bio-101 www.bio-protocol.org/e1010602 DOI:10.21769/BioProtoc.1010602

高级阶元分子系统发育研究未来创新的思考

Thinking on the Future Innovations in the Molecular Phylogenetics at Higher Category Levels

谢强*

中山大学生命科学学院，有害生物控制与资源利用国家重点实验室，生物博物馆

*通讯作者邮箱: xieq8@mail.sysu.edu.cn

引用格式：谢强．(2021)．高级阶元分子系统发育研究未来创新的思考．Bio-101 e1010602．Doi: 10.21769/BioProtoc. 1010602.

How to cite: Xie, Q. (2021). Thinking on the Future Innovations in the Molecular Phylogenetics at Higher Category Levels. Bio-101 e1010602. Doi: 10.21769/BioProtoc.1010602. (in Chinese)

摘要：以 1970 年前后氨基酸序列比对方法的初步建立和碱基替换模型的初步研究为标志，分子系统发育研究已经走过了 50 余年的历史。20 世纪 80 年代，高级阶元系统发育研究所采用的主要分子标记是核基因 rDNA；20 世纪 90 年代，线粒体基因组数据的使用逐渐增多（植物研究中叶绿体基因组比线粒体基因组更常用）；进入 21 世纪之后，在原核生物、原生生物、植物、真菌、动物等各大类群的系统发育研究中都陆续出现了基于基因组和/或转录组等高通量测序数据所做的分析。在后基因组时代分子标记的数量即将达到极限的今天，或许是时候从整体上思考高级阶元分子系统发育研究未来的创新方向了。事实上，相对于在测序平台与技术方面已经发生的升级换代而言，分子系统发育底层方法论的重塑可能还在持续发展的过程中，未来在分类单元选取原则与方式、数据质量、建模方式、算法、化石记录在分歧时间标定中的标准化使用等各方面仍然存在巨大的进步空间。在高级阶元分子系统发育研究本身得到持续完善的时候，其服务于生命科学其他领域研究的能力也将获得更大的提升。

关键词：高级阶元；分子系统发育；创新；分类单元选取；模型；算法；化石；形态

研究背景

本文分为两个部分，第一部分是高级阶元分子系统发育分析的流程，从温故而知新的角度考虑，先回顾目前高级阶元分子系统发育分析的一般方法流程，其中包括过去几十年来一些重要创新的持续积累，以此来认识高级阶元分子系统发育方法论体系的基本特点和目前所达到的水平。然后，思考未来如何通过一些关键步骤的优化进一步提升高级阶元分子系统发育分析的表现。第二部分是高级阶元分子系统发育研究的成果如何发挥“生命参照系”的作用，放大到生命科学领域的发展主线，思考高级阶元分子系统发育研究如何更好地为生命科学其他学科领域做出贡献。

1 高级阶元分子系统发育分析的流程本身

Hennig (1966) 的《Phylogenetic Systematics》一书中，有一句话描述生物系统学在当时的生物学整体中所处的弱势及其原因，“If in this struggle for survival biological systematics has recently lost ground to other and, as is often heard, younger and more modern disciplines, this is not so much because of the limited practical or theoretical importance of systematics as because systematists have not correctly understood how to present its importance in the general field of biology, and to establish a unified system of instruction in its problems, tasks, and methods”。五十多年后的今天，当我们思考高级阶元系统发育研究目前的发展水平，虽然整个领域已经取得了巨大进步，但是如何更好地去构建一个全面包含目标类群、科学问题、研究任务和可靠方法的统一系统，仍然还有诸多问题待解。以下，我们先来简要回顾目前高级阶元分子系统发育分析的一般方法流程，然后思考其中存在的问题和未来的任务。

注：biological systematics 一词在 Hennig (1966) 中有广义和狭义的理解，此处可以理解为今天研究者们所说的 systematic biology

1.1 一般方法流程

高级阶元分子系统发育研究的一般方法流程包括分类单元选取、建立直系同源基因数据集、选择进化模型、系统发育重建、分歧时间推断 5 个大的步骤（表 1）。基于基因组等大数据的分析流程可以参考 Kapli *et al*. (2020)的总结与回顾。

表 1. 目前对于高级阶元分子系统发育研究主要步骤的一般认知

主要步骤			一般认知
分类单元选取			内群比较完整， 外群中至少要包括内群的姊妹群
整理直系同源基因	核基因或线粒体 rDNA， 线粒体蛋白质编码基因		手动整理
	核基因蛋白质编码基因		从 NCBI 和 ENSEMBL 等数据库 下载基因组序列
			用 OrthoFinder 或 OrthoMCL 等 建立 orthogroup
			用 Orthograph 调取整合直系同源基因
	序列比对		用 Muscle 或 MAFFT 判断碱基/氨基酸位置同源性
选择进化模型		替换	PatitionFinder、ModelFinder (IQ-Tree)
		插入、缺失	rDNA 可参考 rRNA 二级结构 进行一定程度的校正
系统发育重建	数据类型		碱基（核苷酸）、氨基酸
	单基因序列信息利用方式		基因联合、树形合意（溯祖）
	算法		最大似然法、 贝叶斯推断、 最大简约法（隐含模型）
分歧时间推断		进化速率模型	采用松弛型分子钟
		分歧时间标定	化石记录、进化事件

1.1.1 明确目标问题和分类单元选取（taxon sampling）范围

通常来说，这一步看似没有什么技术含量，但是这一步恰恰是非常缺少基于单因素对照研究而不断积累的经验性认识的，也缺少被明确表述的原则和要求。目前，较高质量的研究通常至少要参照一个最主流的分类系统（为方便表述，这里假设该分类系统只

跨 2 个阶元层级，例如某纲内部分目，或某科内部分亚科），对于该分类系统中同一阶元层级的各分类单元各选取至少 1-2 个物种作为代表。即使暂时无法做到每一个分类单元中都有代表物种，有代表物种的分类单元数量应至少达到主流分类系统中所列出的分类单元总数的三分之二以上，并且对于认识目标问题相关的竞争性假设（competitive hypotheses）来说应该包括所有直接相关的分类单元。

1.1.2 建立直系同源基因数据集，并进行序列比对

根据后续的分析需求不同，如果采用树形合意（coalescent，也称为溯祖）的方式，各直系同源基因序列的矩阵文件互相独立；如果采用基因联合（concatenation）的方式，所有基因数据需要被整合为一个矩阵。在分子标记数量不超过 100 个的情况下，手动整合矩阵是可能的；当分子标记数量超过 100 个之后（基于基因组和转录组得到的直系同源基因数量通常在 1000 个以上），手动整合将非常耗时、并可能出现难以察觉的低级错误，通常需要编写或下载计算机语言脚本进行处理。

