生物多样性 2014, **22** (3): 277–284 Doi: 10.3724/SP.J.1003.2014.13267
Biodiversity Science http: //www.biodiversity-science.net

流程化的生态建模方法与科学工作流系统

乔慧捷 林聪田 王江宁 纪力强*

(中国科学院动物研究所, 北京 100101)

摘要: 科学工作流系统是由一系列经过特殊设计的数据分析与管理步骤组成的、按照一定的逻辑组织在一起, 并在给定的运行环境下, 完成特定科学研究的工作流管理系统。科学工作流系统致力于使全世界的科学家可以在一个简单易用的平台上交换思想, 共同设计全球尺度的实验, 共享数据、实验步骤与结果等。每一个科学家可以独立创建自己的工作流, 执行工作流并实时查看结果; 不同科学家之间也可以方便地共享和复用这些工作流。本文以开普勒系统(Kepler system)和生物多样性虚拟实验室(BioVeL)两个项目为例, 介绍了科学工作流的发展历史、背景、现有项目和应用等。以生态位模型工作流为例, 介绍了科学工作流的流程以及特点等。并通过对现有科学工作流的分析, 对其发展方向和存在的问题提出了自己的看法及预期。

关键词: 科学工作流, 生态建模, 生态位模型, 开普勒系统, 生物多样性虚拟实验室

Process-oriented ecological modeling approach and scientific workflow system

Huijie Qiao, Congtian Lin, Jiangning Wang, Liqiang Ji*

Institute of Zoology, Chinese Academy of Sciences, Beijing 100101

Abstract: A scientific workflow system is designed specifically to organize, manage and execute a series of research steps, or a workflow, in a given runtime environment. The vision for scientific workflow systems is that the scientists around the world can collaborate on designing global-scaled experiments, sharing the data sets, experimental processes, and results on an easy-to-use platform. Each scientist can create and execute their own workflows and view results in real-time, and then subsequently share and reuse workflows among other scientists. Two case studies, using the Kepler system and BioVeL, are introduced in this paper. Ecological niche modeling process, which is a specialized form of scientific workflow system included in both Kepler system and BioVeL, was used to describe and discuss the features, developmental trends, and problems of scientific workflows.

Key words: scientific workflow, ecological modeling, ecological niche model, Kepler system, BioVeL

1 背景简介

在生态学研究的过程中一直伴随着各种数学的分析方法和模型(Grinnell, 1917; Elton, 1927; Birch, 1953), 特别是最近20年, 数据采集手段的增加和计算能力的提高, 提升了生态学模型的数量和复杂度。很多新的理论, 如贝叶斯理论、机器学习等, 都被引入到生态建模中(Phillips *et al*., 2004; Patricia & Pamela, 2006; Olden *et al*., 2008; Dietterich, 2009)。而用于生态学模型算法实现的编程语言也各自不同, 如适用于分析生态位保守性与一致性关系的ENMTools使用Perl语言编写(Warren *et al*., 2010), 适用于生态位宽度分析以及生物分布地理分析的工具DIVA-GIS由Delphi语言编写(Hijmans *et al*., 2011), 适用于分析物种生存需求和预测潜在分布地的软件最大熵模型(Maximum En-

收稿日期: 2013-12-26; 接受日期: 2014-05-16
基金项目: 国家自然科学基金(31100390)
* 通讯作者 Author for correspondence. E-mail: ji@ioz.ac.cn

tropy Modeling, Maxent)由Java语言完成(Phillips *et al*., 2004, 2005), 两种集成了多个生态位模型的综合分析系统openModeller(Santana *et al*., 2006; Muñoz *et al*., 2011)和BIOMOD (Thuiller *et al*., 2009)分别由C++和R语言编写。这些软件的运行环境、部署方式各不相同。此外, 由于数据采集手段的增加, 各个领域建立了很多与生态学相关的数据库, 如全球生物多样性信息网络(Global Biodiversity Information Facility, GBIF)、i4LIFE、GenBank、WorldClim等, 它们各自有自己的数据标准, 并通过特定的格式提供对外的查询和下载服务。工具和数据的多样化, 增加了生态学模型的使用难度。此外, 随着气候变化和国际贸易、跨国交流等活动的增加所带来的全球性问题(如外来种入侵、全球性的传染疾病等), 要求生态学研究应着眼于全球尺度。然而由于科研水平的地区性差异, 先进的研究手段与方法很难在科技水平欠发达的地区开展。因此需要建置一个统一的研究平台, 使得不同研究水平的科学家可以在同样的条件下对同一类问题、在不同的地区分别研究, 以取得可进行横向比较的标准结果。科学工作流就是在这个背景下产生的。

