

人类新基因电子克隆的自动化软件系统建立和实验验证与功能研究

Development of an Automatic Software System for in Silico Cloning of Novel human Genes and Their Experimental Validation and Functional Research

张德礼¹ 马大龙¹ 钱敏平²

(北京大学人类疾病基因研究中心、国家人类基因组北方研究中心¹ 北京 100083; 北京大学数学学院、北京大学计算生物学中心² 北京 100871)

中国博士后科学基金资助项目《人类新基因的电子克隆、实验确认与功能研究》获得的结果与结论如下: 电子克隆到 100 个人类新基因, 具有典型基因特点和一定功能结构域, 其中 23 个基因被 GenBank 接受并被国际人类基因组组织 (HUGO) 基因命名委员会批准; 首批选择 10 个基因用 RT-PCR 和 cDNA 测序验证, 结果证明均是真实基因; 这些基因的进一步确认和体内外功能鉴定正在实施中; 采用生物信息学分析与实验确认相结合的技术路线, 用我们克隆的基因在非冗余数据库对比发现, 网上发布的计算机注释人类基因组编码序列存在各种类型的多处错误。我们采用生物信息学方法电子克隆的策略既有助于发现更多的人类新基因, 又有助于纠正美国国家生物技术信息中心 (NCBI) 基因组注释项目公布的参考序列 (REFSEQs) 中所存在的错误。本研究已获第七届国际人类基因组大会 (HGM2002) 论文竞赛 Poster 奖 (参赛论文 680 篇, 获奖论文仅 5 篇)。这是我国首次在国际人类基因组组织 (HUGO) 大会上获奖。在此基础上, 我们认为应当在我国进一步深入开展“人类新基因电子克隆的自动化软件系统建立和实验验证与功能研究”, 并希望能得到国家有关部门的大力支持。

一、采用生物信息学方法电子克隆人类新基因的重要意义和作用

新基因搜寻和蛋白功能分析是人类基因组图谱公诸于世后的热点研究领域, 世界范围内新基因资源的争夺或掠夺已达到白热化程度, 具有自主知识产权的基因医学产业已成为各国追求的目标。电子克隆就是以数学算法为核心, 以计算机和互联网为工具, 发掘人类基因组的未知功能的新基因。研究电子克隆的算法和原则, 并对由电子克隆获得的候选新基因, 通过生物学实验进行序列和功能验

证, 将为人类功能基因组学与蛋白质组学研究提供新的线索和基础。采用生物信息学方法分析人类新基因, 并与实验生物学相结合, 可发现重要的人类新的功能基因, 取得源头创新的成果, 具有重要的理论意义和实用价值。

国际上, 基因发现研究有 3 类基本方法: (1) 基于碱基成分、特殊信号等的统计方法 (例如 Z 曲线法, Codon usage, Amino acid usage, Position asymmetry 等); (2) 基于基因结构的方法 (其中最有效的是隐马氏模型, 即 HMM 方法, 著名的 Genescan 软件的关键算法就是 HMM); (3) 相似性同源分析方法 (例如利用一个物种的已知基因在另一种物种中进行同源比对的 GLASS 和 Rosatta Stone 算法等)。对原核生物基因组和部分简单的真核生物基因组国际上已有的外显子预测软件, 最高敏感性和特异性均可达到 95.0% ~ 99.8%; 而对于真核生物基因预测, 国际上已有的软件, 按外显子计算敏感性和特异性最高可达 81% ~ 82%。但是对第一外显子, 最好算法的准确性不足 50%; 而对小分子肽类基因则基本上预测不到。可见, 目前国内外已有的基因发现软件要付诸实用, 尚有待于大大改善。

