

哺乳动物驱动蛋白的计算基因组学分析与鉴定

薛宇^{①*} 刘丹^{①*} 符传孩^① 窦震^① 周庆^{①②} 姚雪彪^{①†}

(^① 中国科学技术大学细胞动力学实验室及合肥微尺度物质科学国家实验室, 合肥 230027; ^② 浙江大学农学院, 杭州 310003.

* 同等贡献. † 联系人, E-mail: yaoxb@ustc.edu.cn)

摘要 提供了一种新颖的、基于比较基因组学方法的全长驱动蛋白预测方法(full-length kinesin prediction program, FKPP), 用于哺乳动物中驱动蛋白质组的鉴定与研究. 之前的预测认为, 哺乳动物共含 94 个驱动蛋白, 而用 FKPP 从哺乳动物的基因组中总共鉴定出 134 个可能的驱动蛋白基因. 基于数据库中存在的片段序列, 用 FKPP 鉴定出 25 个可能的全长驱动蛋白. 此外, 用 FKPP 方法发现人的动点马达蛋白 CENP-E 应包含 2701 个氨基酸, 而不是当初根据克隆预测的 2663 个氨基酸. 通过对 CENP-E 的 cDNA 进行重测序, 并同时采用特异性识别 FKPP 预测的 CENP-E 多肽抗体予以实验评估, 结果表明, FKPP 预测的 CENP-E 氨基酸序列是正确的. 为此, 本研究利用 FKPP 对哺乳动物的驱动蛋白进行了重新分类. 鉴于目前公共数据库含有较多非全长序列的蛋白片段, FKPP 可提供一个具有显著效率且准确新颖的全长序列预测手段, 用于驱动蛋白以及其他蛋白家族的分子鉴定.

关键词 驱动蛋白 比较基因组学 CENP-E 全长驱动蛋白预测方法(FKPP)

驱动蛋白(kinesin)是基于微管的分子马达蛋白, 参与囊泡、细胞器、染色体、蛋白复合物以及 RNA 结合蛋白等的转运^[1-6]. 驱动蛋白通过调节微管的动力学变化, 从而决定细胞的动力学和可塑性^[7-10]. 驱动蛋白包含 3 个结构功能域, 即分子马达结构域、颈部以及茎部^[11]. 分子马达结构域包含了在真核生物中高度保守的氨基酸序列, 这些残基序列组成了一个 Walker A 的 ATP 结合域(ATP-binding motif)以及一个微管结合域^[11]. 在分子马达结构域之外, 驱动蛋白分子的序列同源性较低. 据此可以假设, 驱动蛋白的颈部和茎部决定着驱动蛋白与被运输物的特异性结合. 近期研究表明, 一些驱动蛋白与被运输物之间结合的特异性, 是通过结合在驱动蛋白的颈部和茎部的适配蛋白所介导的^[11]. 基于分子马达结构域的位置, 驱动蛋白可以分为 3 类: N-型、I-型以及 C-型^[11]. 尽管对于一些驱动蛋白的功能研究已经取得了很大的进展, 但是目前仍然不清楚哺乳动物的基因组中究竟存在着多少驱动蛋白的基因, 它们的功能特异性如何.

以前关于哺乳动物中驱动蛋白的研究, 主要局限于对小鼠中驱动蛋白分子的 ATP 结合域(约 160 个氨基酸)的鉴定^[12-14]. 然而, 完整的驱动蛋白分子马达结构域(motor domain)长约 300 氨基酸, 除了包括 ATP 结合域外, 还有一个紧邻的微管结合域(microtubule-binding motif). 仅对驱动蛋白的 ATP 结合域进行分析,

无法深入地理解驱动蛋白在分子水平上的功能保守性以及特异性. 对于驱动蛋白功能的分子生物学研究需要使用驱动蛋白的全长序列, 而通过实验来鉴定全长序列常常是需要耗费大量的人力、物力和财力, 并且实验进展常常受制于所使用的 cDNA 文库. 此外, 对于目前在基因组数据库中潜在的驱动蛋白序列, 还没有一个可以验证其正确性的方法. 因此, 开发一种有效准确的计算生物学辅助方法(*in silico*), 对于实验学家来说, 可以达到事半功倍的效果. 通过这样的计算生物学方法, 一旦新颖的、全长的驱动蛋白被预测之后, 这些可能的驱动蛋白的 cDNA 序列可以通过实验测序获得, 并且该蛋白在细胞内的功能也可以通过实验生物学的手段来鉴定. 因此, 计算生物学方法在驱动蛋白超家族分子功能的辅助鉴定方面具有两个基本的任务: (i) 对目前存在于公共数据库中的序列进行检验, 并计算构建已知驱动蛋白的全长序列; (ii) 寻找哺乳动物中尚未发现的驱动蛋白.

