物种在谱系树上亲缘关系的量度

黄建雄

导师：米湘成

研究目的

1. 在生物研究中传统的研究方法很多都是在物种水平上，如常用的多样性指数等。自从达尔文提出“进化论”以来，大家逐渐认识到在一个生态系统中物种之间都不是孤立存在的。
2. 传统的以形态及解剖学特征为基础的分类学方法提供的信息非常有限，随着目前技术的发展，我们已经有了更多的精确重建物种间的谱系树的方法，这就要求我们有一个更好的系统来表态物种之间的亲缘关系。

Part1 模型构建
介绍谱系信息描述模型

进化论的基本观点

1. 生物是由简单到复杂不断进化的。
2. 所有的物种都有一个共同的祖先。
3. 自然选择是生物进化的推动力。

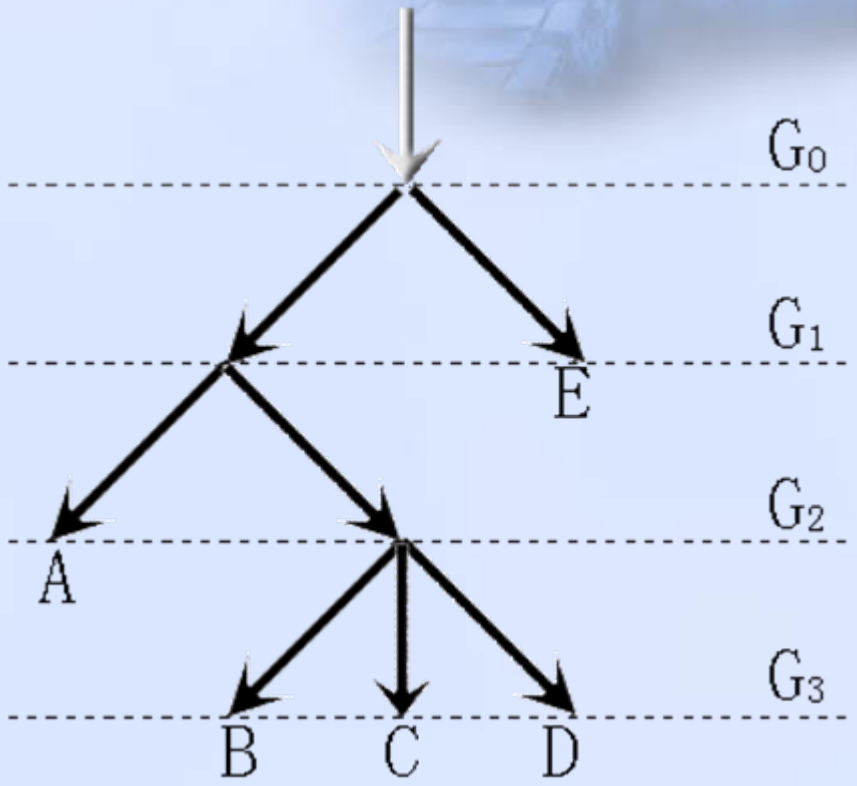


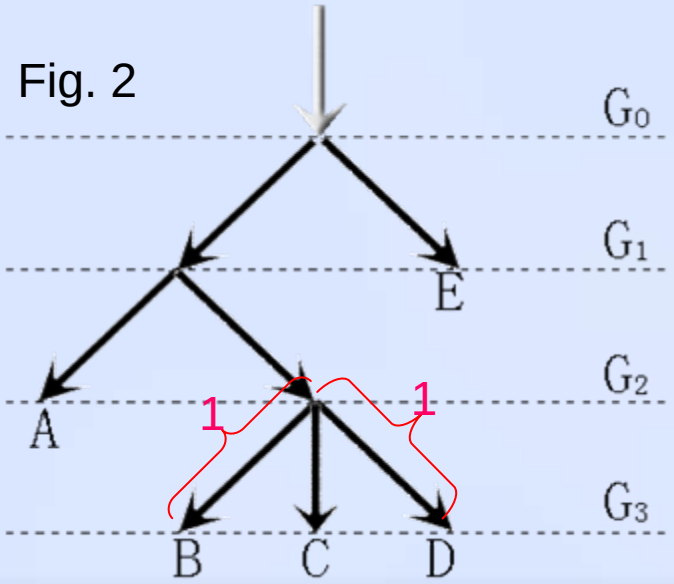
Fig. 1

研究现状

- 越来越多的研究人员已经认识到了这个问题，并提出了各种新的方法来描述谱系信息。例如：
  - Taxonomic Diversity (TD): Clarke & Warwick
  - Net Relative Index (NRI)/Net Nearest Taxonomic Index (NTI): Webb.
  - Phylogeny Diversity (PD): Faith.
  - QE based Index: Pavoine, S.
  - PSE/PSV/PSR: Matthew R. Helmus
  - …

上述指数存在的不足之处

- 这些指数从不同角度使用不同方法对谱系关系的简单描述，每一种方法都是在不同程度上对真实谱系关系的一种近似，因此通常在统计检验中能够获得比传统研究方法更好的结果。
- 但是这些方法存在一个共同的不足：都是利用谱系树上的节点及分枝信息做一些只有数学意义但缺少生物学意义的数据转换。



AvTD——实例分析

$$AvTD = \frac{2 * \sum\sum_{i<j} \omega_{ij}}{s(s-1)}$$

ω	A	B	C	D	E
A	0	3	3	3	3
B		0	2	2	4
C			0	2	4
D				0	4
E					0

- AvTD: 平均分类学距离。
- s: 物种数量。
- ω_{ij}:在谱系树上连接物种i与物种j的分枝总长度。

性状分解的基本思想

- 我们认为任何一个物种都是一系列性状的集合，这些性状可能是在它进化过程中的任意一次成种事件中产生的。
- 我们的基本思想就是将一个物种所有性状按照进化过程的性状的遗传与变异进行分解，进而发现多少性状分别来自哪一次进化。
- 进而研究一个物种的一个性状同时出现在谱系树上的其它物种中的可能性(UC)；或者研究2个物种包含共同祖先性状的可能性(PS)。

性状分解模型

- 根据进化论的观点“物种是由简单到复杂不断进化的”，一个物种在一次进化过程中，物种的性状总量是增加的，增加的量反映在进化树上就是分枝长度。
- 根据遗传学的观点，父本通过遗传把性状传递到子代，在成种事件中，该遗传过程又包含了一定的性状变异率。
- 因此一次进化过程形成的新性状包含新增性状及由祖先性状变异而形成的新性状2部分。如表1所示。