1.1.3 选择碱基或氨基酸替换模型

一般对实验获得的各片段序列分别先进行分子进化模型选择，然后再根据具体模型类别进行适当的合并，将具有相同或相近进化模型的片段序列合并处理。分子进化模型通常是指碱基或氨基酸替换模型，有的时候也涉及密码子替换模型或碱基对替换模型；目前几乎所有的分子进化模型都不包括对于插入缺失（indels）的处理，从目前系统发育基因组学的发展形势来推测，由于数据类型的局限性，提取插入缺失事件中系统发育信号的研究可能未来也不会成为主流。此外，这里所说的替换模型通常是指基于替换类型、碱基或氨基酸频率、多重替换校正等参数的，被明确表达的模型，后续可以被直接用于最大似然法或贝叶斯法的计算，而不是指简约类算法中隐含的模型。当分子标记数量超过 100 个之后，通常需要计算机语言脚本调用模型测试程序进行处理。实际上，碱基或氨基酸的替换模型是进行了较为完备参数化的进化假设，而从进化假设这个本质来说，很显然简约类算法也是含有进化假设的。

1.1.4 基于不同数据类型、不同数据集、不同算法进行系统发育重建

不同数据类型通常包括氨基酸、碱基（核苷酸），对于蛋白质编码基因而言，碱基的

使用还可以区分为使用或不使用密码子第三位的碱基。数据集的不同可以因多种不同角度的考虑而形成区别，例如树形合意和基因联合这两种不同的数据利用方式，氨基酸和碱基这两种不同的数据类型，碱基数据中密码子的第三位是否使用，从进化速率等不同角度对蛋白质编码基因所做的区分和筛选，rRNA 和 tRNA 的使用与否，等等。不同算法主要包括最大似然法（maximum likelihood）、贝叶斯法（Bayesian method）、最大简约法（maximum parsimony）、距离法（distance methods）。从使用频次来看，目前最主流的是最大似然法，贝叶斯法次之，最大简约法再次，而非加权组平均法（UPGMA）、邻接法（neighbor joining）和最小进化法（minimum evolution）等遗传距离类算法（genetic distance methods）目前很少在高级阶元分子系统发育研究的建树步骤使用。

1.1.5 分歧时间推断

近年来，非线性（松弛型 relaxed）分子钟模型的建立、发展和持续优化，及其在相关软件中的实现（Drummond and Rambaut, 2007; Bouckaert *et al*., 2019）和在分歧时间推断中的应用，是分子系统学发展的一大亮点，大大提高了不同分歧时间推断研究结果之间的一致性。略显遗憾的是，由于计算量的限制，目前对于大型矩阵的处理还比较困难（可操作分类单元数量比基因数量的影响更大）；对于什么样的基因更适于用于分歧时间推断，也缺少案例积累和较为一致的观点；而对于化石记录的标准化使用，可能还有很长的路的要走。

需要注意的是，虽然分歧时间推断在整体流程中几乎处于分析流程的最后环节，但是如果要提升分歧时间推断的效果，实际上在最初进行分类单元选取的步骤就已经需要开始进行相关考虑和项目设计（Wang *et al*., 2016）。

1.2 高级阶元分子系统发育分析流程的未来创新思考

1.2.1 到底应该如何进行分类单元选取最恰当？

对于高级阶元分子系统发育分析来说，最恰当的分类单元选取有 2 个可能的答案，一个是物种或 OTU 越多越好，另一个则是并非越多越好。对于这个问题的回答，既关系到研究者们最终建成生命之树（Tree of Life）的方式，也关系到如何去做好常规的高级阶元分子系统发育分析。

物种越多越好，这是一个极具吸引力的答案，因为如果是这样的话，那么当未来计

算能力足够的情况下，人们可以在一棵进化树中一次性容纳下地球上所有已知的物种。但是，即使算力足够，这种方式仍可能面临一系列问题的挑战。

第一，树形空间分布的扁平化（flattening）。当一棵树中包含的 OTU 越来越多的时候，不同拓扑结构树形的总数也会极大增加，这有可能会造成最优的树形在树形整体分布中不够突出、不容易被找到。这样的扁平化是否一定会发生，目前还缺少有力证据，似乎更多是一种担心。我们曾经进行过一些测试，在相同分子标记体系、相同类群范畴、相同模型、相同算法的情况下，当 OTU 总数超过 1000 之后，树形在多个深部节点都会出现变化。但是考虑到分子标记数量的局限、类群代表性的完整程度的局限，我们无法判断这样的结果所给出的指向性是否足够牢靠。

第二，物种数量在高级阶元水平分布的不均衡。物种数量在高级阶元水平分布的不均衡可以达到很高的程度。以昆虫纲为例，鞘翅目 Coleoptera 已知现生物种数量超过 38 万，而蛩蠊目 Grylloblattodea 和螳螂目 Mantophasmatodea 已知的现生物种数量分别为 33 个和 23 个，鞘翅目与后两者的物种数量相差一万倍以上；在科级水平，物种数量超过 1 万的科不在少数，而单型科（仅含 1 属 1 种）也并不少见。也就是说，不论在目级或科级阶元水平，不同分类单元之间的物种数量差异都可以高达 10000 倍以上。在物种分布格局偏异（bias）如此巨大的情况下，如果以物种越多越好作为原则，是否会造成系统发育重建结果的偏异？目前这个问题似乎没有确定的答案。我们曾经进行过一些测试，在相同分子标记体系、相同类群范畴、相同模型、相同算法的情况下，对物种丰富的分类单元进行相对大量的取样（例如，在各门、纲、目、科级类群中同步向属级推进）似乎并不利于系统发育重建的整体表现。但是，类似上述情况，考虑到分子标记数量的局限、类群代表性的完整程度的局限，我们无法判断这样的结果所给出的指向性是否足够牢靠。

第三，可用直系同源基因数量的减少。Thomas *et al.* (2020) 对节肢动物 21 目 76 个物种（其中六足动物 12 目）基因组的比较研究显示，基因家族（gene family）生成和灭绝的发生在高级阶元水平是频繁的，有时还较为剧烈；这 76 个物种之间只存在 150 个共有的单拷贝蛋白质编码基因。可以设想，未来在所对比研究的目级分类单元更多、所包括的物种也更多的情况下，对于节肢动物高级阶元系统发育重建而言，严格意义上直系同源基因的数量很有可能下降到不足以构建基本框架（backbone）的规模。

