生物学能引出新的数学定理吗?

Bernd Sturmfels

数学对于21世纪生命科学的重要性已经得到了大量阐述. 各所大学都热衷于启动旨在加强这两个领域之间相互渗透的初步计划, 各个国立研究所(AMI,IMA,IPAM, MBI, MSRI, SAMSI)则竞相设立以数学和生物学的交叉为主题的研究纲领和工作营. Clay数学研究所也没有游离于这一潮流之外。例如,在2005年夏天, 两位权威专家, Charles Peskin和Simon Levin, 作为Clay高级学者参加了在IAS/PCMI举办的数学生物研究计划,而在2005年11月, Lior Patcher, Seth Sullivant以及作者本人在剑桥的Clay数学研究所组织了主题为代数统计与计算生物学的工作营.

是的, 尽管这些普遍存在的研究设想和研究纲领已经展示了一个事实,但许多数学家仍不觉得信服,其中一些甚至私下希望这场"生物热"会很快烟消云散. 他们没有注意到数量生物学在他们从事的专业方向上的任何显著影响,因而他们提出一个响亮的问题: **生物学给出了新的数学定理了么?**

鉴于这些持续的怀疑, 一些长期观察者不禁想起20年前Gian-Carlo Rota写下的下面这一被广为引用的名句[16], "数学与生物学之间缺少实质的接触,很难判断这究竟是一个悲剧,一个丑闻,还是一个挑战", 他们想要知道事情是否在今天已发生任何变化. 当然, Rota很清楚数学在生物学中发挥作用的悠久历史,比如说在20世纪初Fisher,Hardy,Wright及其它人对人类遗传学的推动. 尽管如此, Rota并不认为这些工作是"实质的接触".

但是, 就在最近, 其它的一些声音出现了. 一些学者指出"实质的接触"意味着成为地位平等的伙伴, 而有意义的智力支援实际上是可以双方向进行的. 一幅乐观的景象可以很简洁地被J.E. Cohen的一篇文章[6]的题目所概括: "数学是生物学的下一个显微镜,只会更好; 生物学是数学的下一个物理学,只会更好".

物理依然是数学家心中的金本位, 因为在相当长时期内数学与物理之间已经存在着"实质的接触"和相互尊重. 历史上,数学对物理作出大量贡献,而在过去20年里,数学也从物理获得了远远超过预期的回报. 当代数学中许多最激动人心的进展受益于理论物理的研究. 难以想象, 倘若没有弦理论,镜对称和量子场论, 今天的几何与拓扑会是什么模样. 物理可以催生新的数学定理,这是非常"显然"的. 任何一个担心上座率的数学系的研讨会的主持者只要安排上一场一流物理学家的讲座就可以确保会场爆满. Clay高级学者Eric Zaslow在2005年6月的一次关于物理数学(Physmatics)的公共演讲中作了如下总结[1]: 数学与物理之间的相互影响已经如此深刻,以至于它们之间的界限正在模糊. 这两门一直就互补的学科之间已经开始拥有一种深层次的基本的关系...".

对数学而言,生物学会成为下一个物理学么? 在将来,会有一个理论生物学家摘取菲

原题: Can Biology Lead to New Theorems? 译自: http://www.claymath.org/library/. 原文是Clay数学研究所2005年年度报告的一部分. 本译文获得Clay数学研究所所长James Carlson和作者授权.
译者注: 该演讲全文见http://www.claymath.org/library/

尔兹奖章么? 这些问题的肯定回答都不太可能在最近得到证实, 我们当然也不能对它们在未来发生的可能性作任何估计. 然而, 我最近与计算生物学家的交流让我相信这种可能性远远比许多数学家所意识到的要大. 底下, 我将对一个更合理的问题提供一个个人的答案: **生物学给出了新的数学定理了么?**

我将展示四个灵感来自生物学的定理. 这些定理属于我的专业,代数,几何和组合学. 我把介绍由生物学所启发的在动力系统与偏微分方程中的工作的任务留给其它专家. 在进入本文的技术细节前, 我们必须作如下提醒. 下面所呈现的数学只是很微小的第一步. 这里讨论的对象与结果肯定还不象Zaslow关于物理数学的公开演讲中提到的内容那样深刻和重要. 但是, 跬步千里,毕竟伟业非一日之功.