目前, 国内尚缺少完全采用生物信息学方法电子克隆新基因, 特别是小分子肽类基因的研究和先例。因此, 发展我们自己的具有更好性能的基因发现软件是电子克隆的关键。两年来, 北京大学人类疾病基因研究中心已开展了相关工作, 有 20 多个新基因在进行功能实验, 近 1 年来, 还与北京大学数学学院实验室合作, 共同开发半自动化电子克隆新基因软件的工作, 且取得了一定进展。

采用 EST 数据库发现人类新基因已被证明是电子克隆新基因的一条有效途径, 目前已有 5 个这样发现的人类新基因进入人体临床试验阶段, 有相当数量的这种电子克隆新基因被实验所证实。在人类基因组工作框架图公布之后, 用 EST 拼接校对的

新基因一般可通过框架图验证,初步确定内含子序列并定位于人类染色体上的适当部位,这就更有助于基于 EST 的人类新基因电子克隆。特别是人类基因组图谱公布之后,用 EST 拼接校对的新基因大多(90%)可通过图谱验证,确定内含子序列并定位于人类染色体上,这就为基于 EST 的人类新基因电子克隆提供了更高的准确性。

虽然 1 年前公布的人类基因组图谱的初步分析结果显示,人类拥有 2.6~ 3.9 万个基因,但欧美许多研究者对此还有疑问。Wiemann S 等(2001)采用 EST 数据库拼接出 500 个含有完整开放读码框架(ORF)的人类新 cDNA,通过与已经完全解码的人类 21 和 22 号染色体基因组序列对比,识别出许多完全漏报或误报的人类基因。我们在人类新基因的电子克隆和实验克隆工作中均遇到了类似情况,因此,人类全长 cDNA 测序仍将是精确识别基因的关键。虽然通过人类基因组图谱,结合人类基因进化树、蛋白质功能结构域知识,已初步确定人类约有 2.6~ 3.9 万个基因,但研究者也承认有错误和漏报。该数量是基于超级计算机排序分析的结果,很大一部分取决于研究人员如何向计算机输入原始信息,况且人类对生命活动本质的认识还相当肤浅,现有基因识别软件的准确性最高不足 50%,多软件联合使用预测的准确性最高也不超过 70%,蛋白质功能结构域预测软件的水平更是处于婴幼儿期,并且所有软件均是根据已有知识预测,一般小分子肽类基因根本预测不到,这就为在图谱发布后和精细图谱公布前的相当长一段时期内,通过呈指数增长的海量 EST 数据库挖掘人类新基因提供了更大可能。

正是基于上述这些依据及当时我们实验室已证实却未被精细图包括的几个人类新基因,我们申请了类似课题,获得中国博士后科学基金的资助,现已较好地完成各项技术指标。我们的研究表明,应慎重对待计算机注释的人类基因组编码序列中可能存在的各种类型错误,人类新基因的正确识别和注释仍是一项长期而繁重的任务,这方面的任何研究成果亦都将有助于其它真核生物新基因的正确识别和注释。两年来,我们在人类新基因电子克隆中,从事了大量精细的手工工作,积累了丰富的实践经验,并设计了一些实用软件,以求建立人类新基因电子克隆的全自动化软件系统并不断完善。我们在所参加的多种相关会议中了解到,单凭任何软件几乎都不能克隆到可靠的人类新基因,但我们通过国外不可能去做的劳动密集型辛劳工作却做到了这一点,已有自己的电子克隆标准和技术,组装成全自动化软件系统的时机已经基本成熟。

人类新基因克隆和人类基因组注释是既有联系又有区别的不同研究领域。在进行人类新基因电子克隆和实验确认过程中,特别是在人类新基因电

子克隆的自动化软件系统建立和实验验证过程中,必然会对人类基因组注释有所贡献,因为作为人类基因组的蛋白质编码序列和编码蛋白质功能注释既是人类基因组注释的核心内容,也是人类新基因克隆和验证的主要内容。