通过以上的叙述，或许我们可以认为，即使未来计算资源足够的话，把全部已知物

种都一次性纳入到一棵进化树中也未必是最恰当的方式。如果这样的认识是对的，未来在通向生命之树的过程中势必要进行分部、分步组装，在这样的背景下，到底应该如何进行项目设计和类群选取呢？

第一，综合考虑确定性和不确定性，给可能出现的不同结果留出足够的空间。以细菌 Bacteria、古菌 Archaea、原生生物 Protists（并系群）、植物 Viridiplantae、真菌 Fungi、动物 Metazoa 这个细胞生物进化的最大背景为例，如果要研究细菌内部的系统发育关系，由于细菌整体的单系性（monophyly）是确定的，细菌与古菌或者说细菌与古菌+真核生物之间的姊妹群关系也是确定的，而细菌内部门与门之间的关系存在很大的不确定性。那么，外群选取应该至少包括古菌每个门各 1 个物种（注意在此情况下得到的古菌内部的门与门之间关系未必可靠），如果对于直系同源分子标记的数量并不造成太大困扰的话，还可以再包括一些真核生物物种（注意在此情况下得到的真核生物内部的关系以及古菌与真核生物之间的关系未必可靠）；在内群方面，已知的每个门应该至少有 2 个代表物种（除非这个门仅已知 1 个物种）、并且这两个代表种之间的亲缘关系应该尽量远（这实际上依赖对于门内分类系统或系统发育的认识）、以最大限度地代表该门的遗传多样性，此外，未被纳入已知的门的地位待定物种也应该先按照潜在的新门的可能性进行选取。如果要研究原生生物的内部关系，或者植物、真菌、动物各自在生命之树中的系统发育地位，都不可避免地要同时包括上述细胞生物 6 大类群，其中细菌和古菌作为连续外群（successive outgroups）置根，在原生生物中，除了 SAR（Stramenopiles-Alveolata- Rhizaria）等稳定的大型单系群（monophyletic group）可以适当控制取样规模，其它类群应参照已有的分子系统发育研究和未被明确认识的遗传多样性，在科甚至属级分类单元做充分的物种选取。

第二，作为追求的目标，一项高级阶元分子系统发育研究应该参照待研类群所有已有的分类系统、对照并认识其间异同，最大限度地去包括各局部所涉及的竞争性假设，对于同一阶元层级的各分类单元各选取至少 2-3 个物种作为代表（除非某分类单元仅含 1 属 1 种），并且这些代表种之间的亲缘关系应该尽量互相远离、以最大限度地代表各分类单元内部的遗传多样性。

第三，对于那些在体形和生物学习性方面分化较为显著的类群（通常包含物种较多），尤其是那些从形态学角度难以给出可靠的自有衍征（autapomorphy）或共有衍征（synapomorphy）的类群，应该主动作为、充分考虑其并系群（paraphyletic group）

的可能性，有针对性地在局部进行更为细致的类群选取，对于并系群的发现给予充分的机会。如果得到证实，及时对分类系统做出相应的修订，这样在未来系统发育研究中分类单元选取才能不留死角。

1.2.2 直系同源基因的原则还能走多远？

正如上述所提及的，无脊椎动物基因组虽然未必如植物基因组一样频繁出现多倍体化，但是基因拷贝数的增减在高级阶元水平的发生规模还是大大超出了此前的预计，未来在对比研究的门、纲、目、科级分类单元更多，所包括的物种也更多的情况下，严格的单拷贝基因的数量有可能非常少。而与此同时，在系统发育大的基本框架足够稳定、可靠之前，研究者们可能又无法通过缩小目标类群范围的方式来保住较高数量的单拷贝基因。在这样的两难局面下，有没有可能未来发展出一些方法，将一个基因家族中的多个拷贝整合或混合成一个虚拟拷贝，作为与近缘类群中真正的单拷贝基因形成直系同源关系的基因序列？这可能是一个值得探索的研究方向。对于那些在各拷贝之间分化比较强烈的基因家族来说，或许难以实现；但是对于那些在各拷贝之间分化相对较弱的基因家族来说，应该还是有希望的。毕竟，核基因 rDNA 也是家族而并非单拷贝，线粒体基因组也不是核基因意义上的单拷贝，从这个角度来说，高级阶元分子系统发育研究中其实一直就存在如何更好面对多拷贝基因这个挑战，只是这个挑战此前没有显得那么突出。当然，相关的建模可能需要足够多高质量测序和组装的基因组数据（二代和三代混合，或者三代有更大的升级）作为数据基础。

1.2.3 以基因片段为基本单位进行分析的逻辑基础是什么？

高通量测序数据全面进入动物的高级阶元分子系统发育研究有 10 年左右的时间了。从基因联合的角度来说，在这十年间，研究者对于大数据的利用方式和之前对于核基因 rDNA 以及线粒体基因的利用方式没有本质区别，依然是对于每个片段测定碱基或氨基酸替换模型，然后把所有片段序列和替换模型汇总起来，主要的改变更多只是所使用基因片段数量规模的扩大。以片段作为序列数据的基本操作单元，看似没什么问题，但是其实缺少稳固的逻辑基础，更多只是一种研究过程中所使用方法流程的惯性。当然，人们必然是经由基因时代进入基因组时代的，不是一步跨入基因组时代，但是既然分子系统发育研究已经进入了基因组时代已有十余年，或许是时候考虑更深层次的方法论拓展

了。

第一，不论是系统发育基因组学（phylogenomics）研究，还是之前基于少量基因序列的研究，在很多时候研究者所使用的都是不完整的片段，而不是基因的全长。从分子生物学的角度来说，基因通常是发挥功能的基本单位，分子系统发育研究中理应追求获得全长的基因序列。从现实的分析流程来看，目前的系统学研究通常是拿到什么样的片段范围，就这样用了，没有过程合理性方面的论证。基于不同分析结果之间的相合性（consensus，或者说一致指向）来显示系统发育信号的稳健性（robustness），虽然必要、但是不充分，也无法代替纯粹理性本身，并且在现实中并不总是形成稳定的输出。

