, Benton M J, Chen Zhong-qiang, 2023. Bayesian analyses indicate bivalves did not drive the downfall of brachiopods following the Permian-Triassic mass extinction. Nature Communications, 14: 5566. DOI: 10.1038/s41467-023-41358-8

Hammer Ø, Harper D A T, Ryan P D, 2001. PAST: Palaeontological Statistics software package for education and data analysis. Palaeontologia Electronica, 4: 9.

Hammer Ø, Harper D A T, 2006. Paleontological Data Analysis. Oxford: Blackwell Publishing. 1–351.

Hart L J, Campione N E, McCurry M R, 2022. On the estimation of body mass in temnospondyls: a case study using the large-bodied *Eryops* and *Paracyclotosaurus*. Palaeontology, 65: 12629. DOI: 10.1111/pala.12629

Hey T, 2009. The Fourth Paradigm: Data-Intensive Scientific Discovery. Redmond: Microsoft Research Press. 1–284.

Huang Bing, 2007. Primary exploration of new statistical methods in systematic palaeontology—Example of early Silurian *Glyptorthis* from the Zhejiang–Jiangxi border area. Acta Palaeontologica Sinica, 46: 278–292 (in Chinese with English abstract). DOI: 10.3969/j.issn.0001-6616.2007.03.002

Huang Bing, Harper D A T, Zhan Ren-bin, Rong Jia-yu. 2010. Can the Lilliput Effect be detected in the brachiopod faunas of South China following the terminal Ordovician mass extinction? Palaeogeography, Palaeoclimatology, Palaeoecology, 285: 277–286. DOI: 10.1016/j.palaeo.2009.11.020

Huang Bing, 2011. Preliminary discussion on similarity measures with an example of Rhuddanian global brachiopod palaeobiogeography. Acta Palaeontologica Sinica, 50: 304–320 (in Chinese with English abstract). DOI: 10.19800/j. cnki.aps.2011.03.003

Huang Bing, Rong Jia-yu, 2011. Statistically differentiating *Katastrophomena* from *Strophomena* (Ordovician-Silurian strophomenid brachiopods). Memoirs of the Association of Australasian Palaeontologists, 39: 245–259.

Huang Bing, 2012. Rarefaction and its application to the study of diversity of palaeocommunities. Acta Palaeontologica Sinica, 51: 200–208 (in Chinese with English abstract). DOI: 10.19800/j.cnki.aps.2012.02.005

Huang Bing, Harper D A T, 2013. Ontogenic study of the brachiopod *Dicoelosia* by geometric morphometrics and morphing techniques. Lethaia, 46: 308–316. DOI: 10.1111/let.12009.

Huang Bing, Harper D A T, Hammer Ø, 2013. Introduction to PAST, a comprehensive statistics software package for paleontological data analysis. Acta Palaeontologica Sinica, 52: 161–181 (in Chinese with English abstract). DOI: 10.19800/j.cnki.aps.2013.02.003

Huang Bing, Harper D A T, Zhan Ren-bin, 2014. Test of sampling sufficiency in palaeontology. GFF, 136: 105–109. DOI: 10.1080/11035897.2014.882976

Huang Bing, 2015. Species-abundance models for the *Cathaysiorthis* fauna (Silurian brachiopods) and its significance, with an application of R in palaeoecology. Acta Palaeontologica Sinica, 54: 472–480 (in Chinese with English abstract). DOI: 10.19800/j.cnki.aps.2015.04.006

Huang Bing, Zhan Ren-bin, Wang Guang-xu, 2016. Recovery brachiopod associations from the lower Silurian of South China and their paleoecological implications. Canadian Journal of Earth Sciences, 53: 674–679. DOI: 10.1139/cjes-2015-0193

Huang Bing, Chen Di, Harper D A T, Rong Jia-yu, 2023. Did the Late Ordovician mass extinction event trigger the earliest evolution of 'strophodontoid' brachiopods? Palaeontology, 66: e12642. DOI: 10.1111/pala.12642

Huang Yuan-geng, Chen Zhong-qiang, Roopnarine P D, Benton M J, Yang Wan, Liu Jun, Zhao Lai-shi, Li Zhen-hua, Guo Zhen, 2021. Ecological dynamics of terrestrial and freshwater ecosystems across three mid-Phanerozoic mass extinctions from Northwest China. The Royal Society Proceedings of the Royal Society B, 288: 20210148. DOI: 10.1098/rspb.2021.0148

Hunt G, 2006. Fitting and comparing models of phyletic evolution: random walks and beyond. Paleobiology, 32: 578–601. DOI: 10.1666/05070.1

Jones B, Smith G P, 1985. Paleoecology of the brachiopod faunas in the Lower Devonian Eids Formation of Southwest Ellesmere Island, Arctic Canada. Journal of Paleontology, 59: 957–974.