对驱动蛋白超家族成员进行准确的亚家族的分类, 有赖于对所有驱动蛋白基因的鉴定、进化分析及结构与功能相关性研究. 之前关于驱动蛋白系统发育的进化分析还局限在人、小鼠、果蝇、线虫和酵母^[12-14]. 而且, 进化分析所采用的序列, 仅仅是包含驱动蛋白分子的 ATP 结合域的片段序列. 最近, 研究人员对大鼠的基因组序列进行测序(大鼠基因组测序计划联盟, <http://www.hgsc.bcm.tmc.edu/projects/rat/>),

这使得通过比较基因组学的方法来鉴定哺乳动物中的驱动蛋白超家族可行且有效。

鉴于基因组学数据的日臻丰富,本研究开发了一套新颖的、基于比较基因组学的计算方法,即全长驱动蛋白预测方法(full-length kinesin prediction program, FKPP)。在目前的公共数据库中,许多蛋白都没有全长而只有片段序列,而 FKPP 方法则可以作为一个有效的分析手段预测出马达蛋白家族新成员的全长序列。基于 FKPP 对全长马达蛋白序列的计算结果,我们对驱动蛋白超家族重新进行了系统发育的进化分析并完成了新的亚家族分类。这些分析结果对于进一步的实验生物学研究具有重要的指导意义。

1 材料和方法

(i) 冗余序列的去除。为了尽可能地鉴定驱动蛋白,本研究使用了一个简单的规则去除冗余序列。通过序列比对,我们发现,目前数据中的驱动蛋白存在着很多的冗余序列:有一些仅仅只是很少的残基的不同,其实是同一个基因(如测序错误或者单核苷酸多态性(SNP)),而其他则可能是可变剪接异构体(alternative splicing variant)。本研究使用 BLAT 工具^[15]检验了所有预测出来的驱动蛋白。BLAT 是一个快速、准确的局部比对工具,有核酸序列比对和蛋白质序列比对两种模式。不考虑嵌合基因(nested gene),并且如果被预测出来的两个驱动蛋白在基因组上的坐标存在着重叠现象,则认为这两个蛋为同一个基因的表达产物。只有序列最长的蛋白被保留以进行接下来的分析。

(ii) 使用 FKPP 进行全长驱动蛋白的预测。同源序列的比对搜索之后,我们发现有一部分驱动蛋白的序列是不完整的。因此,本研究使用了一个新颖的比较基因组学的方法,命名为全长驱动蛋白预测方法(FKPP),通过片段的驱动蛋白序列的进化保守性来预测全长的序列。我们将这个方法应用到人、小鼠、和大鼠的基因组上,流程如下:

(1) 在同源比对搜索和去除冗余之后,从人、小鼠、和大鼠的基因组中分别抽取驱动蛋白的氨基酸序列。用 Clustal W/X 工具对 3 个物种中每一组的直向同源物(orthologs)进行比对,参数设置为缺省^[16]。如果这 3 条序列中某一条较短,并且只能被比对到其他序列的部分区域上,跳转到步骤(2)。如果 3 条序列都可以在全长范围内进行比对,则跳转到步骤(3)。如果步骤(3)结束,则结束整个计算的循环。