第二，对于高级阶元分子系统发育研究来说，追求基因序列的全长或许仍然不够，原因在于基因的新生、拷贝数变化、丢失远比此前想象的更加频繁。不论是对于单拷贝基因集合来说，还是对于 1.2.2 中所提及的各物种基因组中的多拷贝的单一化处理，或许是时候从基因组整体的角度考虑建立碱基或氨基酸替换的进化模型了，也就是为基因组进化建模而不只是为基因进化建模。这是一个极具挑战也极具吸引力的探索方向，考虑到酵母基因组已知的序列数据比较丰富、结构相对于动物和植物基因组简单，或许可以先从酵母基因组做起。

第三，不论是以核基因 rDNA 和线粒体基因组为主要分子标记的研究案例，还是数百或数千基因规模的系统发育基因组学研究案例，研究者们普遍有一种感觉，就是基因联合的表现总是比各个单基因的表现要好得多。在系统发育基因组学的研究中，基因树形合意的表现也比各个基因的单独表现要好得多，大量的单基因甚至不具备再现一些毫无疑问的单系群的能力。这一现象令人困惑，但是既然多基因整体的表现总是更好，或许研究者们确实应该认真考虑如何把“多基因整体”提升到基因组的层次了，毕竟，目前的系统发育基因组学研究仍然留下了不少表现欠佳的树形局部解析。

总的来说，以基因片段为基本单位进行数据处理更多只是一种习惯、缺少逻辑基础，未来应该在追求基因全长序列的基础上探索为基因组进行整体建模。

1.2.4 二维矩阵与目前分子系统发育分析流程隐含的基本假设

到目前为止，绝大多数分子系统发育分析的数据集基本模式都是经过序列比对之后的二维矩阵。事实上，这其中很可能存在一个内在矛盾，并且在高级阶元水平更为突出。分子系统发育分析可以被视为一个从二维矩阵中提取系统发育信号（phylogenetic

signals）的过程，但是有一个问题很少被提及，那就是，系统发育信号是否可以被精确描述和分析？目前，几乎所有的高级阶元分子系统发育研究案例都止于树形相合性和自展检验（bootstrap）等数值评估，而这些其实都是“黑箱”输出的结果，当前的研究实践似乎都在默认，如果有比较强而稳健的信号展示出来，就算是已经得到了系统发育信号。从本质上说，二维矩阵中每一列碱基或氨基酸就是信号和噪音的基本载体，而树形结构是一个包含众多内部节点的层级结构。这就意味着，某一列数据对于某一个或某几个节点是信号，而对于其它节点可能是噪音或无影响的中性数据。而在一个二维矩阵中，所有这些信息是混合在一起的。与此相关的另外一些问题是，从二维矩阵整体上来说，其中包含信号、偏异噪音（biased noise）、白噪音（white noise），在从二维矩阵形成树形的过程中，研究者们是否可能全面抑制偏异噪音、避免其形成假阳性（false positive）结果？对于信号和偏异噪音，可靠的判定依据应该是什么？分类单元选取的完整程度与信噪比格局之间是否存在关联？当然，对于这些问题的回答需要足够多高质量测序和组装的基因组数据（二代和三代混合，或者三代有更大的升级）作为数据基础，目前主要基于二代测序的系统发育基因组学研究中所形成的数据集在数据覆盖度（尤其是位点覆盖度）方面还有较大的不足。

总的来说，当数据规模所带来的红利释放殆尽，如果仍然存在一些没有得到良好解析的树形，或许是时候考虑深层次的攻坚克难了。随着二代测序解放了数据数量的快速获取，随着三代测序（及其与二代测序和一代测序的结合）即将在数据数量的基础上大幅度提高数据质量，研究者们或许有机会、也应该向着内在的、深层的、精准的方向去进行探索，而不是过早满足于目前基因联合与树形合意方式下的最大似然法分析结果之间的相合性的提升，以及基因联合方式下的最大似然法、贝叶斯法、最大简约法分析结果之间的相合性的提升。

1.2.5 算法的未来

在 1.1.4 中曾经提及，目前最主流的算法是最大似然法，贝叶斯法次之，最大简约法再次。这其中主要是出于 2 个方面的原因，第一，在对进化量的估算方面，最大似然法和贝叶斯法所基于的碱基或氨基酸替换模型被认为比最大简约法在模型表述方面的隐蔽和简单相比，有更好的表现。这个思考角度不无道理，毕竟，即使对于相同的后续算法，以 Lartillot and Philippe (2004) 在 PhyloBayes 中对氨基酸替换模型所做的优化

（CAT）为例，对于基于少量核基因蛋白质编码基因或/和线粒体蛋白质编码基因的一些高级阶元分子系统发育研究案例来说，使用 PhyloBayes 比使用 MrBayes 可以得到与其它证据一致性更强、更容易解释的结果。类似的，近来也出现了面向最大似然法的研究，被称为 GHOST 模型（Crotty *et al.*, 2020）。第二，在计算效率方面，最大似然法虽然不如最大简约法、但是比贝叶斯法有明显的优势。近年来发展起来的基于最大似然法的建树软件 IQ-Tree (Nguyen *et al.*, 2015)，在某些案例中可以做到在相同或相近系统发育结果表现的情况下比 RAxML 有明显的效率提升。综合这两个方面的因素，使得最大似然法成为目前最主流的算法。

算法在未来会如何发展呢？这里思考的第一个问题是，最大简约法还有没有未来？Misof *et al.* (2014) 在六足动物系统发育基因组学研究的建树环节将最大简约法作为后续最大似然法搜索过程中生成起始树的算法。Zhang *et al.* (2020) 利用基于最大简约法得到的向导树加速贝叶斯法中马尔科夫链的收敛。此外，越来越多基于核基因转录组数据的系统发育研究表明，最大简约法可以与最大似然法在结果方面达成高度的一致。这样看来，最大简约法未来仍然会是一种主流算法，或者至少以幕后英雄的方式作为一种主流算法，比如，有没有可能将其用来提升其它更多软件进行系统发育或分歧时间推断的计算效率，值得探索。

接下来的问题是，在最大似然法、贝叶斯法、最大简约法之外，还可以做些什么？回顾分子系统发育分析的整个流程，在二维矩阵的方式不做改变的情况下，最有可能发生可期的重大改进的环节可能还是碱基和氨基酸各自的替换模型。除了 1.2.2 和 1.2.3 中所提及的建模问题，这里所说的模型优化是指是否可以对进化历史上真实发生的分子进化改变量做出更贴近实际的描述和估算，如果这一环节无论如何都是一个黑箱，是否可以让基于深度学习（deep learning）的人工智能（artificial intelligence）发挥作用？相对于系统发育信号这样的纯粹理性话题，模型相关的探索更贴近系统发育研究的现实需求。事实上，在种群遗传学领域中的相关研究之后，在高级阶元分子系统发育方面也已经有相关研究开始出现（Suvorov *et al.*, 2020）。也就是说，在分子系统发育研究领域诞生 50 年左右之后，最需要被提升的还是模型方面的研究，正是 50 年前这个领域诞生之初的代表性工作的研究方向。