2 科学工作流系统

通常, “假说验证”类型的科学研究需要经历确立假说(寻找科学问题)→实验设计→搜集数据→整理数据→分析数据→结果解释几个步骤。传统的做法一般是在野外工作之前设计好实验的步骤、需要搜集的数据以及搜集方法等; 然后有目的地去搜集和整理所需要的数据; 利用现有的工具分析实验数据; 得到结果后, 尝试解释实验结果并验证假说。一个假说需要很多次不同时间、地点及研究对象的验证才能被认可。因此, 这类科学研究工作中存在着很多重复内容。

所谓科学工作流系统, 是指将模块化的实验步骤按照一定的逻辑集中在一起, 这些模块包括数据库的存取和查询步骤, 数据分析和挖掘步骤, 以及其他一些在高性能计算集群上的高强度计算工作步骤(Taylor *et al*., 2006; Ludascher *et al*., 2009; McPhillips *et al*., 2009)。科学工作流软件的出现就是为了寻找生态学研究中可以重复的部分, 并加以整理和统一, 以期能够尽量简化和规范科学行为, 为科学家节省更多的精力(McPhillips & Bowers, 2005)。科学工作流系统作为生态学模型分析的工具和框架, 是生态建模工作中新兴的概念。采用该技术可以进行应用软件和模型的开发和集成工作。

3 现有的科学工作流系统

3.1 开普勒系统(Kepler system)

开普勒计划是由加州大学戴维斯分校、加州大学圣巴巴拉分校及加州大学圣地亚哥分校发起并维护的, 其产品开普勒系统采用Java编程语言开发, 可运行在Windows、Linux和Mac OS X等多种操作系统上, 用于生物学工作流开发与共享(Pennington & Michener, 2005; Ludäscher *et al*., 2006)。

开普勒系统构建于另外一个开源建模系统——托勒密II(Ptolemy II)基础上, 为科学家提供了一个方便易用的工作平台。该工作流系统是一个用户友好的程序, 允许科学家通过简单地拖拉操作来连接一些特定的组件, 建立满足条件的科学工作流。用户即使没有计算机科学背景, 也可以使用标准组件来生成工作流, 或者修改现有的工作流模型以满足需要。与此同时, 系统还可以调用其他数据分析软件中已有的分析方法, 如Matlab、R等, 并提供了常用的数据库调用接口和数据格式转换模块。而且, 开普勒系统可以使用基于网格的分布式计算方法来执行这些工作流, 以达到有效利用计算资源的目的。此外, 某位科学家创建的工作流可以保存并以文件的形式交换、发送给其他科学家重复使用。

目前, 开普勒系统已经广泛地应用在多个科学与科普研究项目中, 如系统发育过程的计算和模拟的pPOD(Bowers *et al*., 2008), 实时环境分析处理的REAP(Barseghian *et al*., 2010), 管理野生动物种群变化的SANParks(Swemmer & Taljaard, 2011), 以及依托于中国科学院网络中心e-Science项目的SEEK(生态知识科学环境)等。

3.2 生物多样性虚拟实验室(BioVeL)项目

BioVeL是由欧盟第七框架计划资助、通过网络技术实现海量数据与跨学科数据分析的虚拟生物学实验室。在BioVeL中, 有2个主要的组成部分: 服务(services)和工作流(workflow)。“服务”指的是存在于BioVeL服务器上, 为用户提供各种数据与通讯接口的功能, 包括数据下载服务、功能模块接口以及工作流管理服务等。用户通过这些服务可以下载多个数据源的数据, 转换数据格式, 进行数据质量检

查, 调用不同的分析模块, 组合分析模块形成工作流, 利用工作流分析数据和模型结果, 保存及共享工作流等多项工作。可以说, 服务是BioVeL项目的核心组件, 一切的活动均以各种服务为基础。而“工作流”则是在BioVeL提供的服务基础上, 通过BioVeL门户网站运行的科学研究工作流程的抽象化概念。与开普勒项目不同, BioVeL项目不允许用户自主创建工作流, 而只能在BioVeL提供的工作流模块基础上, 修改、管理、共享和保存工作流的结果。通过BioVeL的门户网站, 用户可以交互式的检测和管理运行中的工作流, 修改工作流的参数。此外, BioVeL还可对工作流的数据分析结果(模型运行结果)进行辅助分析。BioVeL的工作流程如图1所示。

目前BioVeL中已经搜集、整理并建立了包括Biome-BGC生态系统指标监管工作流(Biome-BGC Ecosystem Service Indicators Regulation Workflow)、数据整理工作流(Data Refinement Workflow, DRW)、生态位模型工作流(Ecological Niche Model Workflows, ENMW)、宏基因组特征统计分析工作流(Metagenomic Traits Statistical Analysis Workflow)、系统发育工作流(Phylogenetic Workflows)和种群模型工作流(Population Modeling Workflows)等多个生物学不同领域的工作流供科学家使用。其中部分工作流尚处于试验阶段, 下面重点介绍已经得到应用的三个工作流系统。