Kocsis Á T, Reddin C J, Alroy J, Kiessling W, 2019. The R package divDyn for quantifying diversity dynamics using fossil sampling data. Methods in Ecology and Evolution, 10: 735–743. DOI: 10.1111/2041-210x.13161

Lenz A C, 1967. *Thliborhynchia*, a new Lower Devonian rhynchonellid from Royal Creek, Yukon, Canada. Journal of Paleontology, 41: 1188–1192.

Liu Xiao-kang, Song Hai-jun, 2020. Automatic identification of

fossils and abiotic grains during carbonate microfacies analysis using deep convolutional neural networks. Sedimentary Geology, 410: 105790. DOI: 10.1016/j.sedgeo.2020.105790

Liu Xiao-kang, Jiang Shou-yi, Wu Rui, Shu Wen-chao, Hou Jie, Sun Yong-fang, Sun Jia-rui, Chu Dao-liang, Wu Yu-yang, Song Hai-jun, 2023. Automatic taxonomic identification based on the Fossil Image Dataset (>415,000 images) and deep convolutional neural networks. Paleobiology, 49: 1–22. DOI: 10.1017/pab.2022.14

Long C A, 1985. Intricate sutures as fractal curves. Journal of Morphology, 185: 285–295. DOI: 10.1002/jmor.1051850303

Mayer-Schönberger V, Cukier K, 2014. Big data: a revolution that will transform how we live, work, and think. Boston: Harper Business. 1–272.

Moseley H, 1838. On the geometrical forms of turbinated and discoid shells. Philosophical Transactions of the Royal Society of London, 1838: 351–370.

Newell N D, 1949. Phyletic size increase, an important trend illustrated by fossil invertebrates. Evolution; International Journal of Organic Evolution, 3: 103–124. DOI: 10.1111/j.1558-5646.1949.tb00010.x

Novack-Gottshall P M, 2007. Using a theoretical ecospace to quantify the ecological diversity of Paleozoic and modern marine biotas. Paleobiology, 33: 273–294. DOI: 10.1666/06054.1

Oksanen J, Kindt R, Legendre P, O'Hara R B, Stevens M H H, 2010. Vegan: Community Ecology Package. R package version 2. http://r-forge.rproject.org/projects/vegan

Pauly D, Holmes J D, 2023. Reassessing growth and mortality estimates for the Ordovician trilobite *Triarthrus eatoni*. Paleobiology, 49: 120–130. DOI: 10.1017/pab.2022.22

Phillips J, 1860. Life on the Earth: its origin and succession. Cambridge: Macmillan. 1–270. DOI: 10.5962/bhl.title.22153

Raup D M, 1967. Geometric analysis of shell coiling: coiling in ammonoids. Journal of Paleontology, 41: 43–65.

Raup D M, 1975. Taxonomic survivorship curves and Van Valen's Law. Paleobiology, 1: 82–96. DOI: 10.1017/s0094837300002220

Raup D M, Schopf T J M, 1978. Stochastic Model in Paleontology: A primer. National Museum, 1–130.

Reitan T, Ergon T H, Liow L H, 2023. Relative species abundance and population densities of the past: developing multispecies occupancy models for fossil data. Paleobiology, 49: 23–38. DOI: 10.1017/pab.2022.17

Rohlf F J, 2006. TpsDig, program for digitizing landmarks and outlines for geometric morphometric analyses, version 2.32. New York: Department of Ecology and Evolution, State University of New York at Stony Brook.

Rong Jia-yu, Li Rong-yu, Kulko N P, 1995. Biogeographic analysis of Llandovery brachiopods from Asia with a recommendation of use of affinity indices. Acta Palaeontologica Sinica, 34: 428–453 (in Chinese with English summary). DOI: 10.19800/j.cnki.aps.1995.04.003

Schopf T J M, Raup D M, Gould S J, Simberloff D S, 1975. Genomic versus morphologic rates of evolution: influence of morphologic complexity. Paleobiology, 1: 63–70. DOI: 10.1017/S0094837300002207

Schopf T J M, 1974. Theory in Paleoecology-Models in Ecology. Cambridge: Cambridge University Press. 1–146.