我们的技术讨论从进化生物学对度量空间的一个贡献开始. 这是Andreas Dress和他的合作者们发展出的一个更广大的理论的一部分[2,9,10]. 一个有限度量空间是一个n阶非负对称方阵$D = (d_{ij})$,其对角元全都为零,且适合三角不等式$(d_{ik} \leq d_{ij} + d_{jk})$. 每一个$\{1, 2, \ldots, n\}$上的度量空间$D$对应于$\mathbb{R}^{\binom{n}{2}}$中的一个点. 所有这些度量空间一起形成$\mathbb{R}^{\binom{n}{2}}$中的一个多面锥,被称为度量锥[8].

对于度量锥上任一点$D = (d_{ij})$, 我们可以附上一个凸多面体

$$P_D = \{x \in \mathbb{R}^n : \ x_i + x_j \geq d_{ij}, \forall i, j\}.$$

倘若$D_1, \ldots, D_k$为度量空间,则$D_1 + \cdots + D_k$亦然,并且有

$$P_{D_1+D_2+\cdots+D_k} \supseteq P_{D_1} + P_{D_2} + \cdots + P_{D_k}.$$

如果上面的多面体包含关系中等式成立,我们就说和式$D_1 + D_2 + \cdots + D_k$是凝聚的. 一个$\{1, 2, \cdots, n\}$的分离(split)是一对适合$\alpha \cup \beta = \{1, 2, \cdots, n\}$的不相交的非空子集$(\alpha, \beta)$. 每一个分离$(\alpha, \beta)$定义一个分离度量$D^{\alpha,\beta}$如下:

$$\text{如果 } \{i, j\} \subseteq \alpha \text{或者 } \{i, j\} \subseteq \beta, \text{ 则 } D_{ij}^{\alpha,\beta} = 0; \text{ 否则 } D_{ij}^{\alpha,\beta} = 1.$$

相应于分离度量$D^{\alpha,\beta}$的多面体$P_{D^{\alpha,\beta}}$恰有一条有界边, 而该边的两个顶点分别对应于α和β的0-1关联向量. 一个有限集合X上的度量D被称为是分离素(split-prime)[2]的, 如果不存在一个X的分离$\{\alpha, \beta\}$以及一个正实数λ, 使得$D = (D - \lambda D^{\alpha,\beta}) + \lambda D^{\alpha.\beta}$ 成为一个D的凝聚分解; 这等价于说对X的任一分离(α, β)都成立$\min(\max(d_{ik} + d_{j\ell}, d_{i\ell} + d_{jk}) - (d_{ij} + d_{k\ell}) : i, j \in \alpha, k, \ell \in \beta) \leq 0$.

定理1 (Dress-Bandelt Split Decomposition [2]) *每一个有限度量空间D都允许一个唯一的凝聚分解$D = D_1 + \cdots + D_k + D'$,使得$D_1, \cdots, D_k$是线性无关的,而$D'$是分离素的.*

这一定理在进化生物学中的用处在于它为系统发生重建提供了一个多面体组合学的研究框架. 假设我们有n个分类单元, 譬如说, n种生物体的基因组, 并且令D为这些分类单元之间的距离矩阵. 在标准的应用场合, d_{ij}取为将染色体i和染色体j逐对对齐时所常用的Jukes-Cantor度量[21]. 我们接着考虑多面体复形$Bd(P_D)$,它的胞腔由多面体P_D的

译者注: 分离素的定义依据Andreas Dress教授的意见作了若干改动.

所有面所组成. 这是一个可缩复形, 它也是度量空间D的紧跨越(tight span)和内射闭包(injective hull). 度量D是一个树度量当且仅当其紧跨越$Bd(P_D)$是一维的,而且,在这种情形,这个一维可缩复形$Bd(P_D)$恰是给出树度量D的一颗系统发生树.

n个分类单元上的系统发生树所组成的树空间为Billera,Holmes和Vogtmann所引入[4]. 由于每个树度量唯一决定一颗树,树空间可以自然地嵌入度量锥,它由下述结论所刻画.