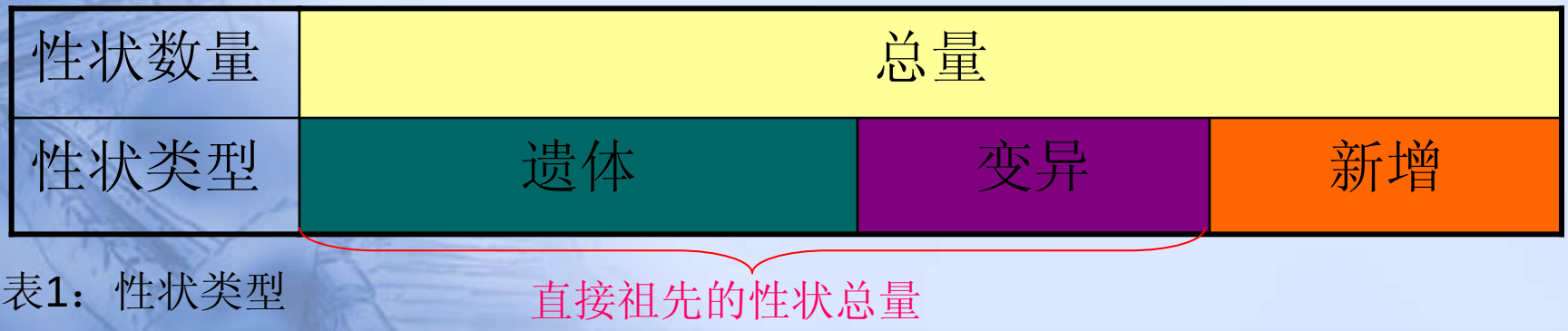


表1：性状类型

性状分解演示

以图1中的假想的进化树中的物种B为例，表2是分解过程。设q为0.5

进化阶段 (G_i)	G_3	G_2	G_1	G_0
新增性状量(R_i)	$R_3=1$	$R_2=1$	$R_1=1$	$R_0=1$
性状总量 (C_i)	$C_3=C_2+R_3$ $=4$	$C_2=C_1+R_2$ $=3$	$C_1=C_0+R_1$ $=2$	$C_0=0+R_0$ $=1$
变异性状量(M_i)	$M_3=C_2*q=1.5$	$M_2=C_1*q=1$	$M_1=C_0*q=0.5$	$M_0=0$
形成的新性状 (T_i)	$T_3= M_3+ R_3$ $=2.5$	$T_2= M_2 +R_2$ $=2$	$T_1= M_1 + R_1$ $=1.5$	$T_0= M_0 +R_0$ $=1$
物种B是保留的新性状(U_i)*	$U_3=T_3*p^0$ $=2.5$	$U_2=T_2*p^1$ $=1$	$U_1=T_1*p^2$ $=0.375$	$U_0=T_0*p^3$ $=0.125$

表 2：性状分解过程，p为遗体率,q为变异率(=1-p) sum(Ux)=2.5+1+0.375+0.125=4

谱系相似度(PS)

$$PS(A,B)=\sum_{i=0}^{k} T_i * P_A * P_B$$

- PS（A，B）：物种A与物种B的谱系相似度
- K：物种A，B在谱系树上共同祖先级数
- Ti：第i个共同祖先形成的新性状数量
- P_A,P_B：物种A与物种B分别包含Ti的概率，其值等于Ui(sp)/Ti;

PS的应用

- 谱系Alpha多样性

$$Simpson(PS) = 1-|\mathbf{p}*\mathbf{PS}*\mathbf{p}'|$$

  - PS为物种相似度矩阵，当PS=E时，该指数即退化为传统的物种水平上的Simpson指数(E为单位对角矩阵)

- 谱系Beta多样性

$$Sorenson(PS) = [\sum_{i=1}^{a} \max(PS(i, j \mid j \in b)) + \sum_{i=1}^{b} \max(PS(i, j \mid j \in a))]/(a+b)$$

  - PS为物种相似系数，max(PS(i,j∈b))代表群落A中的物种i与群落B中所有物种相似度的最大值。当只考虑物种水平的相似度时，该指数即退化为传统的物种水平上的Sorenson指数。

特异性系数 (UC)

- 很自然地，我们可以利用性状分解模型来描述群落中的一个物种相对于其它物种的特异性。
- 我们将一个物种中的含有的不被群落中其它物种包含的性状称为独特性状，并将UC定义为该物种期望的独特性状数量与该物种总性状数量的比值。

计算UC

- 在完成性状分解后，关键就在于获得每一个性状在群落中其它物种中都不存在的概率。
- 还以图1中的物种B为例，为了获得B在G0阶段形成的一个性状Ch_{b0}的独特率，就需要知道Ch_{b0}不在其它所有物种中的概率。
- Ch_{b0}不在E，A，C，D中的概率就分别是(1-p)，$(1-p^2)$,$(1-p^3)$, $(1-p^3)$因此Ch_{b0}的独特率就是(1-p)*$(1-p^2)$*$(1-p^3)$ *$(1-p^3)$

$$UC_i = \sum_{x=0}^{g} Ch_{ix} * P_{ix}$$

UCi:第i个物种的UC
G：进化次数
Chix:物种i在保留的第x次进化过程形成的新性状数量
Pix：Chix中的性状的独特率

UC的特点

- UC基于前述的性状分解模型，每一个过程都有较好的理论依据，而不仅仅是对谱系树的拓扑结构的简单概括。
- 同时UC在计算时充分考虑了一个物种与群落中其它每一个物种的关系，信息量更加完整。
- 此外，UC是一个不包含单位的系统，计算方便，能够较好的与当前广泛使用的其它指数结合。