(2) 假设存在 3 个互为直向同源物的驱动蛋白 HsKIF-A (人)、MmKIF-A (小鼠)和 RnKIF-A (大鼠)。其中 HsKIF-A 和 MmKIF-A 比 RnKIF-A 要短得多,只能比对到 RnKIF-A 上的一部分区域。用 NCBI 的 RnKIF-A 的序列,在人和小鼠的蛋白质组数据库中(nr 数据库)进行 BLASTP 的比对,参数为缺省。如果同源性 > 70%,则结果被保留用做进一步的分析。这样就从人和小鼠的蛋白质组中分别得到了关于 RnKIF-A 的 N 和 M 条同源的片段序列。在人中,这些序列分别为 HsKIF-A, Hs-F₁, Hs-F₂, ..., Hs-F_{N-1}, 而在小鼠中,这些序列分别为 MmKIF-A, Mm-F₁, Mm-F₂, ..., Mm-F_{M-1}。这些同源的片段序列可以是不连续的,也可以是相互重叠的。把人和小鼠的 N'和 M'不连续序列的重叠片段序列拼在一起,组成一个较长的序列,并直接使用 RnKIF-A 中存在的序列来填补在人和小鼠中是空白的区域,这样可以得到拼接的序列 HsKIF-A'和 MmKIF-A'。用序列 HsKIF-A'和 MmKIF-A'继续进行同源的搜索,如果更多的同源片段序列被发现,则直接用这些对应 RnKIF-A 的互补区域来产生新的拼接序列 HsKIF-A'和 MmKIF-A'。然后把 HsKIF-A'和 MmKIF-A'序列通过 BLAT 工具分别定位到各自的基因组上^[15]。尽管其他的用来确定外显子/内含子结构的工具也同样可以使用以达到类似的目标,但我们在此仅使用 BLAT 工具。

如果全长的拼接序列不能很好地定位在基因组上(如存在不能匹配的片段、填补空白的部分片段序列不能定位在基因组上、原始的基因结构不完整、BLAT 的比对结果分数下降超过 10%),则将拼接出来的序列视为不合格样品而舍弃。显然,在进化过程中,有一些位点会发生突变,并且产生少量的插入或缺失片段^[17]。因此,以基因组本身的内容作为参考,可以修正在填补空白片段上已经变化的残基或核苷酸。这样可以保证用 RnKIF-A 中对应 HsKIF-A'和 MmKIF-A'相应的空白部分进行填补,并且在修正之后能够分别与人和小鼠的基因组本身的内容保持一致(100%相同)。然而,如果在修正的序列中发现了终止密码子(stop codon),将放弃这些样本,否则将保留结果。这样得到了序列 HsKIF-A''和 MmKIF-A''。接着跳转到步骤(1)。

(3) 检验序列中可以进行比对的区域。如果一个驱动蛋白相对于其他的另外两个序列具有更多的小的片段,则将这些片段填补在其同源序列的对应

位置上. 然后, 得到的拼装序列要通过 BLAT 被定位到自身的基因组上. 这里的结果接受的判定规则与步骤(2)是一样的. 接着跳转到步骤(1). 由于 FKPP 是一个非常简便的方法, 实验生物学家可以很容易地手工实现这个计算的流程.

(iii) FKPP 的实验验证. FKPP 预测的 CENP-E 分子包含一段在最初克隆的 CENP-E 分子中不存在的 36 个残基序列. 为了验证 FKPP 的准确性, 我们测定了 DNA 序列并人工合成了一个对应此段序列中 15 个残基的多肽 (PYLQTKHIEKLFTANC; BACHEM America), 然后采用标准的多肽抗体制备方法获得有效抗体, 进行免疫沉淀、Western blotting 及免疫组化的实验验证^[9]. 为了验证通过 FKPP 的计算生物学方法推导的 CENP-E 与过去克隆的人的 CENP-E 蛋白呈现一致的亚细胞定位, 我们使用免疫荧光标记的方法来观察微管与 CENP-E^[9]. 为了与原始发现 CENP-E 蛋白的工作相比较, 我们使用了鼠抗 177(mAb 177)作为对照^[6].