该方面的研究是创新性前沿研究,将在以下几个方面产生积极的作用和影响。

1. 目前,虽然国外已经有了可用于基因注释的拼接、对比、搜索软件和基因预测软件,但是它们大都只能用于大中型计算机,精确度尚有很大问题。对这些软件进行的改进和研发,将加快新基因的鉴定过程。

2. 迄今为止,人类基因组中能标注其功能的蛋白质只占 50% 左右,对基因功能及所参与的生物过程的注释还很有限。增加对启动子和蛋白质结构域的注释,对发现新基因、探索基因的功能以及了解所参与的细胞过程大有益处。

3. 可以使许多从事基因及基因组研究的技术人员,能够在自己的实验室中,对感兴趣的几种生物的全基因进行搜索、查询。利用该技术平台,研究人员可以对组织特异性表达的、含某一结构域(Domain)的、参与某一生物过程的有关基因和蛋白质序列进行搜索、查询和比对;还可以对一些未知功能的 DNA 片段进行注释以确定其身份(Identity),高精度地发现 ORF、外显子、内含子、启动子、转录因子结合位点;也可为基因组注释提供标准的语言和规范化的数据处理和可视化软件。这套系统亦必将成为我国功能基因组研究人员的重要工具。

4. 目前理论与计算生物学部门所开发的有限基因预测软件不能解决实际问题的说明,大规模实验验证是确保开发的基因克隆软件系统先进实用的关键,这已成为今后发展的方向。本研究进行的电子克隆,着重克隆细胞因子基因、癌症相关基因和信号传导基因等生物活性强、易于在体内组织细胞间穿梭、中下游开发相对容易的人类小分子肽类基因,并通过逆转录聚合酶链反应和基因表达谱分析等实验生物学手段验证其存在和功能,力求至少开发成功几个有价值的人类新基因并形成产业。因此研究中应理论与实验并重,生物信息和蛋白功能分析并重,编码蛋白结构、功能和表达调控机制研究并重,以尽快建立先进实用的人类新基因电子克隆的自动化软件系统,并获得一批国际专利保护的重要功能基因。

二、确定正确的研究目标、研究方法和技术路线

(一) 研究目标、研究内容和拟解决的关键问题

1. 研究目标

在深入研究电子克隆的原理及算法的基础上,建立人类新基因电子克隆的高精度、专家级、全自动化软件系统,从而获得一批具有自主知识产权的人类重要功能新基因,纠正网上公开的计算机注释人类基因组编码序列的错误,对功能明显的人类 RABL3 和 ZNF362,初步研究其功能,并进行调控序列电子克隆研究。

2. 研究内容及拟解决的关键问题

(1) 建立新的 PC 机上基因注释和搜寻软件平台 关键是最终建立人类新基因电子克隆的高精度、专家级、全自动化软件系统,适用于全球科学家网上克隆任何真核生物新基因。

①精度较高的基因及其调控序列的预测方法及原理 力求在编码区识别的精确度上超过国际上通用的基因预测软件(如 GENSCAN, GeneMark 等)并具抗干扰能力(能检测编码区删除、插入低于 0.3% 的基因, SN—敏感度、SP—特异度均超过 80%)。启动子预测软件精度达到: SN90%、SP50% 或 SN80%、SP80% (这个指标已超过国际现有的最好水平)。这些软件应能在 PC 机上运行并有直观的 User Interface。

②快速的可在 PC 机上运行的、进行批处理的 DNA 序列比对和基因组搜索软件 发展可在微机上运转的、界面友好的、可选择项目的,并且可以进行批处理的基因分析预测软件,至少能涵盖已经分布的 EST 库、公共的人类基因组序列库、无冗余序列库、SWISSPROT 数据库,也可以与以上注释了的基因组组合运用。其关键技术为改进的 Hidden Markov Model、Smith—Waterman 和 EULER 算法以及语言结构表征的启动子识别算法。

