

分子流行病学研究及其应用

III. 主要研究手段

中国预防医学科学院流行病学微生物学研究所*

段广才 祁国明

生物(群)体间遗传距离的计算

一、原理:在分子流行病学研究中,经常要判断不同生物群体或个体间的遗传关系;同时,对生物体进行准确适当的分型和分类常是一切研究工作的基础和开端。既往,主要根据表型特征分型和分类。近年来,更多的研究集中在分子水平。但由于生物体的复杂性,不管是表型或基因型及其它分子水平的特征,都不仅是双相的,且要进行多个特征的综合评判。所以,数值分类在这一领域显示着极大的优越性。在表型分类中,由于各表型特征的可变性很大,检测手段的灵敏度和特异度又很不一致,所以,数值分类是使用等权数的。数值分类有两个最大的优点:一是可以进行多特征的数值化综合评判;二是各种特征以等权重加入不偏不倚。进入分子水平的研究以后,仍有一些学者这样认为就不妥了。第一个优点确实在任何情况下都能成立,但第二条则不然。这主要是检测手段和检测特征有了变化。所以,在分子水平分类和计算遗传距离常需使用不同的权重,而且还要符合数值分类的要求和分子遗传学基本原理。原因如下:

1. 分子及基因水平的特征是相对稳定的,不象表型特征那样受环境可变因素的很大影响。

2. 分子及基因水平的检测手段有很高的灵敏度和特异度,一致性好。如用基因探针检测不同的基因片段,对于不同的个体及不同基因位点之间,灵敏度和特异度相当一致。再如用MEE法对不同基因位点的酶进行分析,由于方法的一致性好,检出的各位点的变异度是可比的。

3. 分子和基因水平的变异符合分子遗传学基本原理,即在基因突变中,突变位点有难易之别。另一方面,单点突变较两点突变容易,两点突变较多点突变容易。因此,每个遗传位点变异所代表的遗传距离应有所不同,且要符合多点突变原则。当然,这里有一

个假设前提,即被检测不同位点代表相同的碱基数,其包含的易变碱基也相同,而且环境选择也相同。如没有这个前提,则每个遗传位点所代表的遗传距离计算就很不容易。

二、方法:

1. 权重系数的计算:普通数值分类相似及不相似系数的计算在专著中有详细论述,此处仅简要介绍一下分子流行病学研究中用到的距离系数。不同生物群体或个体间遗传距离的计算可以用不同遗传位点的等权重或不同权重计算。在等权的计算中,每个遗传位点所代表的遗传距离是相同的。如共检测五个遗传位点,两群体或个体间完全不同,则其遗传距离为1。如果有四个位点不同,则遗传距离为0.8。每个遗传位点所代表的遗传距离为: $D=r/n$; D -遗传距离, n -检测的遗传位点数, r -不相同位点数。这是最简单的计算方法。类似的方法还有: $D=1-S$; $S=2t/(t_1+t_2)$; D -遗传距离, S -相似系数, t_1 -个体或群体1所具有的特征数, t_2 -个体或群体2所具有的特征数, t -两个个体或群体相同的特征数。这个公式适于蛋白质分类和核酸酶切图谱分类以及质粒谱分析,上述的特征可以是不同的条带。

对于不同权重系数的计算也有不同的方法。Sela-nder等提出的权重系数为: $W_i = \sum_j F_{ij}^2 / \sum \sum F_{ij}^2$; $D = \sum W_i$; W_i -权重系数, F_{ij} -第 i 个遗传位点第 j 个等位基因的等位基因频率, D -遗传距离, r -遗传位点不同数。考虑到遗传多态度是表示每一遗传位点变异程度大小的客观指标,作者曾使用与遗传多态度直接有关的权重系数: $W_i = (1-H_i) / \sum (1-H_i)$, $H_i = (1 - \sum F_{ij}^2) / [N / (N-1)]$, $D = \sum W_i$; W_i -权重系数, H_i -第 i 个遗传位点的遗传多态度, F_{ij} -第 i 个遗传位点第 j 个等位基因的等位基因频率, N -样品数, t -检测遗传位点

*北京, 邮政编码 102206

数, r —两个个体或群体间不同遗传位点数, D —遗传距离。这两个方法都有一定的合理性和实用性, 但其最大缺点是不同位点之间权重悬殊太大, 以致在有些情况下不符合分子遗传学原理, 应用受到限制。最近作者经进一步探讨, 提出了与遗传多态度有关, 且符合数值分类要求和分子遗传学原理的权重系数计算方案。