Biome-BGC生态系统指标监管工作流: Biome-BGC是由蒙大拿大学林学院数字地球动态模拟研究组(Numerical Terradynamic Simulations Group, NTSG)开发的用于模拟水体与不同陆地生态系统的碳、氮等物质循环通量的模型。Biome-BGC生态系统指标监管工作流通过模拟单一情况下的碳或氮通量, 量化其模型结果, 并由此派生出重要的生态系统监控指标。

数据整理工作流: 主要用于搜集和整理多个来源的包括物种分类、标本、观测等用于科学分析的生物相关数据。这些数据可用于后续的科学分析, 如物种分布、物种丰富度和多样性、物种发生历史等时空分析。

生态位模型工作流包括生态位模型工作流和数据统计工作流(ENM Statistical Workflow, ESW)。前者基于openModeller提供建模服务。使用者可以

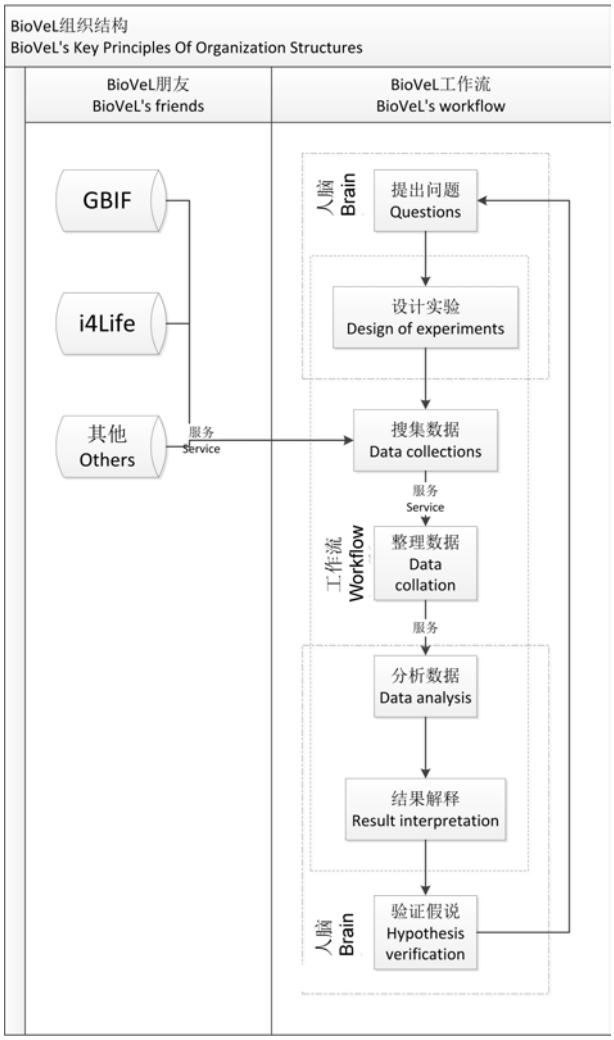

图1　BioVeL的组织结构及工作原理

Fig. 1　The organization design and operating principles in BioVeL

通过BioVeL的门户网站提供物种发生数据和物种分布相关的地理图层以及建模区域, 选择相应的模型并设定模型参数。该工作流可自动进行迭代建模并选择最优的模型结果, 并将该结果应用到另外的场景(如其他研究区域或其他年代)。在得到模型结果后, 用户可以通过ESW进行结果分析。ESW建立在R语言基础上, 包括了模型结果的交叉验证、受试者反应曲线(receiver operating characteristic curve, ROC)下的面积(area under curve, AUC)、模型的Kappa值、真实技巧统计值(true skill statistic, TSS), 以及模型的特异性与敏感性等多个评价指标。用户可以通过这些指标对模型进行分析和比较以得到最优的结果。此外, ESW还可以比较两个不同场景

结果的差异(如现在和未来气候场景下的差异等)。

3.3 两个科学工作流系统的比较

由于现存的两个科学工作流系统设计理念、出现时间、发展思路以及服务的对象均有所不同, 因此其应用程度也不相同。我们从运行方式、使用方便程度、现有工作流和用户数量等几个方面分别对两个系统进行比较。通过表1可以看出, 开普勒系统更加关注于完善的功能、灵活的可定制性和丰富的用户交互体验, 在工作流的数量、复杂程度上均高于BioVeL。但由此带来的部署困难和较高的使用门槛, 阻止了部分对计算机使用不熟悉的用户。而BioVeL的用户可通过在线方式管理和使用工作流, 成功地规避了开普勒系统安装部署难的缺陷, 但却也由此带来功能单一, 操作不便等缺陷。