此外，在更长远的未来，当计算能力不再是瓶颈的时候，对于系统发育重建本身来说，作为一个更接近终极的目标，有没有可能实现对于树形概率分布格局的类地形可视

化？如果不可能，为什么？如果可能，应该怎么进行？不同的树形（拓扑结构）之间是互相离散的，如果能够在一个以似然值（类比于地形学中的海拔高度）、进化量（模型估算）、自由度（由分类单元选取和分子或形态特征信息共同决定）为坐标的几何空间中考察树形整体所形成的分布格局，或许有望对于系统发育推断这个“黑箱”的理解前进一大步。在自由度方面，按目前的系统发育实践的一般经验来看，数据集所使用的有效信息（碱基、氨基酸、形态特征）越少，越难得到一个可靠的完全二叉树；所包括的末端分类单元过少、过多，或者相对于分类系统而言取样不够完整、不平衡，都不利于得到一个可靠的完全二叉树。到目前为止，探索树形分布格局的类地形可视化的研究似乎极少（Sundberg *et al.*, 2010），可能主要是受困于目前算力瓶颈的局限。

1.2.6 化石记录的标准化使用

回顾高级阶元分子系统发育研究的基本流程，可以发现除了分类单元选取之外，几乎其它所有环节的分析都已经标准化了，另一个还没有标准化的环节就是在进行分歧时间推断的过程中，如何恰当地使用相关的化石记录。造成这个情况的原因主要有三个。

第一，很多类群中依然缺乏足够可靠的高级阶元系统发育研究结果，进行后续分歧时间推断的研究基础还不够牢固。

第二，很多高级阶元系统发育研究在进行分类单元选取的时候并未事先考虑到化石记录的使用问题，只是在系统发育分析完成之后才开始考虑，造成最后可选化石记录的缺少，或者所选化石记录的不当。

第三，化石记录相关信息的可靠性存在不确定性，尤其形态学鉴别特征、分类地位、以及相对于现生类群的所处的系统发育地位。

此外，一些研究在试图解答分歧时间的树形局部进行分歧时间标定，造成循环论证和很大程度的人为结果。对于标定点来说，虽然其后验结果与先验设定之间存在一定的数值差异，但是这种差异通常非常有限，对于结果的定性讨论几乎不形成影响。

表 2　高级阶元分子系统发育研究未来可能的创新方向

主要步骤	创新方向
分类单元选取	充分参考不同的分类系统（竞争性假设）
	充分甄别并系群、修订分类系统
	每个分类单元至少选取 2-3 个代表物种（单型情况例外）
整理直系同源基因	充分提取利用非单拷贝基因中的系统发育信号
建立碱基或氨基酸替换模型	探索基因组规模的建模
	引入深度学习的方式
系统发育重建	最大简约法的潜力挖掘
	树形概率分布的可视化
	系统发育信号的理论思考
分歧时间推断	现生类群及其与化石类群之间的比较形态学研究
	对分歧时间标定中使用的化石记录建立标准化体系

回顾分子系统发育和分歧时间推断近年来的发展历史，可以发现，在高通量测序技术大大解放了数据数量、还将大幅提升数据质量的大背景下，高级阶元分子系统发育这个领域从来没有像现在这样有机会去融汇多个学科领域的贡献，同时，也从来没有像现在这样特别需要去融汇多个学科领域的贡献，以期获得更深层次、更高质量的发展。这其中包括，需要计算生物学（数学、计算机科学、生物信息学）在模型构建和优化方面做出贡献，同时也包括，需要分类学、形态学、古生物学在分类单元选取和化石记录选取的标准化方面发挥支撑作用（表 2）。

2 高级阶元分子系统发育的“生命参照系”作用

Hennig (1966) 在《Phylogenetic Systematics》一书中提到，“a phylogenetic system is to be preferred among all conceivable biological systems… by ‘phylogenetic system’ we mean a system that expresses the phylogenetic relationships of

organisms… the phylogenetic system as a general reference system for biological systematics”，认为一个基于系统发育关系的生物系统在生物学所有的系统中是最可取的一个作为其它系统（如形态系统、生理系统、生态系统等）参照系的系统。做出这样判断的主要原因在于，相对而言，以系统发育关系作为参照系去关联其它生物系统最为容易、直接。此外，支撑系统发育系统的亲缘关系是可以相对精确量度的，并且通常与形态特征相似性的对应程度较好，因而实用性比较好。事实上，当系统发育基因组学（phylogenomics）这个词最初被提出的时候（Eisen and Fraser 2003），其本意就是试图表明，一方面，基因组数据可以帮助进行系统发育解析，而另一方面，更加可靠的系统发育树的构建也可以从进化角度为基因组比较研究提供强有力的支持，包括基因家族进化，以及从大量报告的可疑的横向基因转移（lateral gene transfer, LGT）事例中鉴别出那些真实的案例，等等。

2.1 依托高级阶元分子系统发育研究结果的常见后续研究

尽管高级阶元分子系统发育研究本身依然存在不少需要改进的环节，但是其作为“生命参照系”的作用已经在若干领域有了很多体现，其中最主要的有三个，一个是生物地理学（biogeography），一个是选择压力分析、尤其是正选择压力位点的探测（Yang, 2007），一个是特征进化与祖征重建。

*注：确切地说是分支生物地理学（**cladistic biogeography**），但由于生态生物地理学（**ecological biogeography**）越来越多地用生态地理学（**ecography**）一词表述，似乎生物地理学与分支生物地理学的意义正在接近。*

2.2 生命科学研究整体的发展主线

20 世纪，在生命科学进入分子生物学时代之前，学科整体的发展主线是“生理或病理过程-表型”；在生命科学进入分子生物学时代之后，学科整体的发展主线是“分子遗传型-表型”，这两条线也可以被合并为“内在原因-外在表现”一条线。这一特点在学科发展历程、诺贝尔奖颁奖历史、以及平时具有较大影响力的论文发表等几个方面都可以得到充分的体现。