2 结果

2.1 FKPP 预测全长驱动蛋白

在进行同源比对搜索之前, 我们从公共数据库, 如 NCBI, GenBank, Swissprot, Kinesin Home Page (<http://www.proweb.org/kinesin/>)及 Ensembl 中得到各个物种中的驱动蛋白的序列. 这些序列包括酵母(*S. cerevisiae* 和 *S. pombe*)、线虫(*C. elegans*)、果蝇(*D. melanogaster*)、小鼠(*M. musculus*)、大鼠(*R. norvegicus*)以及人(*H. sapiens*)共 7 个物种的驱动蛋白氨基酸序列. 使用标准的 BLASTP 工具, 将每一个已知的驱动蛋白序列在这 7 个物种的蛋白质组数据库中进行序列比对, 以发现同源的蛋白. 对于在酵母、线虫以及果蝇中设置的准则是, 序列的相似性要大于 30%, 并且 E 值小于 e^{-10} . 而在小鼠、大鼠和人中, 设置了两个准则: (i) 在两两比对中, 能够互相对应的区域需要大于短一些的序列的 80%; (ii) 两条序列可以比对的部分, 其相似性需要大于 30%.

由于基因组计划并没有同时提供全长的 cDNA 序列, 虽然目前有一些哺乳动物的基因组已经测序完成, 但是现在仍然不清楚在小鼠、大鼠和人的基因组中究竟存在着多少驱动蛋白的基因. 目前收集了部分驱动蛋白的数据库仅有 Kinesin Home Page, 共列出了关于这 3 个物种的 94 个驱动蛋白, 其中人驱动蛋白 36 个、小鼠 47 个以及大鼠 11 个. 在这些已

经列出的驱动蛋白中, 大约有 20 个驱动蛋白仅仅只有部分的氨基酸序列(~160 个残基).

由于 cDNA 文库中驱动蛋白基因拷贝数较少, 鉴定全长的哺乳动物驱动蛋白序列较为艰难. 如果有一个行之有效的计算生物学手段, 即可准确确定哺乳动物驱动蛋白组及其结构与功能相关性. 基于此设想, 我们开发了一套新颖的全长驱动蛋白预测方法, 得到可能的全长的哺乳动物, 以驱动蛋白序列. FKPP 是一种基于比较基因组学的方法, 其理论依据是在人、小鼠、大鼠中的马达蛋白序列非常保守. 在啮齿类的蛋白编码区域存在着少量的插入或缺失片段(~21%, 1743/8148), 并且这些短的插入或缺失片段一般以 5% 的点突变率出现, 长度一般是 1~10 个核苷酸^[17]. 驱动蛋白在小鼠和人中的高度保守^[12-14], 表明比较基因组学的计算生物学方法准确可靠.

基于数据库中已有的片段序列, 通过计算得到了 25 个全长的驱动蛋白序列. 例如, MmKIF16A 蛋白 (GI: 2443266, kinesin-3) 即使在同源的序列比对搜索之后仍然只有 160 个残基的序列, 但是其在大鼠中可能的直向同源物 RnKIF16A (GI: 34857644), 则有 4614 个残基. 通过 FKPP, 我们可以得到可能的 MmKIF16A 序列是 4529 个残基. 图 1 中给出了对于 KIF16A, FKPP 是如何对其进行全长序列的预测分析的. 首先通过使用全长的 RnKIF16A 序列在小鼠蛋白质组中进行 BLASTP 同源搜索, 这样可以得到包括原始 MmKIF16A 在内 3 的条与 RnKIF16A 同源的序列(图 1(b)). 将这 3 条片段序列, 包括 mKIAA1300 (GI: 28972710), 未命名蛋白(unnamed protein; GI: 26325666)以及 MmKIF16A (GI: 2443266), 拼接成一个完整的序列, 同时空白缺失的部分直接使用大鼠 RnKIF16A 上对应的序列来填补, 这样便产生了一个嵌合的驱动蛋白序列. 接着, 利用 BLAT 工具的蛋白比对模式, 将这个嵌合的驱动蛋白序列定位到小鼠的基因组上. 序列在基因组上定位之后, 按照基因组的内容对嵌合的序列进行修正, 从而保证了该序列与基因组的内容 100% 的匹配(图 1(c)).