Part2 模型有效性检验
利用UC检验幼苗死亡率与谱系的关系

目前关于幼苗死亡的主要观点

- Webb在06年AmNat上的文章指出：
  - 1、由于密度制约，同种的邻居对幼苗死亡的影响显著。
  - 2、由于病菌在亲缘关系相近的种间更易于传播，因此不同种的个体也会因为进化上的亲缘关系而对幼苗的死亡有显著影响。

Webb近年来在ecology、AmNat等高影响力的期刊上发表了一系列关于谱系的文章。

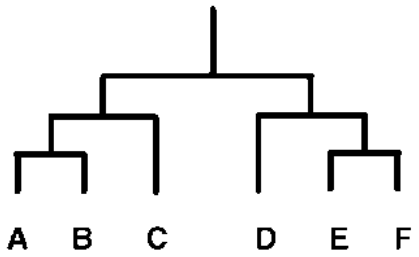
Webb的方法

- 使用多因子线性模型来检验因子的影响力
- 使用NRI、NTI来代表不同种的亲缘关系
- 使用AIC来比较模型的拟合度

Net Relatedness Index(NRI) and Nearest Taxa Index(NTI)

Webb 2000 AmNat

Phylogeny



Greatest possible mean *pairwise* nodal distance for a community of 4 taxa (given this phylogeny) = 3.66 nodes (for A, B, E, F)

Greatest possible mean *nearest* nodal distance for a community of 4 taxa (given this phylogeny) = 2.00 nodes (for A, C, D, F)

Community 1: A, B, C, D

Nodal distances:

	A	B	C	D
A		1	2	4
B			2	4
C				3

Mean *pairwise* nodal distance = (1 + 2 + 4 + 2 + 4 + 3) / 6 = 2.66

Net Relatedness Index = 1 - (2.66 / 3.66) = 0.273

Mean *nearest* nodal distance = (1 + 1 + 2 + 3) / 4 = 1.75

Nearest Taxa Index = 1- (1.75 / 2.0) = 0.125

Community 2: A, B, E, F

Nodal distances:

	A	B	E	F
A		1	5	5
B			5	5
E				1

Mean *pairwise* nodal distance = (1 + 5 + 5 + 5 + 5 + 1) / 6 = 3.66

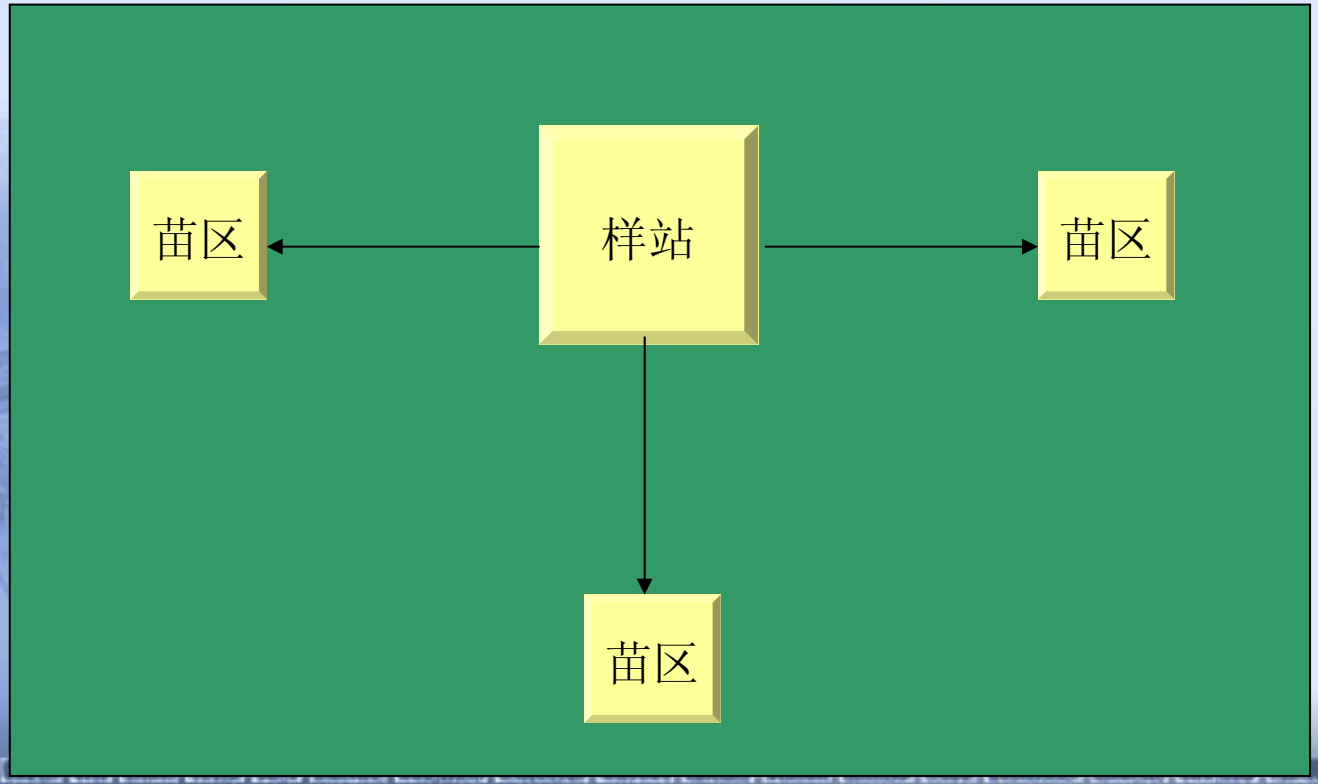
Net Relatedness Index = 1 - (3.66 / 3.66) = 0.0