以具有悠久历史、近年来快速发展的生物地理学为例，之所以受到广泛重视，其主要原因在于相关研究既可以揭示基因型和/或表型的分化与地理因素之间的关联，又可以

经由地理角度与相关生态学研究形成进一步的关联和整合。以近年来快速发展的共生微生物研究领域为例，之所以受到广泛关注的主要原因在于，人们越来越意识到动物或植物的诸多表型无法单纯由基于对其自身基因组的研究而得到很好的认识，某些表型的成因甚至主要在于共生微生物。此外，进化发育生物学（EvoDevo）是很典型的从进化视角出发以发育生物学的方法论研究分子遗传型-表型关联的一个领域，其中最著名的代表性发现是以黑腹果蝇 *Drosophila melanogaster* 为模式生物对体节分化和平衡棒-后翅进行的研究，相关成果于 1995 年获得诺贝尔生理医学奖。以分子系统发育研究自身为例，已有的动物高级阶元分子系统发育研究表明，那些可重复性较好、认可度较高的单系群在很多情况下都有形态学衍征的支持。

以诺贝尔生理医学奖和化学奖的颁奖历史为例，包括 DNA 双螺旋结构的解析等著名发现在内，与中心法则及其扩展相关的发现有 30 项左右；而像 2005 年获奖的“幽门螺杆菌-胃炎胃溃疡”等与中心法则无关的发现，在很大程度上也可以被视为“分子遗传型（消化道微生物）- 表型（病理）”的一例。

再以古生物学领域为例，这一领域看上去似乎远离上述主线，但其实这主要是由于古生物学的数据很少涉及生物大分子层面。众所周知，每当带有奇特表型（直接证据通常是形态特征）的化石类群被发现，大都会引起广泛关注，其背后的原因主要在于相关化石的发现和研究可以大大拓展人们对于形态表型多样性本身的认识；并且，人们还会不禁去联想是什么样的分子遗传型支撑了那些奇特的表型，这也是为什么古 DNA 与古蛋白质的相关研究会引人关注（虽然其实很难在古序列与化石类群的表型之间找到有效的关联）。

在很多研究中，表型未必是形态学层面的，也可以是代谢、细胞、生物大分子层面的。此外，在涉及“功能（function）”研究的时候，生命科学各领域的研究还会大量涉及“适应（adaptation）”这个话题；在涉及不同类群、不同结构、不同遗传系统之间的关联研究的时候，还会大量涉及“协同”和“互作”等话题。

2.3 与高级阶元分子系统发育相关的若干创新

2.3.1 物种多样性研究本身

新物种的不断发现、物种界限的精准判定、以及以属的修订为基础的分类学工作依然具有重要意义，因为只有通过面向各类群的分类学工作的持续积累，人们才有可能最

终回答目前地球上到底有多少个物种这个问题；与此同时，也正是由于分类学家在发表新物种的时候对于形态学表型的细致区分，研究者们的点滴积累已经汇聚成了一个巨大的形态学表型变异的知识库。在主要基于形态特征进行分支分析学（cladistics）研究的时候，研究者们通常一方面关注生物有机体类群之间的亲缘关系，另一方面也十分重视形态等表型的进化过程。在分类学与系统学研究中，从生命科学整体的角度来看，那些具有此前未知的奇特表型的、系统发育地位比较特殊的新物种会受到更多关注，一般来说这样的新物种的发现还通常伴随新的高阶分类单元的建立。例如，近年来对细菌界中CPR（candidate phyla radiation，候选门适应辐射）类群的认识（Brown *et al.*, 2015），以及对古菌界中与真核生物具有更近缘关系的类群的认识（Spang *et al.*, 2015）。在六足动物中，Klass *et al.* (2002)基于分别采自 1909 年和 1950 年的标本建立了新属螳蝽属 *Mantophasma*（含 2 新种），并在此基础上建立了新目螳蝽目 Mantophasmatodea。

类似的，对于已知物种，如果其奇特属性此前未被充分认知，后来发现其在高级阶元系统发育格局中的位置比较特殊，也会受到较多关注，并伴随新的高阶分类单元的建立。例如，在后生动物界中，*Xenoturbella bocki* 是 Westblad (1949)建立的一个新种（标本最早采自 1915 年）、最初被置于扁形动物门，Bourlat (2006)通过反复进行系统发育研究认为其应该作为一个新的门。

总的来说，由于可以丰富人们对于生物表型的认识、提升未来高级阶元系统发育研究中分类单元选取的完整程度，那些可以支撑新建高阶（门、纲、目、科）分类单元的新物种的发现更受关注。相应地，对于中国领土领海范围内物种多样性的认识，除了可以支撑新建高阶分类单元的新物种的发现之外，新记录的高阶分类单元（如纲、目、科，基于新种或已知种）通常也会受到较多关注。

2.3.2 比较形态学研究的价值

由于化石类群缺少生物大分子序列方面的信息，但同时又携带了大量有趣的、有研究价值的形态表型，因此，如何持续增强对于化石类群与现生类群之间形态学特征的比较研究，是将两者恰当纳入一个分类系统的基础，也是未来在分歧时间推断研究中准确使用化石记录对分歧节点进行标定的基础。

2.3.3 高质量系统发育基因组学的桥梁作用

进入 21 世纪以来，分子生物学及相关学科依然是持续揭示“分子遗传型-表型”的相关性的主力；与此同时，相关研究也存在一定的局限性，其中最主要的一点可能是目前的方法论体系更多是对一个物种内部的分子遗传型变异-表型变异关联研究比较有效，种间水平的相关研究的数量要少得多，而属间水平的相关研究则非常困难。值得注意的是，很多长期受到多方高度关注的表型间断变异都是在宏进化尺度上存在的，例如植物界中的被子植物 Magnoliopsida（有花植物 flowering plants）与其它植物、动物界中的两胚层动物 Diploblastica 与三胚层动物 Triploblastica（两侧对称动物 Bilateria）、两侧对称动物中的原口动物 Protostomia 与后口动物 Deuterostomia、六足动物 Hexapoda 中的有翅类昆虫 Pterygota 与无翅类群（进化上原生无翅）、有翅类昆虫中的完全变态类 Holometabola 与非完全变态类，等等。这些两两对比的生物大类之间，尽管各自内部存在千差万别的其它变异，但是对于关键表型的有无，各自内部的表型都高度稳定，而在两两之间则形成间断。对于这些表型间断的分子遗传机制有效研究，很可能需要包括高级阶元分子系统发育在内的多个学科的合力。