2.2 哺乳动物及其他真核基因组中驱动蛋白的系统发育分析

用 FKPP 计算得到所有哺乳动物驱动蛋白序列后, 采用系统发育的分子进化方法对人、小鼠和大鼠中的驱动蛋白进行亚家族的分类. 将这些驱动蛋白保守的分子马达结构域序列抽出, 用 Clustal W/X 进

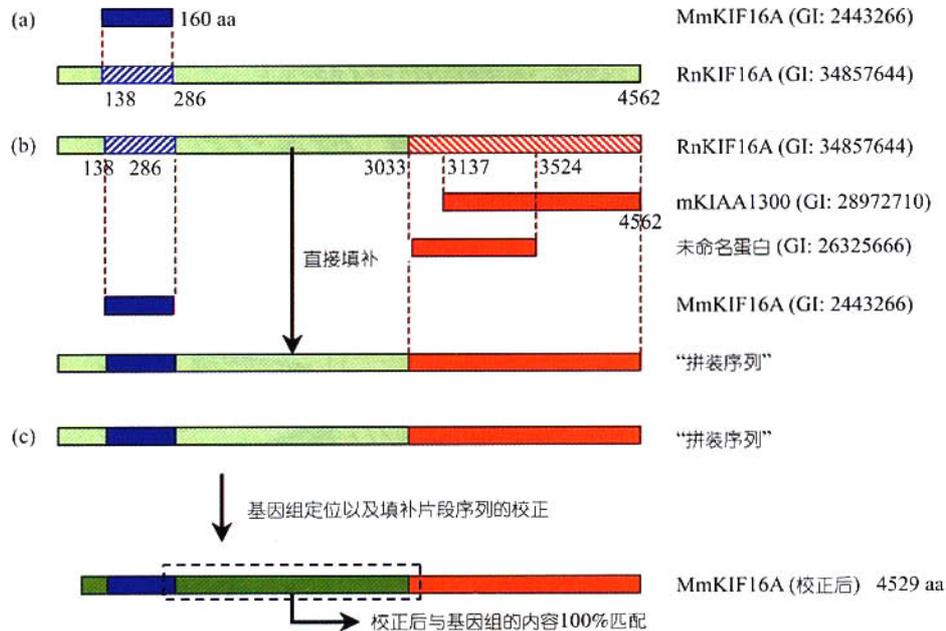


图 1 使用 FKPP 来发现驱动蛋白的全长

用 MmKIF16A (160 个残基, GI: 2443266) 演示如何将一个较短的序列通过 FKPP 方法得到一个完整的全长序列。(a) 将 MmKIF16A 与 RnKIF16A 进行序列比对; (b) 用 BLASTP 在小鼠的蛋白质数据库中进行同源比对搜索 RnKIF16A 的同源序列。为了简化计算, 这里只列了 3 条主要的序列 MmKIF16A (GI: 2443266), mKIAA1300 (GI: 28972710) 和未命名蛋白 (GI: 26325666), 其中 mKIAA1300 和未命名蛋白首先可以合并成一条序列, 然后通过 RnKIF16A 上互补的序列填补空白缺失的部分, 这样得到一条嵌合的序列; (c) 依据小鼠基因组的内容, 对嵌合的序列进行修正, 从而保证与基因组的内容达到 100% 的一致。这样计算出来的全长的 MmKIF16A 包含了 4529 个残基

行多序列分析并予手工校正。与文献[18]报道的方法一致, 进化树的建立使用 MEGA 程序(2.1 版)^[18], 并进行了手工校正。图 2(a)中显示了哺乳动物中的驱动蛋白分子进化树, 所使用的建树方法是邻接法(Neighbor-Joining)、Bootstrap 检验和泊松修正(Poisson Correction)。图 2(b)中则展示了从酵母到人的 7 个物种中的驱动蛋白的分子进化树, 所使用的方式是最小进化法(Minimum Evolution)以及伽玛距离模型(Gamma Distance)。对两树进行了 Bootstrap 检验, 结果表明, 所构建的进化树是可靠的。与之前的报道相一致^[12~14], 本研究建立的进化树都属于无根树(un-root)。

通过 FKPP 鉴定, 我们总共鉴定了 134 个驱动蛋白(表 1), 其中小鼠含 45 个, 大鼠含 45 个, 人含 44 个。我们使用了标准的驱动蛋白名称命名法^[19], 同时加上了前缀“Hs”(人)、“Mm”(小鼠)、“Rn”(大鼠), 以此来命名 3 个哺乳动物中的驱动蛋白。与以前哺乳动物驱动蛋白的鉴定(Kinesin Home Page)相比, FKPP 鉴定出 49 个新的驱动蛋白, 其中包括人的新驱动蛋白 8 个, 小鼠 6 个, 大鼠 35 个。为了基于驱动蛋白所包含的结构域(domain)与结合域(motif)对驱动蛋白进