$D = \sum W_i$, $W_i = (H_i^2 - 2H_i + 2) / \sum (H_i^2 - 2H_i + 2)$, 各符号的意义同上。上述方法可计算任意两个样品间的遗传距离。如要进行多个样品间的相互比较及分类, 可进行如下的数值分类分析。

2. 数值分类: 数值分类是将生物体特征(表型或基因型)的差异用数值来表示, 如1, 2, 3……, 先计算其相互间的相似系数(遗传距离), 如上所述, 获得距离矩阵。然后进行Q型系统聚类分析。其基本原理是, 构成独特性状的个体或群体作为分类单元(OTU), 将遗传距离最近的两个或几个OTU合并为一个新的OTU, 然后再计算这个新的OTU与其它OTU之间的遗传距离, 获得新的遗传距离矩阵, 再将距离最近的OTU合并为一个新的OTU, 如此逐步进行, 直到所有进行分析的群体合并为一个群体, 则聚类结束, 这是平均距离法的基本过程。此外还有最短距离法, 最大距离法等, 详见有关专著及文献。

蛋白质研究

用于蛋白质的鉴定、分离、纯化及特性研究的技术主要有蛋白质电泳技术, 如聚丙烯酰胺凝胶电泳、毛细管电泳等; 色谱技术, 如高效液相层析、亲和层析等; 分子免疫技术, 如蛋白质转印技术等。生物体内的蛋白质(如细菌的外膜蛋白等)是生物体特定基因的产物, 也是某些特征和功能的分子基础, 分离纯化这些特定的蛋白质, 然后研究其分子量及电荷的差异, 可以进行分子水平的分型和分类, 也可借助于分子免疫学方法进行快速诊断及抗原研究。

酶学方法

多位点酶电泳(简称MEE)法, 目前已发展成为一种研究生物体遗传关系及分型分类的重要方法应用于分子流行病学、分类学、群体遗传学等领域。基因的多态度也表现为酶分子的多态现象。因此, 对酶分子进行多态度的研究, 可以了解不同生物体之间的遗传关系、进化变异、分型分类、及正常和异常代谢等。在分子流行病学研究中, MEE法主要用于生物体的分

子生物学分类、分型, 不同生物群体或个体之间的遗传关系及生物群体(包括病原体或非病原体)的群体变异规律的探讨等。

核酸研究方法

用于分子流行病学研究的核酸技术主要有: 限制性内切酶图谱分析。由于限制性内切酶的在DNA上特异的酶切位点, 可以把大小相似但序列有差异的DNA切割成不同的片段, 在经过电泳等方法分离以后, 可以通过DNA谱带的相似性来判断不同来源DNA的异同性。基因探针分子杂交技术。通过标记特异的基因片段, 来探测待检样品中互补片段的有无, 与酶切电泳技术结合可以判断不同来源DNA或RNA片段间的异同及其相邻序列的异同。核酸序列分析。可以直接分析测定DNA或RNA序列的异同。多聚酶链反应技术。用合成特定引物的方法, 来扩增特异的DNA片段, 进而进行各种核酸分析。如电泳分析、分子杂交分析等。也可直接合成寡核苷酸作为探针应用。

一、核酸图谱及核酸酶切图谱: 染色体和质粒DNA以及病毒的RNA等是生物体遗传信息的载体, 比较其异同性是判断生物体间遗传关系的最直接的证据。

核酸图谱是分离纯化的核酸(DNA或RNA), 由于分子量的差异, 在电泳中迁移的速率有所不同, 用电泳的方法将其分为不同的带型, 通过带型异同度的分析, 直接判断生物体间的遗传关系。如质粒图谱分析, RNA图谱分析等都是这一原理。

核酸酶切图谱分析: 由于限制性内切酶在DNA分子上有特异的酶切位点, 如HindⅢ是由流感嗜血杆菌中分离出的核酸限制性内切酶, 其识别的特异序列为“AAGCTT”, 具有相同序列的DNA可被切割成相同的片段, 而序列不同的DNA则可产生不同的片段。通过核酸电泳可将分子量不同的核酸片段分离开来而获得的谱型称为核酸酶切图谱。用相同的核酸内切酶消化的DNA所产生片段的异同程度, 直接反应生物体间的遗传关系, 同时用不同的内切酶消化DNA可分析基因间的连锁关系等。核酸酶切图谱与基因探针结合起来, 用于研究特定的基因片段及其连锁基因的多态性等。