Mean *nearest* nodal distance = (1 + 1 + 1 + 1) / 4 = 1.0

Nearest Taxa Index = 1 - (1.0 / 2.0) = 0.5

古田山幼苗数据

- 共169个样站，每个样站设置3个幼苗库。
- 总共进行6次调查，总共获得5次分析数据。

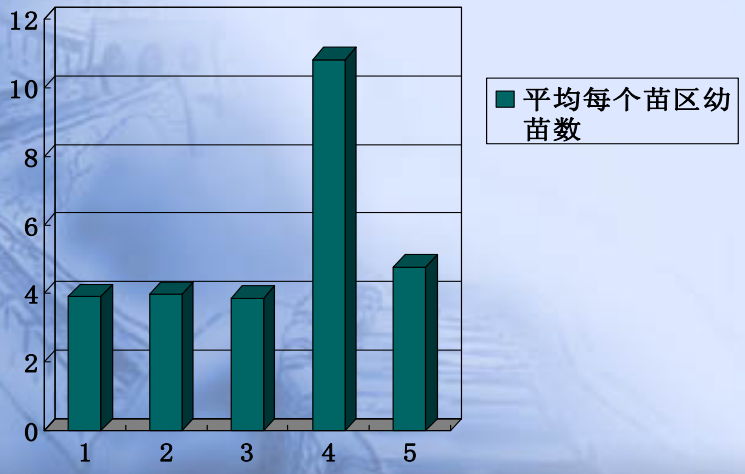


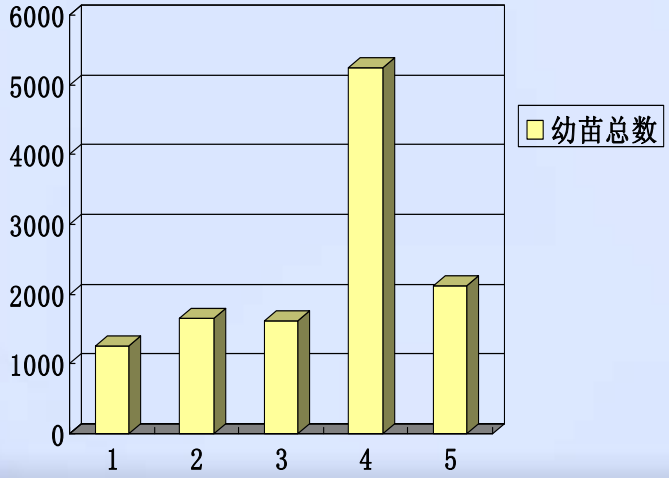
感谢陈磊提供数据

数据基本情况

- 从6次调查中获得5次比较数据：

调查时间	2006.5	2006.8	2006.10	2007.5	2007.8
幼苗总数	1264	1648	1612	5245	2118
平均每个苗区幼苗数	3.91	3.97	3.87	10.84	4.77