③搜索系统的完善与自动化 研究并设计适合我国研究人员使用的、节省时间和空间的搜索算法。

(2) 提交一批纠错的人类基因及其编码蛋白序列并经 RT—PCR 验证 研究中宜采用生物信息学方法,利用人类 EST 公共数据库和自行设计的软件进行 EST Contig 序列的全自动化拼接和校对,以 BLAST 比对为基本方法进行人类新基因大规模搜寻工作,力求至少搜寻到 1 000~3 000 个包含编码 100~300 氨基酸残基完整开放读码框架的新的人类 cDNA(400~3 000 bp);采用 cDNA 序列基因预测软件、包含校对过的 cDNA 的复原基因组序列预测基因软件、蛋白质功能结构预测软件,筛选出 300~500 个具有典型基因特点和一定功能结构域的人类新基因候选对象;再通过逆转录聚合酶链式反应(RT—PCR)克隆 cDNA 序列,得到重组质粒后并经过基因表达谱分析等实验技术验证其客观存在性,发现有价值的人类新基因、克隆小鼠和大鼠同源基因;进而参照基因存在论(Ontology)及基因组组成,纠正网上公开的人类基因及其编码蛋白质序列错

误,公开发表。

(3) 对已经计算机克隆获得的 2 个基因 RABL3 和 ZNF362 开展初步的功能研究 对已经电子克隆成功并经 RT—PCR 验证序列的一批基因,特别是功能线索明显的人类 RABL3 和 ZNF362(相应的小鼠和大鼠同源物已获得克隆并正式命名)等,及时采用细胞培养、实验动物、生物芯片等技术进行功能研究,重点研究细胞增殖与分化相关癌基因及转录因子功能。目前,SPRYD1、C17orf32 和 CLorf31 已交清华大学孟安民实验室转入模型动物斑马鱼体内合作进行功能研究。

(二) 研究方法和技术路线

1. 人类新基因电子克隆的全自动化软件系统开发与应用研究的方法

主要是开发和建立:高精度和高敏感的注释软件;新基因预测标准和基因组信息注释标准;跨平台、网络化的数据分析、搜索及软件共享系统。

应开发以下 3 个方面的软件。

(1) 高精度的基因预测软件 着重提高第一外显子预测的精确度。GENSCAN, GENEMARK 和 HMM 等著名软件的准确性都低于 50%,而当基因的外显子由很长的内含子隔开,外显子又较短或有删除插入,或者是基因发生重叠时,这些预测软件往往找不到基因;但从这些著名软件中吸取精华,在算法上加以改进以提高精确度,则是完全可行的。第二个方面是应采用基因预测和启动子预测软件合作运行的方法,用启动子预测来增加对第一外显子预测的能力。此外,应利用 EST、cDNA 等数据库的内容来佐证基因编码及调控序列,这将大大提高可靠性。

(2) 启动子的预测软件 宜根据启动子的生物学特征采用语言结构的数学模型,建立全新的启动子识别算法。

(3) 快速且可用于批处理的适应 PC 机运行的搜索、比对、拼接软件 这类软件的传统算法是 Smith—Waterman 或在此基础上增加了 Hashing 概念的算法,如 FASTA 和 BLAST。传统的 Smith—Waterman 的敏感性高但速度慢且内存要求太高,用改进的算法所写的拼接软件在重复序列高的基因组中出错率较高,应采用 PAVZNER, Tang 和 WATERMAN 等人提出的 EULER 图精确比对、拼接思想,设计可在微机上运转的快速的拼接、比对、搜索软件,并在此基础上包装出一个软件系统,以求可以自动地在分散的网站上搜索所需的信息。这些软件都应适合在微机上运行,以满足许多生物实验室的需要。