核酸酶切图谱较常用于比较不同群体基因组的异同, 从而判断它们的遗传关系。也用于某特定基因片段及相邻序列的多态度和缺失与否的研究。

二、核酸分子杂交: DNA是由两条多核苷酸链通

过碱基之间的氢键联接起来的双螺旋结构分子。其碱基之间遵循着严格的配对原则，即A-T，G-C。核酸分子杂交技术是应用便于检测的物质（如放射性同位素、生物素、地高辛等）标记已知核酸分子的一条链或特定片段作为分子探针，然后与处理后的样品中的核酸分子进行反应。如果样品中有标记核酸分子的同源序列或同源片段，则通过碱基互补的原则，样品中的同源片段可以与探针分子结合为双链分子。这个过程称为分子杂交。所形成的分子称为杂交分子。然后将未结合的探针分子除去，检测杂交分子的存在与否，即可判断样品中的同源片段。

原位杂交主要是检测样品中某特定基因片段的有无，如微生物的毒力基因等，判断样品中微生物等生物体的性状，为病源诊断提供直接证据。转印杂交（Southern blot检测DNA和Northern Blot检测RNA）不仅可研究某特定基因片段及相邻序列的有无，而且也可研究其群体多态度及生物体的遗传异同性。

三、核酸序列分析技术：生物体的各种遗传信息都储存在DNA（RNA）的特定核苷酸序列中，欲揭示生物体（或生物群体）间的遗传关系，直接测定核酸序列的异同程度，是最为准确有效的方法。目前应用DNA序列分析的方法甚多，如加减法、末端终止法，M13噬菌体法，以及最近的PCR法等，但其基本原理都是通过一定途径获得与目的DNA互补的不同长度的单链DNA，然后用凝胶电泳将其进行分离，再通过放射自显影等方法将各条带记录、分析。如ddNTP末端终止法是，在DNA合成中，DNA聚合酶不断将4种dNTP按碱基配对的原则加到引物的3'-OH末端，使引物延伸，合成出新的互补DNA链。当反应体系中有ddNTP（2', 3'双脱氧核苷酸5'-三磷酸）存在时，聚合酶将其加到新合成的链中，因为ddNTP的3'端没有-OH，使链不能继续延伸，合成链得以终止。用4中

不同的ddNTP分4组进行反应，即可使合成新链终止在不同的特定位置，产生不同长度的DNA片段，通过凝胶电泳分离及放射自显影，可以读出被测的DNA序列。

四、PCR技术：聚合酶链式反应（Polymerase Chain Reaction，简称PCR）技术是近几年来发展起来的一项新的分子遗传学技术。由于其不仅可以用于基因分离、克隆和核酸序列分析，还可用于突变体和重组体的构建、基因表达调控研究、基因多态度分析、遗传病和传染病的诊断、肿瘤机制探讨，以及法医学鉴定等很多方面，因而这一技术产生仅数年时间，就已广泛应用于生物学各领域。

PCR是体外酶促合成特异DNA片段的一种方法。众所周知，DNA在体内合成过程是以单链DNA为模板，在DNA聚合酶的作用下，引物可以沿5'→3'的方向不断延伸，从而合成新的互补DNA链。早期的PCR技术应用大肠杆菌DNA Pol I Klenow大片段酶进行体外合成，由于此酶不耐热，所以合成的每个周期中随着DNA的变性酶也失去活性，因而需要不断加入新酶，这种操作异常烦琐，未能普及。近年来，学者们从水生栖热菌（*Thermus aquaticus*，简称Taq）中分离到耐热的DNA聚合酶，应用于PCR，解决了上述难题，使PCR技术得到飞跃发展。其基本原理是，待扩增的DNA双链在高温（92-95℃）下变性为单链，再在低温（40-64℃）下使合成的特异引物DNA片段与其复性，然后在适当的温度（70-80℃）下进行引物链的延伸，即可体外合成新的互补DNA链。如此反复循环多次，由于Taq DNA聚合酶在70-80℃的条件下，可使链延伸60-150个核苷酸/秒，所以在短时间内可以合成大量的目的DNA片段。这一方法主要用于检测含量甚微的样品中某一特定片段的有无及其不同个体间此片段的异同。其可与酶切图谱法和分子杂交法联合应用从而获得更好的效果。