Sepkoski J J Jr, 1984. A kinetic model of Phanerozoic taxonomic diversity. III. Post-Paleozoic families and mass extinctions. Paleobiology, 10: 246–267. DOI: 10.1017/s0094837300008186

Sepkoski J J Jr, 1991. Population biology models in macroevolution. Short Courses in Paleontology, 4: 136–156. DOI: 10.1017/s2475263000002166

Sepkoski J J Jr, Bambach R K, Raup D M, Valentine J W, 1981. Phanerozoic marine diversity and the fossil record. Nature, 293: 435–437. DOI: 10.1038/293435a0

Shi G R, 1993. Multivariate data analysis in palaeoecology and palaeobiogeography—a review. Palaeogeography, Palaeoclimatology, Palaeoecology, 105: 199–234. DOI: 10.1016/0031-0182(93)90084-v

Shi Yu-kun, 2017. Introduction of morphometrics and a case study on fusulinid foraminifera. Acta Micropalaeontologica Sinica, 34: 179–191. DOI: 10.16087/j.cnki.1000-0674.2017.02.006 (in Chinese with English abstract)

Sidor C A, Vilhena D A, Angielczyk K D, Huttenlocker A K, Nesbitt S J, Peecook B R, Steyer J S, Smith R M H, Tsuji L A, 2013. Provincialization of terrestrial faunas following the end-Permian mass extinction. Proceedings of the National Academy of Sciences of the United States of America, 110: 8129–8133. DOI: 10.1073/pnas.1302323110

Signor P W, 1991. Random models in paleobiology. Short Courses in Paleontology, 4: 123–135. DOI: 10.1017/s2475263000002154

Signor P W, Gilinsky N L, 1991. Introduction to analytical paleobiology. Short Courses in Paleontology, 4: 1–3. DOI: 10.1017/s2475263000002087

Song Hai-jun, Huang Shan, Jia En-hao, Dai Xu, Wignall P B, Dunhill A M, 2020. Flat latitudinal diversity gradient caused by the Permian–Triassic mass extinction. Proceedings of the National Academy of Sciences of the United States of America, 117: 17578–17583. DOI: 10.1073/pnas.1918953117

Stanley S M, 1975. Why clams have the shape they have: an experimental analysis of burrowing. Paleobiology, 1: 48–58. DOI: 10.1017/S0094837300002189

Sucar L E, 2021. Probabilistic Graphical Models: Principles and Applications. Berlin: Springer Press. 1–384.

Temple J T, 1992. The Progress of Quantitative Methods in Paleontology. Palaeontology, 35: 475–484.

Tipper J C, 1991. Computer applications in paleontology: balance in the late 1980s? Computers & Geosciences, 17: 1091–1098. DOI: 10.1016/0098-3004(91)90070-t

Ubukata T, Tanabe K, Shigeta Y, Maeda H, Mapes R H, 2014. Wavelet analysis of ammonoid sutures. Palaeontologia Electronica, 17: 9A. DOI: 10.26879/381

Wang Cheng-shan, Hazen R M, Cheng Qiu-ming, Stephenson M H, Zhou Cheng-hu, Fox P, Shen Shu-zhong, Oberhänsli R, Hou Zeng-qian, Ma Xiao-gang, Feng Zhi-qiang, Fan Jun-xuan, Ma Chao, Hu Xiu-mian, Luo Bin, Wang Juan-le, Schiffries C M, 2021. The Deep-Time Digital Earth program: data-driven discovery in geosciences. National Science Review, 8: nwab027. DOI: 10.1093/nsr/nwab027

Wang Qian, Huang Bing, 2020. Network analysis and its application

in paleontology—a preliminary introduction. Acta Palaeontologica Sinica, 59: 380–392 (in Chinese with English abstract). DOI: 10.19800/j.cnki.aps.2020.014

Wang Qian, Huang Bing, 2022. An ontogenetic study of *Eospirigerina putilla* (Brachiopoda) surviving the Late Ordovician mass extinction. Palaeoworld, DOI: 10.1016/j.palwor.2022.06.001

Wang Yong-dong, Huang Cheng-min, Sun Bai-nian, Quan Cheng, Wu Jing-yu, Lin Zhi-cheng, 2014. Paleo-CO_2 variation trends and the Cretaceous greenhouse climate. Earth-Science Reviews, 129: 136–147. DOI: 10.1016/j.earscirev.2013.11.001