有初步的研究表明，分子衍征（molecular apomorphy）在从界与界之间到物种复合体（species complex）与物种复合体之间的高级阶元水平可能普遍存在（Xie *et al.*, 2012; Wu *et al.*, 2016）。所谓分子衍征，这里是指类群特异性的碱基或氨基酸，可以被理解为：在一棵二叉树中（主要是系统发育基因组学研究），在每一个二分支节点，总是在某个或某些直系同源基因中存在若干碱基/氨基酸位点，在这些位点上，两个姊妹群（sister group）之间的碱基或氨基酸状态完全不重叠，而在两个姊妹群各自内部完全一致或高度一致（根据氨基酸的理化性质类别判断；图 1）。

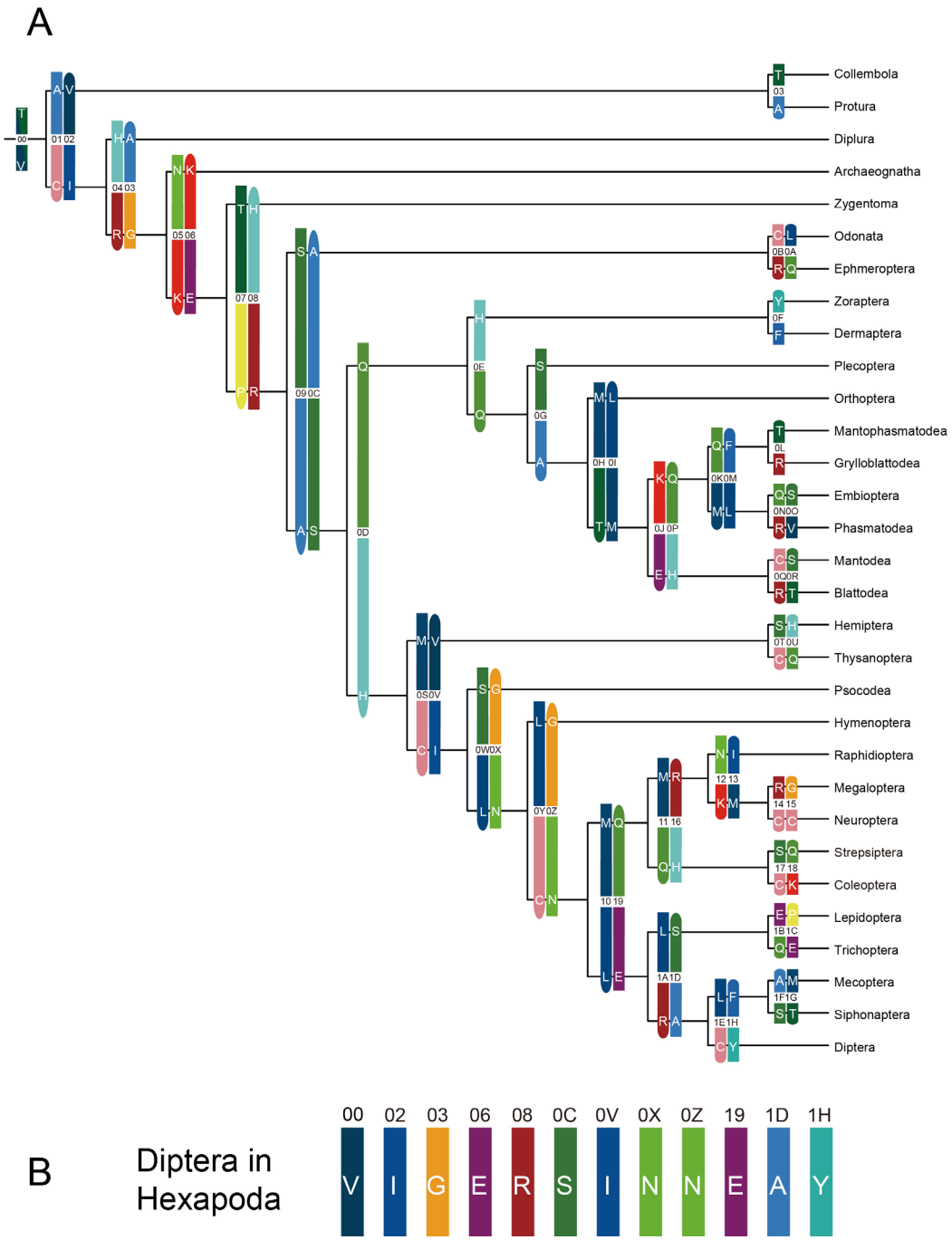

图 1. 依托系统发育基因组学得到的二叉树形结构提取基因组或转录组中的氨基酸分子衍征信息（A）；并以双翅目在六足动物亚门（节肢动物门）中的系统发育地位为例，基于分子衍征信息编制分子鉴定条形码（B）。图中长方形色块表示衍征，末端呈弧形的色块表示祖征或姊妹群的衍征（引自 Wu *et al*. 2016）。

依托树形结构，普遍存在的分子衍征可以被组织成层级式的分子衍征体系，用于支撑高阶分类单元和物种的标准化描述。这样的标准化描述可以有 2 个方面的用途，第一，当生命之树（Tree of Life）建成的时候，还可以同步生成生命检索表（Key to Life），也就是一个可以容纳地球上所有物种特征描述的检索表，在这个二分式检索表（dichotomous key）中的检索项可以被标准化设置为“某基因某位点的碱基或氨基酸为 X”；第二，当 2020 年刚刚获得诺贝尔化学奖的基因编辑技术未来大幅升级到基因组编辑技术的时候（例如一次实验操作可以编辑 100 个以上的碱基位点），或许人们将有机会看到类群特异的碱基或氨基酸位点与类群特异的表型之间是否存在关联。至少，基于“结构与功能相关”这一在生物学各领域普遍适用的假设，基因组中类群特异性的碱基或密码子（氨基酸）应该具有类群特异性的功能。此外，对于人类基因组中所发生的那些比较少见的、通常表现为疾病的位点变异，人们将有机会从宏进化角度形成更为直观的认识，也就是说，某个碱基或氨基酸位点上所发生的变异，此前到底是在什么阶元级别水平表现为保守，是真核生物、动物、两侧对称动物、后口动物、脊索动物、脊椎动物、四足动物、羊膜动物、哺乳动物、或者灵长目？虽然对于疾病治疗本身未必有价值，但是至少可以提供一个参照系，评估各位点变异在宏进化角度发生的罕见程度。

小结与建议

随着高级阶元分子系统发育研究中所使用的分子标记数量大幅提升、分析流程的标准化逐步完善、非线性分子钟模型在分歧时间推断中的应用，该领域研究成果所给出的树形结构可靠性得到了大幅提升。