总之, 现有的科学工作流都存在一些问题。如何在功能强大与使用方便之间寻找一个平衡点, 将是科学工作流系统今后发展的主要目标。

4 工作流软件的发展现状与前景分析: 以生态位模型为例

生态位模型以已知样本点(如野外调查或标本记录等)为基础, 分析物种在生态位空间(niche space)或环境空间(environmental space)中的特征, 进而研究物种的环境耐受能力(species' environmental tolerances)。该类模型中使用了环境数据和地理信息数据, 并且涉及到了多种数学分析方法, 其工作流程的复杂性和数据的多样性均符合工作流软件的基本要求。因此, 在现有的各个工作流软件中, 均以生态位模型为最基本的原型研究, 并提供了例子。本文以开普勒系统和BioVeL中生态位模型工作流的应用为例, 简单地介绍工作流软件的发展现状。

4.1 生态位建模的方法介绍

生态位的保守性(Peterson *et al*., 1999; Wiens & Graham, 2005; Peterson & Ammann, 2013)是生态位模型研究领域的基本理论之一, 也是近期讨论最多的热点问题。"生态位保守主义"概念由Townsend Peterson在1999年的*Science*中提出(Peterson *et al*., 1999), 是指物种的基础生态位在时间尺度上呈现出大尺度上缓慢变化、小尺度上保持稳定的"保守主义"的特征(Peterson *et al*., 1999; Hadly *et al*., 2009; Soberón & Peterson, 2011)。这一概念直接催生了物种潜在分布地预测工作(乔慧捷等, 2013; 朱耿平等, 2013)。包括生物气候包络(Bioclimatic Envelope Algorithm, BIOCLIM)(Busby, 1991; Walker & Cocks, 1991; Mbogga *et al*., 2010)、生态位因子分析(Ecological Niche Factor Analysis, ENFA)(Hirzel *et al*., 2002)、最大熵模型(Phillips *et al*., 2004, 2006; Phillips & Dudík, 2008)、遗传算法(Genetic Algorithm for Rule-set Production, GARP)(Stockwell, 1999)等, 经典或流行的物种分布模型(Species Distribution Model, SDM, 或称为生态位模型都是基于这一理论完成的。生态位的保守主义是SDM的理论基础。也正是基于物种在小时间尺度上的保守性, 才有了SDM在包括物种栖息地保护与保护区规划(Graham *et al*., 2004; Franklin, 2010), 外来入侵物种防控(Thuiller *et al*., 2005; Ebeling *et al*., 2008; Václavík & Meentemeyer, 2009), 疾病的传播模式(Peterson *et al*., 2002; Peterson, 2006; Costa & Peterson, 2012), 以及物种随时间的分布趋势变化等(Franklin *et al*., 2013)多个方面的应用。另一方面, 针对生态位的保守性与一致性的讨论(Warren *et al*.,

表1 现有的两个科学工作流系统的特征比较
Table 1 Comparison of the two scientific workflow systems

	开普勒系统 Kepler system	BioVeL
运行方式 Operating mode	单机运行 Stand-alone	在线运行 Online
是否可自由组合 Can be combined freely?	是 Yes	否 No
是否可重复使用 Can be reused?	是 Yes	是 Yes
复杂程度 Complexity	复杂, 多变, 可自由组合 Complex, varied, can be combined	复杂与否与提供的服务相关 Associated with the provided services
共享方式 Way to share	通过文件 Via files	在线服务 Online services
使用方便程度 Usability	复杂 Complex	简单 Simple
已有的数量 Number of instances	丰富 Abundance	有限 Limited
用户数量 Number of users	丰富 Abundance	测试阶段, 用户数量未知 Unknown

2008)促进了如ENMTools(Warren *et al*., 2010)、SDMTools (van Derwal *et al*., 2011)等针对SDM结果进行比较性研究的工具的发展, 这部分工具也逐渐扩展到生物地理学、进化生物学等多个领域(Liu *et al*., 2013)。

伴随SDM研究的进展, 研究者先后开发出近20种不同的模型。按照所基于的数学理论, 大体可分为以下3种类型(Rangel & Loyola, 2012): 第一类是以BIOCLIM、ENFA为代表的简单包络模型; 第二类是以数理统计、概率为理论基础的统计学模型, 如广义线性模型(Generalized Linear Model, GLM) (Guisan *et al*., 2002)、广义可加模型(Generalized Additive Models, GAM)、多变量自适应回归条样模型(Multivariate Adaptive Regression Splines, MARS) (Friedman, 1991)等, 这是一类更加复杂的包络模型; 最后一类是基于复杂的统计机器学习方法的机器学习模型, 如Maxent、GARP等。