2. 人类新基因的电子克隆、实验验证、细胞增殖与分化相关癌基因和转录因子的功能及表达调控的方法

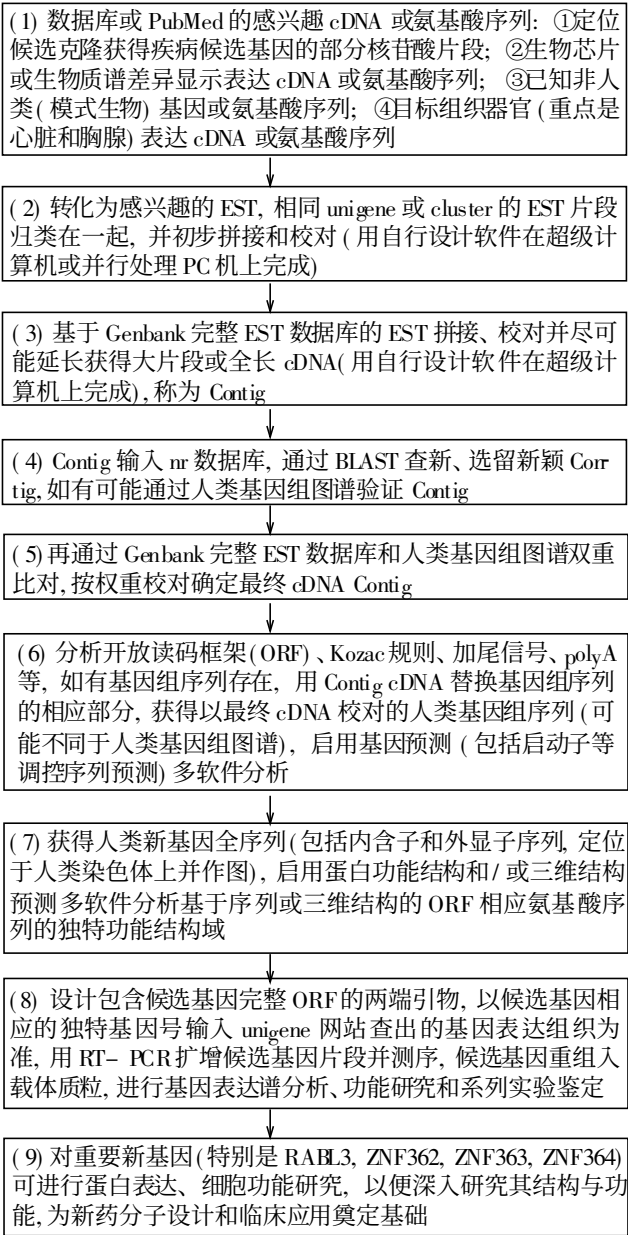
应采用以下 3 类研究方法。

(1) 生物信息学方法 包括数据库 (美国 ncbi, 英国 ebi, 日本 nigr, 加拿大 gdb, 北大 cbi, Tigr, Starford, Selera, 欧洲 embnet, 瑞士 expasy, Genbank, Swisprot, pdb, est, nr, htgs, human genome resource)、软件 (Pcgene, DNAsis, ORFfinder, E-PCR, Hidden Markov Model, Z curve 及上述数据库或网站中众多基因识别、蛋白质功能分析、蛋白质空间结构预测的软件, 应逐渐以至最终几乎全部使用自行设计的主要软件、BLAST 算法。

(2) 分子生物学、功能基因组学与蛋白质组学方法 包括 RT-PCR, RACE, DNA 序列分析, 基因重组, Northern blotting, Western blotting, 蛋白表达, 蛋白纯化, 转染技术等。

(3) 医学和实验动物学方法 细胞培养和动物实验等技术。

3. 技术路线



三、在我国开展该项研究的可行性分析和研究基础

该项研究在国内首次完全采用生物信息学方法电子克隆新基因, 瞄准细胞因子基因、疾病相关基因和信号传导基因等功能活性强的人类小分子肽类基因, 再通过分子医学技术实验确认, 充分利用了国内最好的理论与实验生物学条件, 并配合使用我国计算速度最快的超级计算机, 这种跨学科、跨部门的有效合作为预期研究目标的实现提供了有力保障, 具有明显的可行性。