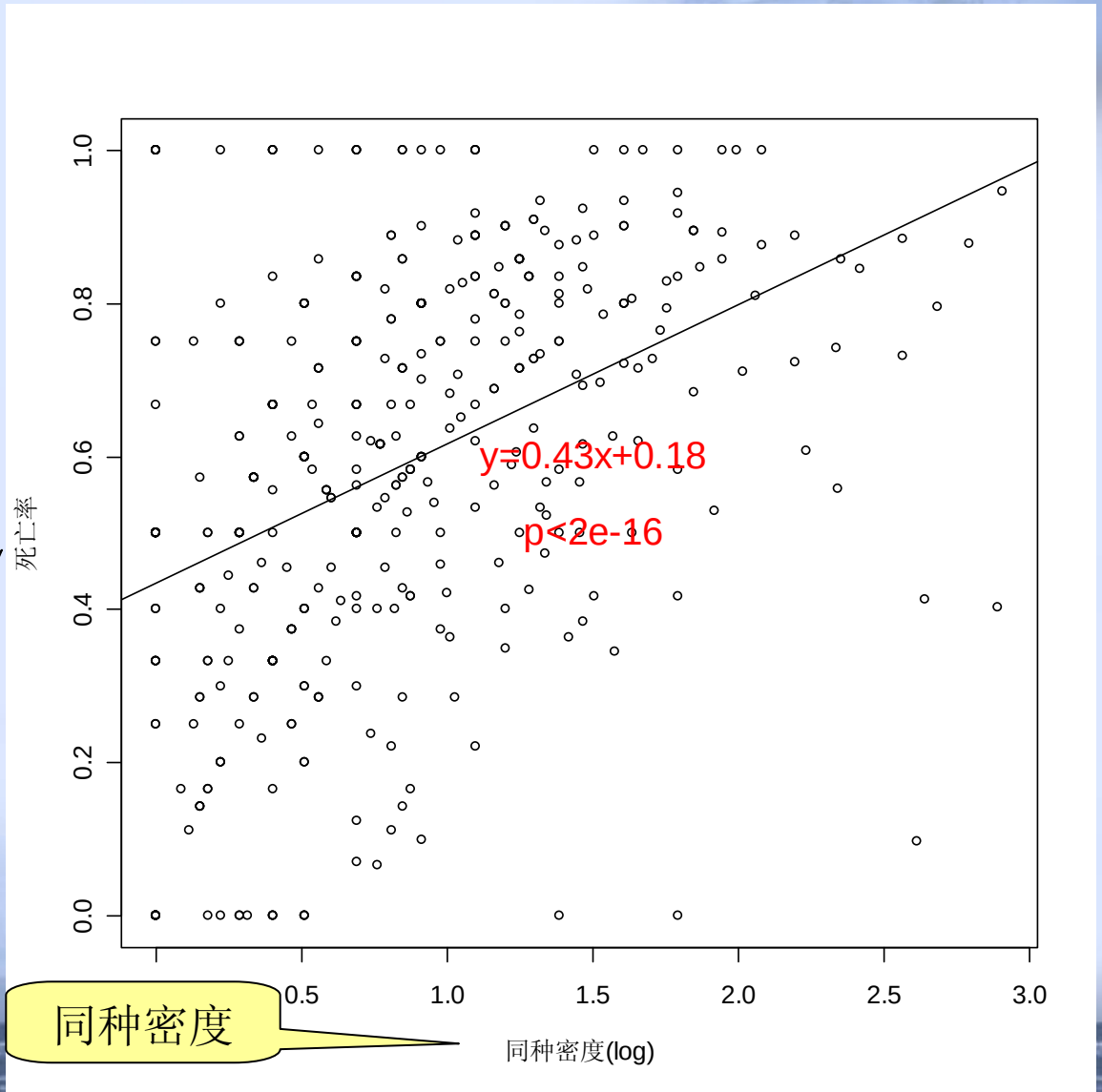


数据选择

- 5次比较数据中除第4次数据每个苗区的幼苗数达到10外，另外4次均不超过5棵。
- Webb的研究结果表明，受密度制约而导致幼苗死亡的主要因素是邻居中的同种个体数。
- 使用同种密度和死亡率做回归分析，发现在1、2、3、5次调查数据中不显著。表明密度制约效应只有在一定密度水平上才有效果。

同种幼苗密度制约

回归分析表明：同种幼苗密度与幼苗死亡率之间存在极显著的对数化线性关系。

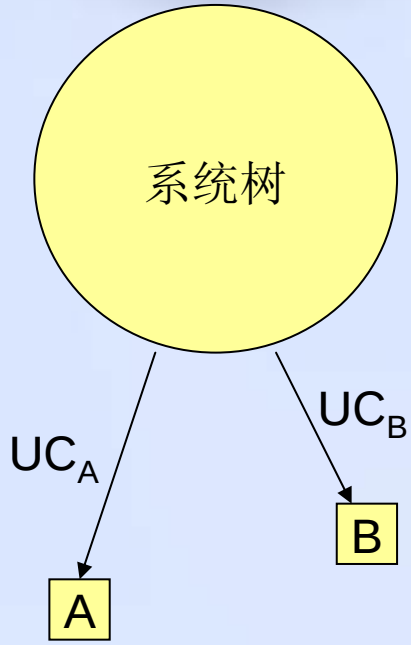


幼苗死亡率

同种密度

样地不同种UC平均值(mUC)

- 由于一个物种的UC与计算UC时所使用的谱系树上的物种数量相关，我们在此采用一个包含苗区所有物种的谱系树来获得每一个物种的UC。
- 使用苗区中两个不同种的UC的和来代表这2个物种在进化上的距离。mUC则代表所有物种对的进化距离的平均值。
- 由于所有的UC都是从同一棵谱系树上获得，从而使得mUC在不同的苗区具有可比性。



$$Dist(A,B) = UC_A + UC_B$$

谱系因子比较

	同种密度	同种密度+谱系因子		
		NRI	NTI	mUC
显著水平	***	***	***	***
AIC	-13	-35	-40	-56
Multiple R-Squared	0.1794	0.2239	0.2345	0.2613

*使用AIC比较线性拟合时，值越小越好

基本结论

- 使用UC来表达物种的亲缘关系时，其拟合结果显著优于NRI/NTI。
- NRI、NTI是Webb于2000年首先提出并加以应用，证明能够有效的代码物种间的谱系关系，我们采用与Webb相同的数据验证方法，获得了更好的结果，我们认为UC能更好的代表物种的谱系关系。
- 更重要的一点是，UC从遗传的角度出发，有完整的理论推断，而不是简单的对于进化树的拓扑结构的概括，因此UC对于物种亲缘关系的研究有重要意义。

讨论

- 从UC的推导过程，我们可以发现，一个物种的UC与群落中的物种数量是相关的，在群落中任意增加或者减少一个物种都将改变群落中所有物种的UC值。

不同q值对于拟合结果的影响			
q	0.1	0.01	0.001
AIC	-55.98	-55.88	-55.89

Thank You!