行重新分类, 本研究用了 Interpro 数据库^[20]分析 FKPP 所得到的新的驱动蛋白。对功能结构域和结合域的预测, 用缺省的参数。并且, 用在 FKPP 中新发现的序列来进行系统发育的进化分析。

本研究计算结果与之前的工作总体吻合^[12~14]。然而, 新加入的驱动蛋白序列信息对于哺乳动物驱动蛋白的分类提供了更多的有效数据。在之前的工作中^[12~14], 由于数据不全, KIF6, KIF7 和 KIF9 都不能得到分类, 因而被看作是孤儿(orphan)蛋白。然而, 通过 FKPP 分析, 我们可以准确地将 KIF6 和 KIF9 划分入 kinesin-9 亚家族, 将 KIF7 划分入 kinesin-4 亚家族。之前的工作由于引入一个可能不存在的驱动蛋白 KIF8 (在小鼠中不能正确地定位到基因组上, 在人与大鼠中也没有发现直向的同源物), 使得之前工作中的 N-2 亚家族的分类出现了较为明显的偏差。而本研究则较好地将该亚家族成员进行再分类。该亚家族的另一成员 KIF11, 按照标准的驱动蛋白名称命名法, 分入 kinesin-5 亚家族中。因之前的工作在动物(细胞)中只鉴定出 KIF19A 和 KIF19B, 这两个驱动蛋白与 KIF22 被错误地归入到一个亚家族中。鉴于与