要走在世界的前沿, 需要发现和培养一批既懂计算机又有生物学背景的研究人员。在生物学的研究中, 国外的软件占了绝大部分, 要把我们自行设计的软件和数据库推向市场, 除需要高精确度的软件外, 还要在网络化和易于操作上下工夫。因此, 生物信息学的人力资源是十分重要的, 为此, 北京大学从 2000 年开始率先在国内试点招收生物信息学跨学科研究生班(班主任是笔者之一钱敏平教授)。

特别是, 本研项目所在北京大学人类疾病基因研究中心是用“985 计划”经费装备起来的大型现代化重点实验室, 是北大计划医学部“十五”期间重点发展的三大新兴学科之一。中心主任是笔者之一北大基础医学主管科研的副院长马大龙教授, 中心学术委员会主任是韩启德院士。马大龙教授已指导多名博士分别实验克隆到多个重要的人类功能基因, 已申请国际专利并发表论文。因此, 总体指导教师具有人才培养的丰富经验, 将为本项目的完成提供卓有成效的指导和丰富的分子医学人力资源。另外, 合作研究者所在的北京大学数学学院可方便地利用校本部的一切优越条件, 并且钱敏平教授实验室几年来一直在开展人类基因组注释的理论与实践研究, 已取得可喜成果。

如前所述, 我们已电子克隆到 100 个人类新基因, 其中经实验验证的一些癌基因和转录因子与细胞增殖分化是密切相关的, 是具有多个重要功能结构域的高效基因, 其发现具有科学意义和重要的学术价值, 将来有可能开发形成产业。两年来, 本研究的初步成果的特色在于基因电子克隆与实验验证相结合的理论实验生物学手段, 并将人脑的主观能动性贯彻于研究过程的始终, 发现了自动化水平高的发达国家容易漏报的人类新基因, 既具有中国特色, 又具有规模效益, 不失为后基因组时代人类基因争夺战中行之有效的中国策略之一。今后, 本研究拟将此特色发扬光大, 进一步采用高通量、规模化的全自动化专家级电子克隆技术与基因功能体内、外筛选平台相结合的理论实验生物学手段, 达到高度集成化, 并定期将基因的生物医药理

论成果和科学知识纳入自动化预测软件系统,使其更新换代,长居国际先进水平。特别应指出的是:北大基因研究中心已在国内建立的人类新基因功能体内、外筛选平台是强有力的实验生物学支撑技术,为高通量电子克隆基因的实验验证提供了快捷有效的手段;在理论生物学方面,可与有良好合作关系的我国生物信息学先驱陈润生教授领导的中科院生物物理所生物信息学组合作,得到最佳的软件设计;还可将 NCBI Genbank 的 EST 数据库定期完全下载到我国计算速度最快的超级计算机——总参三部的超级计算机上进行高通量 EST Contig 序列的全自动化拼接和校对,与北大自行设计的并行处理 PC 机运行系统的工作效率进行比较研究。上述合作条件是本项目得以实施并可望取得成功的有力保障。

还必须看到,在以下几方面我们已具备必要的研究基础。在生物信息软件开发方面,已经建立了一个基因预测软件,在编码区识别的精确度上已经达到或超过国际上通用的基因预测软件的要求;拥

有自行设计启动子预测软件的能力,启动子和可变剪切预测的数学模型研究已经开始,并已有初步的成果,可以在微机上快速地比对和粘贴(将 EST 或 cDNA 序列粘贴到 genome 上去)。在数据库和跨平台的数据搜索方面,我们已经成功地把 ENSEMBL 的人类基因组注释内容及运用软件装载在当地的 Linux 系统上运行。