4.2　工作流系统中的生态位建模方法

生态位模型工作流是开普勒系统中提供的第一批工作流之一。开普勒系统中的生态位模型工作流可通过EcoGrid直接获取建模相关的环境变量、物种分布数据等所需样本信息, 并可通过层级化的组件对获取到的数据进行逐层分析, 完成数据准备工作, 为后续的建模提供高质量的数据。在开普勒系统中提供了生态位建模方法并提供了标准化的数据分析方法。建模后, 开普勒系统还允许用户使用ROC曲线计算AUC值, 从而评价模型的准确率。可以说, 在开普勒系统中提供的生态位建模方法集成了数据下载、质量检查、建模与结果分析等四个主要的要素, 是一个较成熟的生态位工作流。图2展示了用开普勒系统进行生态位建模的流程。用户通过开普勒系统提供的标准工作流, 设置每一个模块的参数后, 点击红色按钮, 即完成生态位建模工作。用户还可以通过调整不同模块的参数, 对模型进行微调, 以达到更好的建模效果。

BioVeL也提供了生态位模型工作流, 该工作流由DRW、ENM和ESW组成。

DRW工作流主要用于数据筛选与质量检查。用户可以物种学名为关键词, 在用户的标本数据库中下载公开的分布数据, 并与自己的分布数据整合, 显示在以Google地图为底图的网页中。用户也可以利用Google Refine工具对上述数据进行整理, 以保

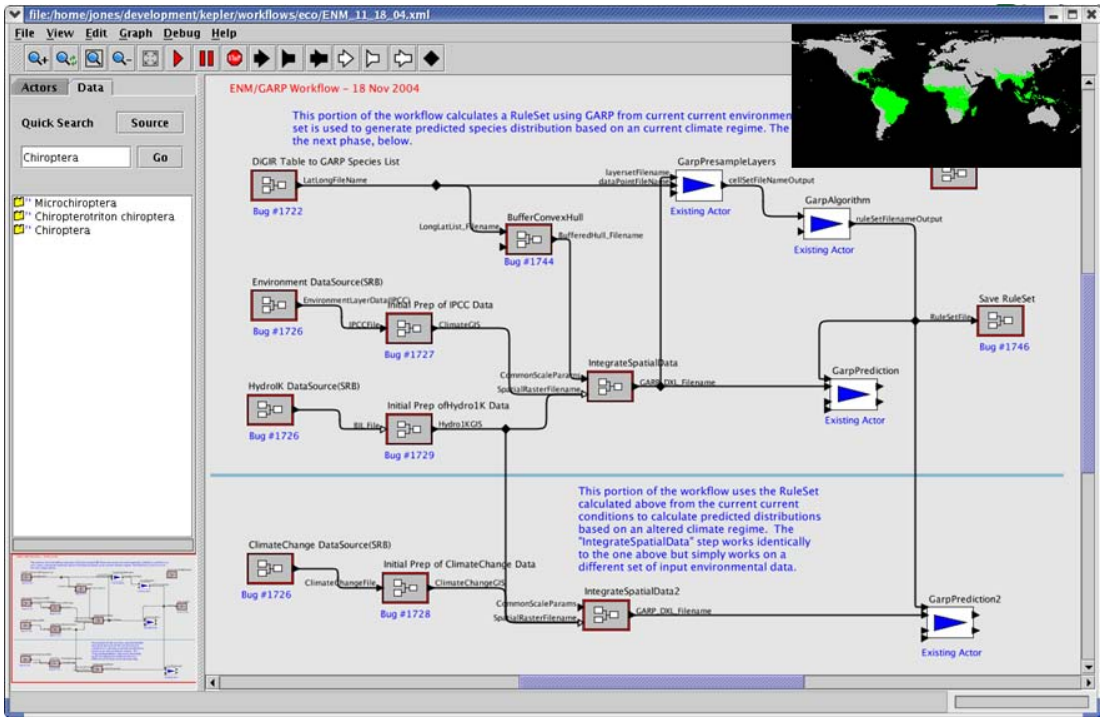

图2　开普勒系统中的生态位模型工作流

Fig. 2　Ecological Niche Modeling Workflow in Kepler system

证数据的准确性。并可按照物种、类群或用户自定义的标准对数据进行筛选，为后续的生态位建模做准备。此外，DRW还可以用于物种拉丁名的拼写错误检查、异名搜索等多个方面的数据质量检查。

数据质量检查后，用户可以选择需要建模的算法，并指定算法的运行参数。ENM工作流中共有包括BIOCLIM、GARP、ENFA等在内的15个常用的生态位模型。用户可以同时选择多个数据集，在多个算法下进行建模。在用户提交建模任务后，所有工作将在后台排队进行，用户可以关闭浏览器甚至关闭计算机，任务仍然在BioVeL的服务器运行直至完成。

任务完成后，用户可登录BioVeL的门户网站查询和下载模型运行结果，同时也可以调用ESW模块进行结果分析。用户可以选择交叉验证、ROC曲线的AUC值、Kappa值、TSS值以及特异性值、敏感性值等多个指标来评价模型，以得到最优的结果。此外，用户还可以利用ESW模块比较两个不同场景结果的差异，以说明气候的变化对物种分布的影响等。