Watkins A J, Wilson J B, 1994. Plant community structure, and its relation to the vertical complexity of communities: dominance/diversity and spatial rank consistency. Oikos, 70: 91. DOI: 10.2307/3545703

Worsley D, Broadhurst F M, 1975. An environmental study of Silurian atrypid communities from southern Norway. Lethaia, 8: 271–286. DOI: 10.1111/j.1502-3931.1975.tb00932.x

Wu Xi-zhi, Zhang Min, 2020. Applied regression and classification with R and Python. Beijing: Renmin University of China Press. 1–332 (in Chinese)

Xie Shu-cheng, Gong Yi-ming, Tong Jin-nan, Shi Xiao-ying, Lai Xu-long, Chen Zhong-qiang, Feng Qing-lai, Wang Hong-mei, Du Yuan-sheng, Wang Yong-biao, Yan Jia-xin, Zhang Ke-xin, Yin Hong-fu, 2006. Disciplinary shift from paleontology to geobiology. Chinese Science Bulletin, 51: 2327–2336 (in Chinese).

Xu Hai-peng, Zhang Yi-chun, Yuan Dong-xun, Shen Shu-zhong, 2022. Quantitative palaeobiogeography of the Kungurian–Roadian brachiopod faunas in the Tethys: Implications of allometric drifting of Cimmerian blocks and opening of the Meso-Tethys Ocean. Palaeogeography, Palaeoclimatology, Palaeoecology, 601: 111078. DOI: 10.1016/j.palaeo.2022. 111078

Xu Hong-he, Niu Zhi-bin, Chen Yan-sen, Ma Xuan, Tong Xiao-jing, Sun Yi-tong, Dong Xiao-yan, Fan Dan-ni, Song Shuang-shuang, Zhu Yan-yan, Yang Ning, Xia Qing, 2023. A multi-dimensional dataset of Ordovician to Silurian graptolite specimens for virtual examination, global correlation, and shale gas exploration. Earth System Science Data, 15: 2213–2221. DOI: 10.5194/essd-15-2213-2023

Yang De-ming, Pisano A, Kolasa J, Jashashvili T, Kibii J, Gomez Cano A R, Viriot L, Grine F E, Souron A, 2022. Why the long teeth? Morphometric analysis suggests different selective pressures on functional occlusal traits in Plio-Pleistocene African suids. Paleobiology, 48: 655–676. DOI: 10.1017/pab.2022.11

Yu Bin, Kumbier K, 2020. Veridical data science. Proceedings of the National Academy of Sciences of the United States of America, 117: 3920–3929. DOI: 10.1073/pnas.1901326117

Zhang Chi, Ronquist F, Stadler T, 2023. Skyline fossilized birth–death model is robust to violations of sampling assumptions in total-evidence dating. Systematic Biology, syad054. DOI: 10.1093/sysbio/syad054

Zhang Lin-na, Fan Jun-xuan, Wang Bo, Zhang Yuan-dong, Liu Jian-bo, Huang Hao, Chen Qing, 2023. Quantitative paleogeographical reconstructions and basin evolution of South China during the Ordovician. Earth-Science Reviews, 241: 104400. DOI: 10.1016/j.earscirev.2023.104400

Zhang Zhi-liang, Topper T P, Chen Yan-long, Strotz L C, Chen Fei-yang, Holmer L E, Brock G A, Zhang Zhi-fei, 2021. Go large or go conical: allometric trajectory of an early Cambrian acrotretide brachiopod. Palaeontology, 64: 727–741. DOI: 10.1111/pala.12568

Zhao Fang-chen, Caron J B, Bottjer D J, Hu Shi-xue, Yin Zong-jun, Zhu Mao-yan, 2014. Diversity and species abundance patterns of the early Cambrian (Series 2, Stage 3) Chengjiang Biota from China. Paleobiology, 40: 50–69. DOI: 10.1666/12056

Zhao Xian-ye, Yu Yi-lun, Clapham M E, Yan E, Chen Jun, Jarzembowski E A, Zhao Xiang-dong, Wang Bo, 2021. Early evolution of beetles regulated by the end-Permian deforestation. eLife, 10: e72692. DOI: 10.7554/eLife.72692

Zhou Min-zhen, Zhang Mi-man, Yu Xiao-bo *et al*., 1983. Translation collection of cladistic systematics. Beijing: Science Press. 1–209.