推论 树空间是度量锥的下述子集:

$$Trees_n = \{D \in \mathbb{R}^{\binom{n}{2}} : D \quad \text{是一个度量，并且 } \dim Bd(P_D) \leq 1\}.$$

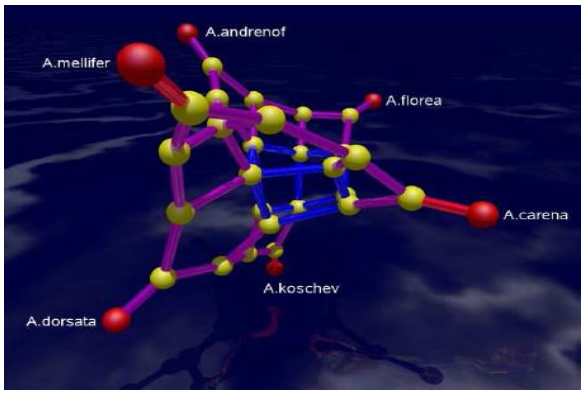

图1: 从蜜蜂6个种的DNA比对序列产生的六点度量空间的紧跨越. 我们感谢Michael Joswig与Thilo Schröder绘制本图并允许我们在此使用. 更详细的描述见[14].

如果一个度量D产生于实际数据,由于各种噪声的影响，它不太可能恰好落入树空间. 生物学家常用的算法,如邻域合并算法, 将计算出度量D在树空间$Trees_n$的投影. 从数学的角度，我们希望通过将树的概念替换为某种高维对象从而可以获得该实际数据的忠实表示. 紧跨越$Bd(P_D)$就是这样一个适用的对象，而且它的计算容易通过软件POLYMAKE来进行. 图1展示了在六个分类单元上的一个度量的紧跨越. 这个度量来自于六只蜜蜂的DNA序列的比对. 关于更多细节和POLYMAKE的介绍，请参见[14]. 我们指出，对于更大的数据集合，其紧跨越往往大得难以直接处理。这时候，定理1就有了用武之地：我们从数据集D中剔除被看作噪声的分离剩余D'. 留下的分离之和$D_1 + \cdots + D_k$可以由Huson与Bryant设计的软件SPLITSTREE来有效计算，并被一个系统发生网络所表示.

Andreas Dress现在担任中德合作在上海建立的计算生物所(www.icb.ac.cn)的所长. 2005年11月，他在剑桥的Clay研究所组织的一个工作营上阐述了他的理论. 当Dress于在柏林召开的1998年国际数学家大会上作邀请报告时，就指出“生命进化树是一个仿射厦”. 仿射厦是一种高度对称的无穷单纯复形，它在好几个数学领域中发挥重要角色，包括群论，表示论，拓扑，以及调和分析.

认识到系统发生树和它的可能高维推广与仿射厦有紧密联系是非常重要的观察. 本文作者热心支持Dress的观点，因为它与在系统发生学与热带几何的交叉研究中的最新进展相一致. [24]给出将树空间看作热带代数几何中的格拉斯曼空间的一种看法: 图2实际上描绘了格拉斯曼空间以及它的重言(tautological)向量丛. 基于这样的观念, Lior Pachter和Clay

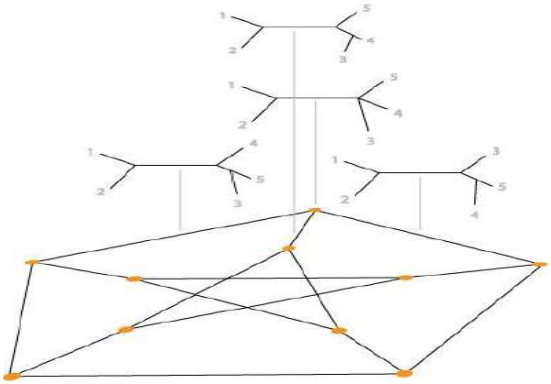

图2： 5个分类单元上的系统发生树空间是10维度量锥中的一个7维多面体. 它以Petersen图为组合结构. $Tree_5$由15个7维锥组成,分别对应于Petersen图的15条边. 它们相交于10个6维锥, 分别对应于Petersen图的10个顶点.

研究员David Speyer[20]在三年前发现了下一个定理.