BioVeL项目目前尚处于内部测试阶段，提供的服务无法被公众使用，在此无法提供更加详细的使用方法的介绍，但BioVeL作为一个新兴的概念和活跃的项目，是值得我们关注和效仿的。

5 工作流软件发展中存在的问题与发展方向

科学工作流旨在简化并规范化科学研究过程中的每一个步骤，以更好地规范科学研究流程及重复研究过程，并且在统一的标准下比较科学研究结果。然而“科学软件中的盲目信任”问题是目前的科学研究，特别是生态学研究中普遍存在的一个问题(Joppa *et al*., 2013)。工作流在为研究者节约很多时间和精力的同时，也会禁锢科学研究的自由思维，从而限制这个研究领域的创新和发展。因此，如何在模型的方便性和自由性之间找到一个平衡点，是工作流系统设计需要注意的首要问题。

近些年来，我国在生物多样性数据库建设方面有了长足的进步，建设了以NSII系统为首的一批生物多样性数据库(许哲平等, 2012)，并且逐步建立了如生物多样性数据库纵向搜索引擎、多数据库联合搜索、生物物种名录管理系统等针对物种数据的检索和管理工具(王利松等，2010；许哲平等，2012)，还建立了如物种分类树比较工具(Lin，2013)、生物图像识别系统(Wang *et al*., 2012a, b)等一批数据分析与挖掘工具。然而，这些工具尚建立在传统的软件架构下，还没有在工作流的概念下实现模块化管理、动态链接和功能组合。而在国际上，尽管工作流的概念已经提出很多年，但实际应用范围并不广。因此，建立一个针对国内生物多样性数据库和研究工作特点的科学研究工作流开发规范和模式项目，是一件有意义的工作，并有机会在短时间内达到国际领先水平。

建立工作流的基础是已有的生态理论和算法。因此，在建立模式工作流之前，需要寻找到合适的切入点。而根据软件开发的“二八原则(80%的用户只使用软件的20%的功能)”，确定工作流系统重点模块是建立工作流的首要任务。此外，建立工作流系统是一个由生态学家、软件开发人员、数据库管理人员以及普通的使用者共同完成的工作。因此，如何在选定的专业领域内，组织一批由专家、软件工程师以及相关工作人员组成的工作团队是建立工作流系统的保障。最后，一个完整的工作流是由数据提供者、方法提供者以及结果分析系统共同组成的，任何一个环节若不能提供完整的支持，都将导致工作流适用范围的单一化或结果出现偏差。因此，如何消除各个环节中的壁垒，达到科学工作流系统高效运行，是建立一个成功的科学研究工作流系统的关键。

综上所述，科学工作流系统尽管存在着屏蔽了算法细节，容易导致科学研究的盲目性等缺点，但作为一项新兴的辅助工具，其在科学研究流程的简单化与规范化，数据与分析结果的标准化等方面都发挥了很大的作用。另外，这方面的研究仍处于起步状态，是一个大家都在关注的热门研究领域，因此，具有重大发展潜力。

参考文献

Barseghian D, Altintas I, Jones MB, Crawl D, Potter N, Gallagher J, Cornillon P, Schildhauer M, Borer ET, Seabloom EW, Hosseini PR (2010) Workflows and extensions to the Kepler scientific workflow system to support environmental sensor data access and analysis. *Ecological Informatics*, **5**, 42–50.

Birch LC (1953) Experimental background to the study of the distribution and abundance of insects. I. The influence of temperature, moisture and food on the innate capacity for increase of three grain beetles. *Ecology*, **34**, 698–711.

Bowers S, Timothy M, Sean R, Manish A, Bertram L (2008)

Kepler/pPOD: scientific workflow and provenance support for assembling the tree of life. In: *International Provenance and Annotation Workshop* (eds Freire J, Koop D, Moreau L), pp. 70–77. Springer, Berlin, Heidelberg.

Busby JR (1991) BIOCLIM—a bioclimate analysis and prediction system. *Plant Protection Quarterly*, **6**, 8–9.

Costa J, Peterson AT (2012) Ecological niche modeling as a tool for understanding distributions and interactions of vectors, hosts, and etiologic agents of Chagas disease. *Advances in Experimental Medicine and Biology*, **710**, 59–70.

Dietterich TG (2009) Machine learning and ecosystem informatics: challenges and opportunities. In: *Advances in Machine Learning* (eds Zhou ZH, Washio T), pp. 1–5. Springer, Berlin, Heidelberg.

Ebeling SK, Welk E, Auge H, Bruelheide H (2008) Predicting the spread of an invasive plant: combining experiments and ecological niche model. *Ecography*, **31**, 709–719.

Elton CS (1927) *Animal Ecology*. University of Chicago Press, Chicago.

Franklin J (2010) Moving beyond static species distribution models in support of conservation biogeography. *Diversity and Distributions*, **16**, 321–330.