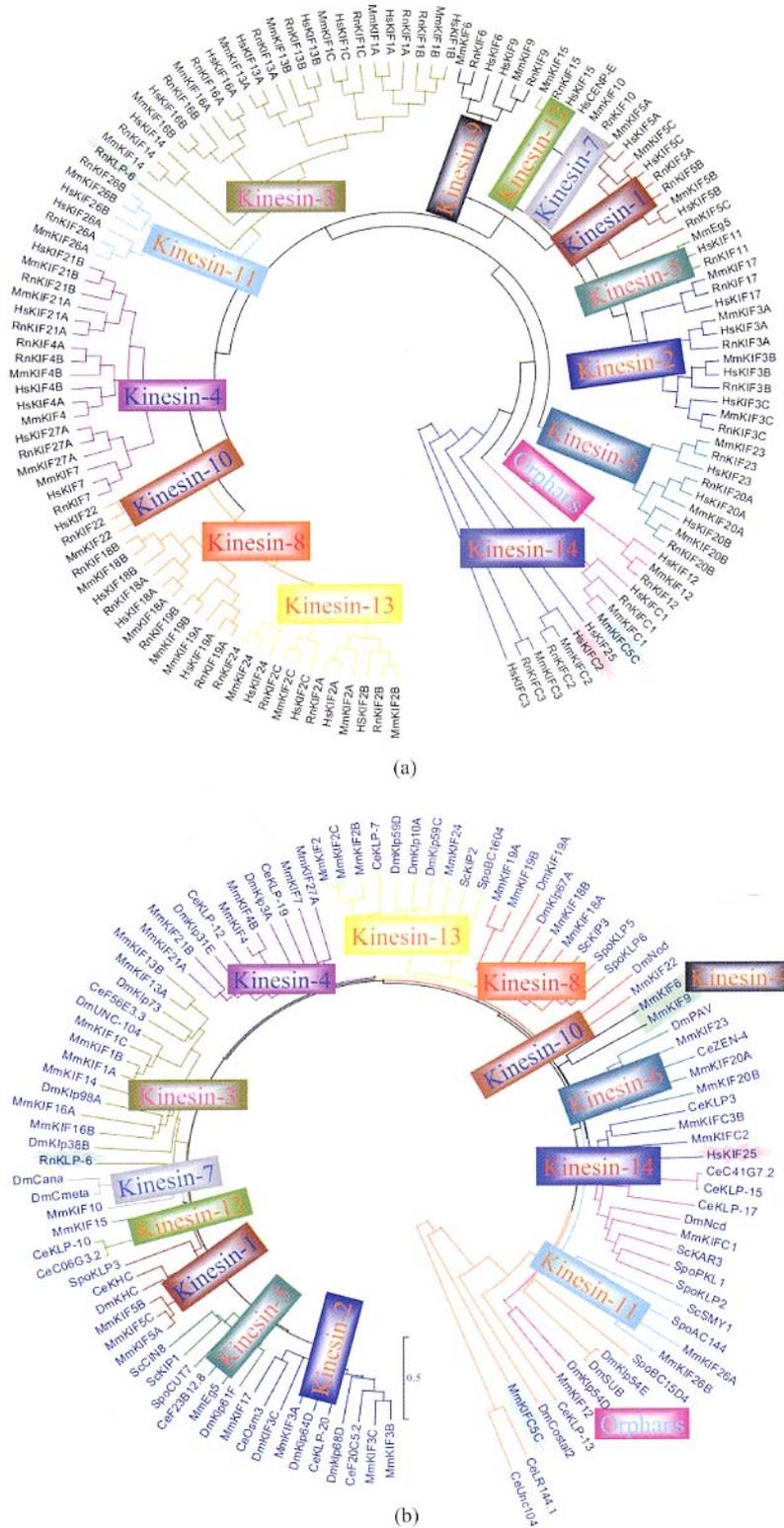


图2 使用全长驱动蛋白的分子马达结构域进行系统发育树分析

(a) 用邻接法构建了人、小鼠、大鼠的驱动蛋白系统发育树, 同时用 bootstrap 进行检验; (b) 用最小进化法构建了包括人、小鼠、大鼠、果蝇、线虫、芽殖酵母和裂殖酵母共 7 个物种的驱动蛋白系统发育树. 这里, 选用小鼠的驱动蛋白来代表哺乳动物

表 1 FKPP 方法得到的完整驱动蛋白序列与之前发表工作的比较^{a)}

驱动蛋白主页			同源搜索之后			FKPP
蛋白名称	蛋白标示符	长度(aa)	驱动蛋白统一命名	蛋白标示符	长度(aa)	长度(aa)
HsCENP-E	399227	2665	HsCENP-E/HsKIF10	399227	2665	2701
HsKSP	1706622	1056	HsKIF11	1706622	1057	
KIF12	NP_612433	513	HsKIF12	32699596	551	618
HsRBKIN1	11761611	1805	HsKIF13A	21361722	1805	
HsGAKIN	8896164	1826	HsKIF13B	29421214	1835	
HsCMKrp	452517	1648	HsKIF14	23396633	1648	
HsKlp7	9910266	1388	HsKIF15	9910266	1388	
			HsKIF16A	34527855	323	4441
HsJ777L9	6522736	412	HsKIF16B	27529917	1393	1796
HsKIAA1405	7243191	791	HsKIF17	34978376	1029	
HsDKFZp434	12053149	898	HsKIF18A	21314742	898	
			HsKIF18B	37544008	870	
HsFLJ3730	NP_694941	548	HsKIF19A	23397458	548	894
HsATSV	2497523	1690	HsKIF1A	19924175	1690	
HsKIF1B	3043706	1338	HsKIF1B	42560524	1816	
HsKIF1C	3913961	1103	HsKIF1C	40254834	1103	
HsRabK6	3978240	890	HsKIF20A/MKLP2	5032013	890	
HsKlpMPP1	5911999	1780	HsKIF20B	15919888	1820	
HsNYREN62	5360129	633	HsKIF21A	33187651	1662	
HsLOC34389	XP_291594	512	HsKIF21B	41114119	1726	
HsKid	4519443	665	HsKIF22	6453818	665	
HsMKLP1	400264	960	HsKIF23/MKLP1	20143967	960	
			HsKIF24	34532133	850	1355
HsKlp6q27	4115553	384	HsKIF25	20138788	384	
HsKIAA1236	6330751	1481	HsKIF26A	20521808	1840	1887
			HsKIF26B	41114119	1726	
			HsKIF27A	30794488	1401	
HsKin2	3024057	679	HsKIF2A	4758644	679	
HsLOC8464	NP_115948	673	HSKIF2B	21707472	673	
HsMCAK	1695882	725	HsKIF2C	20141607	725	
HsKIF3A	3851492	702	HsKIF3A	33112673	702	
HsKIF3B	3913958	747	HsKIF3B	40788226	760	
HsKIF3C	3913957	793	HsKIF3C	3913957	793	
HsKIF4	7266951	1232	HsKIF4A	13959694	1232	
LOC347363	29743725	304	HsKIF4B	41147002	1234	
HsnKHC	2497520	1032	HsKIF5A	2497520	1032	
HsuKHC	417216	963	HsKIF5B	4758648	963	
HsxKHC	3043586	957	HsKIF5C	40788283	999	
			HsKIF6^{b)}	ENSP00000287152	482	525
			HsKIF7	38348350	830	1343
HsKIF9	11275982	725	HsKIF9	18202950	790	3308
HsCHO2	3702453	673	HsKIFC1	33875771	725	
			HsKIFC2	21955174	838	
HsKIFC3	12654739	694	HsKIFC3	34098691	694	
MmKIF10	2443268	160	MmCENP-E/MmKIF10	40388490	2474	
MmKIF11	2443270	170	MmKIF11	45476577	1052	
MmEg5	4160556	1014				
MmKIF12	12858387	642	MmKIF12	33563262	642	
MmKIF13A	10697238	1749	MmKIF13A	30794518	1749	1784
MmKIF13B	2443276	160	MmKIF13B	38076359	1960	
MmKIF14	2443278	166	MmKIF14	38073343	966	1662
MmKIF15	2443280	166	MmKIF15	38173736	1387	