但是鉴于目前主要基于二代测序数据的系统发育基因组学研究结果仍然存在不少表现不如预期的情况，未来应该在多个方向进行深度探索，其中包括数据质量（完整性）、多拷贝基因序列信息的充分利用、基因组进化建模、基于深度学习的碱基或氨基酸替换模型改进、系统发育信号的理论研究、分子衍征体系的构建等侧重数据分析的方向，也包括分类单元取样完整程度的提升、高阶分类单元的建立与高阶分类系统修订、现生类群及其与化石类群间的比较形态学研究、化石记录在分歧时间推断分析中的标准化使用等侧重分类学与形态学的传统方向。

未来，随着系统发育基因组学研究结果表现的持续提升，所有依托树形结构开展的后续研究也将得到提升；随着对于系统发育信号的精准解读，系统发育数据分析过程这

一黑箱可能被揭开；随着基因组水平分子衍征的全面总结和基因组编辑技术能力的大幅提升，系统发育基因组学研究有望更好地连接起生命科学中生理（广义）、进化、生态三大研究传统，充分发挥“生命参照系”的作用。

建议：由于本文涉及话题比较广泛，并且侧重未来创新思考本身，因此只引用了很少的参考文献，以便聚焦对于未来创新的思考本身，如果某些名词或论述显得未能充分展开，建议读者进一步阅读相关书籍和综述获取更为具体的信息。

致谢

衷心感谢中国科学院植物研究所孔宏智研究员、中山大学生命科学学院施苏华教授和贺雄雷教授、南开大学计算机学院王刚教授和任明明博士对本文初稿的阅读和点评。衷心感谢两位匿名审稿专家提出建设性的修改意见。

参考文献

1. Brown, C.T., Hug, L.A., Thomas, B.C., Sharon, I., Castelle, C.J., Singh, A., Wilkins, M., Wrighton, K.C., Williams, K.H. and Banfield, J.F. (2015). Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature* 523: 208-211,
2. Bouckaert, R., Vaughan, T.G., Barido-Sottani, J., Duchêne, S., Fourment, M., *et al.* (2019). BEAST 2.5: An advanced software platform for Bayesian evolutionary analysis. *PLoS computational biology* 15: e1006650,
3. Bourlat, S.J., Juliusdottir, T., Lowe, C.J., Freeman, R., Aronowicz, J., et al. (2006). Deuterostome phylogeny reveals monophyletic chordates and the new phylum Xenoturbellida. *Nature* 444: 85-88,
4. Crotty, S.M., Minh, B.Q., Bean, N.G., Holland, B.R., Tuke, J., Jermiin, L.S. and Von Haeseler, A. (2020). GHOST: Recovering historical signal from heterotachously evolved sequence alignments. *Systematic Biology* 69: 249-264.
5. Drummond, A.J. and Rambaut, A. (2007). “BEAST: Bayesian evolutionary analysis by sampling trees.” *BMC Evolutionary Biology* 7: 214.
6. Eisen, J.A. and Fraser, C.M. (2003). Phylogenomics: intersection of evolution and

genomics. *Science* 300: 1706-1707.

7. Henig, W. (1966). *Phylogenetic Systematics*, pp. 1-27. Urbana: University of Illinois Press.
8. Kapli, P., Yang, Z. and Telford, M.J. (2020). Phylogenetic tree building in the genomic age. *Nature Reviews Genetics* 21: 428-444,
9. Klass, K.D., Zompro, O., Kristensen, N.P. and Adis, J. (2002). Mantophasmatodea: A new insect order with extant members in the afrotropics. *Science* 296: 1456-1459.
10. Lartillot, N. and Philippe, H. (2004). A Bayesian mixture model for across-site heterogeneities in the amino-acid replacement process. *Molecular Biology and Evolution* 21: 1095-1109,
11. Misof, B., Liu, S., Meusemann, K., Peters, R.S., Donath, A., *et al.* (2014). Phylogenomics resolves the timing and pattern of insect evolution. Science 346: 763-767,
12. Nguyen, L. T., Schmidt, H.A., von Haeseler. A. and Minh, B.Q. (2015). IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Molecular Biology and Evolution* 32: 268-274.
13. Spang, A., Saw, J.H., Jørgensen, S.L., Zaremba-Niedzwiedzka, K., Martijn, J., Lind, A.E., van Eijk, R., Schleper, C., Guy, L. and Ettema, T.J.G. (2015). Complex archaea that bridge the gap between prokaryotes and eukaryotes. *Nature* 521: 173-179,
14. Sundberg, K., Clement, M. and Snell, Q. (2010). On the use of cartographic projections in visualizing phylo-genetic tree space. *Algorithms for Molecular Biology* 5: 26,
15. Suvorov, A., Hochuli, J. and Schrider, D.R. (2020). Accurate inference of tree topologies from multiple sequence alignments using deep learning. *Systematic Biology* 69: 221-233.
16. Thomas, G.W.C., Dohmen, E., Hughes, D.S.T, Murali, S.C., Poelchau, M., *et al.* (2020). Gene content evolution in the arthropods. *Genome Biology* 21: 15.

17. Wang, Y., Engel, M.S., Rafael, J.A., Wu, H., Rédei, D., Xie, Q., Wang, G., Liu, X.G. and Bu, W. (2016). Fossil record of stem groups employed in evaluating the chronogram of insects (Arthropoda: Hexapoda). *Scientific Reports* 6: 38939.
18. Westblad, E. (1949). *Xenoturbella bocki* n.g, n.sp, a peculiar, primitive turbellarian type. *Arkiv Zoologi* 1: 3-29.
19. Xie, Q., Wang, Y., Lin, J., Qin, Y., Wang, Y., Bu, W. (2012). Potential key bases of ribosomal RNA to kingdom-specific spectra of antibiotic susceptibility and the possible archaeal origin of eukaryotes. *PLoS ONE* 7: e29468,
20. Wu, H.Y., Wang, Y.H., Xie, Q., Ke, Y. L. and Bu, W.J. (2016). Molecular classification based on apomorphic amino acids (Arthropoda, Hexapoda): Integrative taxonomy in the era of phylogenomics. *Scientific Reports* 6:28308,
21. Yang, Z. (2007). PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Molecular Biology and Evolution* 24: 1586-1591.
22. Zhang, C., Huelsenbeck, J. and Ronquist, F. (2020). Using parsimony-guided tree proposals to accelerate convergence in Bayesian phylogenetic inference. *Systematic Biology* 69: 1016-1032.