Franklin J, Davis FW, Ikegami M, Syphard AD, Flint LE, Flint AL, Hannah L (2013) Modeling plant species distributions under future climates: How fine scale do climate projections need to be? *Global Change Biology*, **19**, 473–483.

Friedman JH (1991) Multivariate adaptive regression splines. *The Annals of Statistics*, **19**, 1–67.

Graham CH, Ferrier S, Huettman F, Moritz C, Peterson AT (2004) New developments in museum-based informatics and applications in biodiversity analysis. *Trends in Ecology and Evolution*, **19**, 497–503.

Grinnell J (1917) Field tests of theories concerning distributional control. *The American Naturalist*, **51**, 115–128.

Guisan A, Edwards TC, Hastie T (2002) Generalized linear and generalized additive models in studies of species distributions: setting the scene. *Ecological Modelling*, **157**, 89–100.

Hadly EA, Spaeth PA, Li C (2009) Niche conservatism above the species level. *Proceedings of the National Academy of Sciences, USA*, **106**, 19707–19714.

Hijmans R, Guarino L, Mathur P, Jarvis A (2011) *DIVA-GIS: Geographic Information System for Biodiversity Research.* http://www.diva-gis.org/. (2014-03-20)

Hirzel AH, Hausser J, Chessel D, Perrin N (2002) Ecological niche factor analysis: How to compute habitat-suitability maps without absence data. *Ecology*, **83**, 2027–2036.

Joppa LN, McInerny G, Harper R, Salido L, Takeda K, O'Hara K, Gavaghan D, Emmott S (2013) Troubling trends in scientific software use. *Science*, **340**, 814–815.

Lin C (2013) *Taxonomic Tree Tool*. http://ttt.biodinfo.org/indexen.asp. (2013-12-05)

Liu J, Möller M, Provan J, Gao LM, Poudel RC, Li DZ (2013) Geological and ecological factors drive cryptic speciation of yews in a biodiversity hotspot. *The New Phytologist*, **199**, 1093–1108.

Ludäscher B, Altintas I, Berkley C, Higgins D, Jaeger E, Jones M, Lee EA, Tao J, Zhao Y (2006) Scientific workflow management and the Kepler system. *Concurrency and Computation: Practice and Experience*, **18**, 1039–1065.

Ludascher B, Altintas I, Bowers S, Cummings J, Critchlow T, Deelman E, DeRoure D, Freire J, Goble C, Jones M (2009) Scientific process automation and workflow management. In: *Scientific Data Management: Challenges, Technology, and Deployment* (eds Shoshani A, Rotem D), pp. 467–508. Chapman and Hall, London.

Mbogga MS, Wang XL, Hamann A (2010) Bioclimate envelope model predictions for natural resource management: dealing with uncertainty. *Journal of Applied Ecology*, **47**, 731–740.

McPhillips TM, Bowers S (2005) An approach for pipelining nested collections in scientific workflows. *ACM SIGMOD Record*, **34**, 12–17.

McPhillips T, Bowers S, Zinn D, Ludaescher B (2009) Scientific workflow design for mere mortals. *Future Generation Computer Systems*, **25**, 541–551.

Muñoz MES, Giovanni R, Siqueira MF, Sutton T, Brewer P, Pereira RS, Canhos DAL, Canhos VP (2011) OpenModeller: a generic approach to species' potential distribution modelling. *GeoInformatica*, **15**, 111–135.

Olden JD, Lawler JJ, Poff NL (2008) Machine learning methods without tears: a primer for ecologists. *The Quarterly Review of Biology*, **83**, 171–193.

Patricia DH, Pamela JW (2006) Modern machine learning for automatic optimization algorithm selection. In: *Proceedings of the INFORMS Artificial Intelligence and Data Mining Workshop*. Citeseer.

Pennington DD, Michener WK (2005) The EcoGrid and the Kepler workflow system: a new platform for conducting ecological analyses. *Bulletin of the Ecological Society of America*, **86**, 169–176.

Peterson AT, Sánchez-Cordero V, Beard C, Ramsey J (2002) Ecologic niche modeling and potential reservoirs for Chagas disease, Mexico. *Emerging Infectious Diseases*, **8**, 662–667.

Peterson AT, Soberón J, Sánchez-Cordero V (1999) Conservatism of ecological niches in evolutionary time. *Science*, **285**, 1265–1267.

Peterson AT (2006) Ecological niche modeling and spatial patterns of disease transmission. *Emerging Infectious Diseases*, **12**, 1822–1826.

Peterson AT, Ammann CM (2013) Global patterns of connectivity and isolation of populations of forest bird species in the late Pleistocene. *Global Ecology and Biogeography*, **22**, 596–606.

Phillips SJ, Dudík M, Schapire RE (2005) Maxent software for species distribution modeling. http://www.cs.princeton.edu/schapire/maxent. (2014-03-21)

Phillips SJ, Anderson RP, Schapire RE (2006) Maximum entropy modeling of species geographic distributions. *Ecological Modelling*, **190**, 231–259.