令T 为一棵系统发生树，它的叶子用$[n] = \{1, 2, \cdots, n\}$来标号,且每条边被赋予一个非负长度. 于是我们对$[n]$的$m$元子集如下定义一个实值函数$\delta^{T,m}$: $\delta^{T,m}(I)$为由标号在I中的叶子所张成的的子树的边的长度之和. 当$m = 2$时，我们回到了树度量$D_T = \delta^{T,2}$. 我们称$\delta^{T,m} : \binom{[n]}{m} \to \mathbb{R}$为一个子树权函数.

定理2 (Patcher-Speyer Reconstruction from Subtree Weights [20]) 假设$n \geq 2m - 1$. 每一个n个分类单元上的系统发生树被它的子树权函数$\delta^{T,m}$所唯一决定. 更明确地说, $\delta^{T,m}$确定了$\delta^{T,2}$.

这个定理的妙处在于其统计应用。当$m = 2$换为更大的m值时，$\delta^{T,m}$可以更可靠地从数据中估计出来. 文章[19]阐述了这一方法在实用中的优点.

系统发生学在现代数学的几个不同研究方向上播下了种子，特别是在组合学与概率论中. 若要获得更深入了解，我们推荐Semple和Steel的书[23],以及将于2007年秋季在英国剑桥牛顿研究所举办的系统发生学特殊学期.

代数学家，几何学家和拓扑学家也许有兴趣了解一点儿系统发生学代数几何. 这儿的想法是生物序列进化的统计模型可以被翻译为张量空间的代数簇. 这一途径已经引导出一系列让代数学家感兴趣的最新进展；见[1,18,25]及其参考文献. 作为一个说明，我们介绍Buczynska和Wisniewski的一个近期结果[5]. 他们的预印本的摘要毫无疑问地说明了它是一篇对数学生物学而言不同寻常的论文：“我们探讨作为二元对称系统发生3-内度树(binary symmetric phylogenetic 3-valent tree)的几何模型的投射簇。我们证明这些簇有Gorenstein终止奇点（带有小的解消），它们是指标4的Fano簇...”.

这儿研究的簇都被嵌入射影空间$\mathbb{P}^{2^{n-1}-1} = \mathbb{P}(\mathbb{C}^2 \otimes \mathbb{C}^2 \otimes \cdots \otimes \mathbb{C}^2)$，其中的坐标$x_I$由$\{1, \cdots, n\}$的偶元子集$I$所标记. 我们固定一个叶子被$1, \ldots, n$所标号的内度为三的树$T$. 这棵树的$2n-3$条边中的任一条$e = u_e v_e$被等同于齐次坐标为$(u_e : v_e)$射影直线. 对$T$的叶子集合的每一个偶元子集$I$，$T$中存在唯一的以$I$为端点集合的不交路的集合，记之为Paths$(I)$. 这个观

察给出如下定义的$(\mathbb{P}^1)^{2n-3} \to \mathbb{P}^{2^{n-1}-1}$的双有理态射$\phi_T$：

$$x_I = \prod_{e \in Paths(I)} u_e \cdot \prod_{e \notin Paths(I)} v_e.$$

ϕ_T的像的闭包是一个射影环面簇，我们记之为X_T.

定理3 (Buczynska-Wisniewski Flat Family of Trees [5]) 当T跑遍有$n+1$片叶子的内度3的树的组合形时，所有的环面簇X_T都落入射影空间上Hilbert概形的同一个连通分支. 特别地，这些环面簇对应的凸多面体有相同的Ehrhart多项式([5,§3.4]给出了这个Ehrhart多项式的表达式).

作者与Seth Sullivant[25]的近期工作指出X_T的齐次素理想有一个由二次多项式构成的Gröbner基. 这些二次式来自一族矩阵的二阶子式，树T的每条边对应其中两个矩阵. 经过重新标号，我们可以假设边e将叶子$1, 2, \ldots, i$与叶子$i+1, \ldots, n$相分离. 我们构造两个$2^{i-1} \times 2^{n-i-1}$的矩阵M^e_{even}和M^e_{odd}如下。M^e_{even}的行由$\{1, 2, \ldots, i\}$的偶元子集I标记，列由$\{i+1, \ldots, n\}$的偶元子集J标记，而其(I, J)元为未定元$x_{I \cup J}$. 利用奇元子集可类似定义M^e_{odd}. 我们对环面簇X_T给出的Gröbner基由所有M^e_{even}及M^e_{odd}的二阶子式组成，这儿e跑遍树T的$2n-3$条边. 鉴于定理3，进一步讨论包含X_T的Hilbert概形的连通分支以及研究它与Keel和Tevelev[17]对模空间$\overline{M}_{0,n}$导出的一个二次方程的联系将会是一个有趣的课题.