另外,关于新基因功能实验验证的深化部分,即,编码蛋白重要功能结构域的三维结构电子预测和实验测定;重要功能新基因调控序列的电子克隆、实验验证和表达调控机制研究,本项目研究者都参与过同类研究,技术路线清楚可行。钱敏平教授预测本中心新细胞因子 CKLF-1(人类趋化样因子-1)的新启动子也已被我们的实验所核实。

总之,本研究项目是《人类新基因的电子克隆、实验确认与功能研究》基金项目的深化和继续,实施时机已经成熟,我国应力争走在世界前列。

(删去全部主要参考文献,外刊上 25 篇、国内刊物上 2 篇) (责任编辑 王宏章)

(上接第 18 页)

在内容相对简单、涉及面相对较窄的课题中。所谓“网式结构”,是指文章内容有两条或更多线索,这两条或多条线索在文中交叉、编织。对于涉及面广、必须多头并进的课题,一般都采用网式结构。它的缺点是不如链式结构那样清晰。所谓“板块结构”,是指文章内容从总体上分成几个大部分。采用什么结构,要看文章的内容而定。因为结构只是形式,形式要为内容服务。一般来讲,运用分析—演绎范式的论文采用链式结构比较好,因为链式结构能够把文中所分析、演绎的内容简洁、清晰地表达出来。然而,如果运用综合—归纳范式的论文采用链式结构,往往效果不太好(不是绝对不可以)。这种范式的论文,采用网式结构或者板块结构就比较好。因为综合—归纳范式的论文讲究系统性,追求完善、周密,其内容必然比较繁杂、头绪较多。

同样,链式结构、网式结构、板块结构或者其他形式的结构,没有高低之分、也没有雅俗之别。它们各自为不同性质的内容服务,都有各自的优点和缺点,都有各自的生存空间和适用范围。然而,在学术界有些单位、有些场合,评委却明显地青睐链式结构,而贬斥、压制网式结构和板块结构。似乎只有链式结构才是正宗的论文结构形式,而网式结构、板块结构,统统都是歪门邪道、不登大雅之堂的。殊不知,在教育领域内,许多具有世界性影响的学术著作,就是用网式结构或板块结构写成的。当然,采用链式结构写成的具有世界性影响的教育学学术著作也很多。

独尊分析—演绎范式而贬斥综合—归纳范式,

独尊链式结构而贬斥其他结构,这难道不是一种“扬李抑杜”现象吗?笔者认为这种现象可能与评委自身的学术专长有关。一般来讲,研究教育经济学、教育心理学、教育社会学等教育学分支学科的研究者,对分析—演绎范式、链式结构比较熟悉、习惯和擅长。因为在这些分支学科各自的学术圈子内,大多数研究者都是按照分析—演绎范式进行研究工作,采用链式结构撰写学术论文的。然而,在另外一些分支学科,例如,在教育原理、课程与教学论、教育史、教育管理等领域内,较多的是运用综合—归纳范式进行研究,较多的是采用网式结构、板块结构来撰写学术论文。每个人都有自己的局限性,对于自己所熟悉、擅长的东西特别钟爱,这也是人之常情。但如果对自己所不熟悉、不擅长的东西,就大加贬斥,就很不妥、很不应该了。

应当营造一种“不拘一格”的环境。所谓“不拘一格”,简言之就是多样化。不论是分析—演绎范式,还是综合—归纳范式;不论是链式结构,还是网式结构、板块结构,都应当一视同仁地给予肯定和鼓励。那么,怎样客观地评价论文的水平呢?笔者认为,对于属分析—演绎范式和链式结构的论文,应按照分析—演绎范式和链式结构的规范要求去衡量和评价;对于属综合—归纳范式和网式结构(或板块结构)的论文,应按照综合—归纳范式和网式结构(或板块结构)的规范要求去衡量和评价。

参考文献

[1] 转引自:王通讯. 宏观人才学. 人民出版社, 1986, 93-94、59-60 (责任编辑 肖庆山)