Phillips SJ, Dudík M (2008) Modeling of species distributions

with Maxent: new extensions and a comprehensive evaluation. *Ecography*, **31**, 161–175.

Phillips SJ, Dudík M, Schapire RE (2004) A maximum entropy approach to species distribution modeling. In: *Proceedings of the Twenty-first International Conference on Machine Learning*, p. 83. ACM. Banff, Alberta, Canada.

Qiao HJ (乔慧捷), Hu JH (胡军华), Huang JH (黄继红) (2013) Theoretical basis, future directions, and challenges for ecological niche models. *Science China: Life Sciences* (中国科学: 生命科学), **43**, 915–927. (in Chinese with English abstract)

Rangel TF, Loyola RD (2012) Labeling ecological niche models. *Natureza & Conservacao*, **10**, 119–126.

Santana FS, Fonseca RR, Saraiva AM, Corrêa PLP, Bravo C, Giovanni R (2006) OpenModeller—an open framework for ecological niche modeling: analysis and future improvements. In: *Proceedings of the World Conference on Computers in Agriculture and Natural Resources*. Orlando, Florida, USA.

Soberón J, Peterson AT (2011) Ecological niche shifts and environmental space anisotropy: a cautionary note. *Revista Mexicana de Biodiversidad*, **82**, 1348–1355.

Stockwell D (1999) The GARP modelling system: problems and solutions to automated spatial prediction. *International Journal of Geographical Information Science*, **13**, 143–158.

Swemmer LK, Taljaard S (2011) SANParks, people and adaptive management: understanding a diverse field of practice during changing times. *Koedoe*, **53**, 199–205.

Taylor I, Deelman E, Gannon D (2006) *Workflows for e-Science: Scientific Workflows for Grids*. Springer, Berlin.

Thuiller W, Lafourcade B, Engler R, Araújo MB (2009) BIOMOD—a platform for ensemble forecasting of species distributions. *Ecography*, **32**, 369–373.

Thuiller W, Richardson DM, Pysek P, Midgley GF, Hughes GO, Rouget M (2005) Niche-based modelling as a tool for predicting the risk of alien plant invasions at a global scale. *Global Change Biology*, **11**, 2234–2250.

Václavík T, Meentemeyer RK (2009) Invasive species distribution modeling (iSDM): Are absence data and dispersal constraints needed to predict actual distributions? *Ecological Modelling*, **220**, 3248–3258.

van Derwal J, Falconi L, Januchowski S, Shoo L, Storlie C (2011) SDMTools—Species distribution modelling tools: tools for processing data associated with species distribution modelling exercises. http://www.rforge.net/SDMTools/. 2014-04-20.

Walker PA, Cocks KD (1991) HABITAT: a procedure for modelling a disjoint environmental envelope for a plant or animal species. *Global Ecology and Biogeography Letters*, **1**, 108–118.

Wang JN, Ji LQ, Liang AP, Yuan DC (2012a) The identification of butterfly families using content-based image retrieval. *Biosystems Engineering*, **111**, 24–32.

Wang JN, Lin CT, Ji LQ, Liang AP (2012b) A new automatic identification system of insect images at the order level. *Knowledge-based Systems*, **33**, 102–110.

Wang LS (王利松), Chen B (陈彬), Ji LQ (纪力强), Ma KP (马克平) (2010) Progress in biodiversity informatics. *Biodiversity Science* (生物多样性), **18**, 429–443. (in Chinese with English abstract)

Warren DL, Glor RE, Turelli M (2008) Environmental niche equivalency versus conservatism: quantitative approaches to niche evolution. *Evolution*, **62**, 2868–2883.

Warren DL, Glor RE, Turelli M (2010) ENMTools: a toolbox for comparative studies of environmental niche models. *Ecography*, **33**, 607–611.

Wiens JJ, Graham CH (2005) Niche conservatism: integrating evolution, ecology, and conservation biology. *Annual Review of Ecology, Evolution, and Systematics*, **36**, 519–539.

Xu ZP (许哲平), Qin HN (覃海宁), Ma KP (马克平), Bao BJ (包伯坚), Li Y (李奕), Zhao LN (赵莉娜) (2012) Research on management, sharing and application of natural science and technology resources: taking Chinese Virtual Herbarium (CVH) for an example. *China Science & Technology Resources Review* (中国科技资源导刊), **44**, 27–33. (in Chinese with English abstract)

Zhu GP (朱耿平), Liu GQ (刘国卿), Bu WJ (卜文俊), Gao YB (高玉葆) (2013) Ecological niche modeling and its applications in biodiversity conservation. *Biodiversity Science* (生物多样性), **21**, 90–98. (in Chinese with English abstract)

(责任编委: 马克平　责任编辑: 闫文杰)