发育生物学家把环面簇称为Jones-Cantor模型. 在某些应用场合, 更自然的研究对象是一般马式模型(general Markov Model). 这是张量空间中的非环面的射影簇，它推广了塞格莱簇(Segre variety)的正切簇(secant variety). 与这些模型相关的代数几何工作的最新进展可参阅Elizabeth Allman和John Rhodes[1].

我们要介绍的最后一个定理与系统发生学无关，而是产生于由生物序列分析中其它问题所激发的数学研究. 这些问题包括基因预测，其任务是辨认基因组中的基因片断，以及基因比对，其目标是探询两个基因组之间的生物关系. [22, §4]是面向数学家的的一个这方面问题的介绍. 目前的采用从头计算方法作基因预测和比对的算法都基于统计学习理论，它们涉及隐马氏模型和更一般的图模型.

从代数统计的观点出发，一个图模型是一个有丰富结构的从低维参数空间到张量积空间的多项式映射，我们在定理3中碰到的$\mathbb{P}^{2^{n-1}-1}$就是一个张量积空间的例子. 下面的定理是在研究图模型的这种代数表示时发现的.

定理4 (Elizalde-Woods' Few Inference Functions [11,12]) 固定一个正整数d. 考虑含d个参数的图模型G，且令E为图G的边数. 那么，该模型的推断函数数目至多为$O(E^{d(d-1)})$.[3]

有必要解释一下什么是推断函数以及定理4的意义. 一个图模型被一个多项式映射$p : \mathbb{R}^d \to \mathbb{R}^N$所规定，这儿$d$为固定的参数，$p$的每一个分量$p_i$是$d$个变元的次数为$O(E)$的多项式. 多项式$p_i$代表了在总共$N$种可能观察中作出第$i$个观察$\#i$的概率. 我们允许参数$N$变

译者注: 根据L. Pacher, B. Sturmfels, Eds., Algebraic Statistics for Computational Biology, Cambridge University Press, 2005,一书的定理9.3对作者的定理陈述作了修正.

大，在生物学应用中，它甚至可以相当大，譬如说$N = 4^{1,000,000}$,带有一百万个碱基对的DNA序列的数目.

p_i中的单项式表示相对应观察的可能解释，而在$\theta \in \mathbb{R}^d$上具有最大取值的单项式对应着参数为θ时的最可能的解释. 记Exp为所有N种观察对应的所有可能解释的集合. 取定一个一般的参数$\theta \in \mathbb{R}^d$, 我们得到一个映射

$$\phi_\theta : \{1, 2, \ldots, N\} \to Exp,$$

它把每一个观察映射到一个参数为θ时的其可能性最大的解释. 当参数θ取自$\mathbb{R}^d$中一个适当的开集时,任何这样的函数ϕ_θ被称为模型p的推断函数. 定理3前面构造的ϕ_T就是一个推断函数的例子. 所有从$\{1, 2, \ldots, N\}$到Exp的函数的个数$|Exp|^N$是一个天文数字. Elizalde和Woods的结果说明这些函数中只有非常非常小的一部分可以成为真正的推断函数. 定理4所保证的多项式增长级使得我们至少在原则上可以对每一个图模型预先计算所有的推断函数. 这对参数推断来说具有重要意义. 参数推断在生物医学中具体应用的两个近期例子可以在[3]和[7]中找到. 生物文章的作者排名次序有特定的意义,因此很少是按字母顺序来排,这也提供给你一种把生物文章从数学文章中区分出来的办法.

这就结束了我对四个由生物学所启发的的数学定理的讨论。它们都来自我所从事的领域，因此定理的挑选非常有倾向性. 定理1,2,3,4的共同特征为它们都是纯数学里头有意义的结论。由于没能对其它许多重要的研究成果加以报道，我必须向我的数学生物学的同行们真诚地道歉. 我只希望他们都会认可我的观点，即对本文题目中的问题的答案是肯定的.

参考文献 (略)

(吴耀琨 译　　丁克诠，苏大卫(David Surowski) 校)