续表 1

驱动蛋白主页			同源搜索之后			FKPP
蛋白名称	蛋白标示符	长度(aa)	驱动蛋白统一命名	蛋白标示符	长度(aa)	长度(aa)
MmKIF16A	2443266	160	MmKIF16A	2443266	160	4529
MmKIF16B	2443262	150	MmKIF16B	38075140	2008	
MmKIF17	2443264	159	MmKIF17	23396634	1038	
MmKIF18A	12862603	151	MmKIF18A	21314852	886	
MmKIF18B	12862605	149	MmKIF18B	37537560	834	
MmKIF19A	12862606	148	MmKIF19A	12862607	148	1000
			MmKIF19B	38081372	748	
MmKIF1A	2506794	1695	MmKIF1A	2506794	1695	
MmKIF1B	2497524	1150				
MmKIF1Bbrain	5081553	1816	MmKIF1B	5081553	1816	
MmKIF1Bbeta	4512330	1770				
MmKIF1C	3913960	160	MmKIF1C	23821040	1100	
MmKlp174	1695174	887	MmKIF20A	6679597	887	
MmKIF20B	12862615	225	MmKIF20B	38085340	1774	
MmKIF21a	6561827	1573	MmKIF21A	6561827	1573	1638
MmKIF21b	6561829	1668	MmKIF21B	6561829	1668	
MmKIF22	2558833	148	MmKIF22	21704182	660	
MmKIFd19	12851512	457	MmKIF23	29568094	953	
MmKIF24	12862611	138				
MmKIFj19	12855902	157	MmKIF24	45708952	1320	1327
			MmKIF26A	38073760	1989	
			MmKIF26B	34328423	1550	1729
			MmKIF27	32401469	1394	
MmKIF2	125402	716	MmKIF2A	125402	716	
MmKIF2beta	2695866	659				
			MmKIF2B	38092011	706	
MmKIF2C	12862613	155	MmKIF2C	29840788	721	
MmKIF3A	125403	701	MmKIF3A	125403	701	
MmKIF3B	3122327	747	MmKIF3B	3122327	747	
MmKIF3C	3913959	796	MmKIF3C	3913959	796	
MmKIF4	1170659	1231	MmKIF4A	1170659	1231	
			MmKIF4B	20858575	1222	
MmKIF5a	3929108	1027	MmKIF5A	3929108	1027	
MmKIF5b	2062607	963	MmKIF5B	2062607	963	
MmKIF5c	3929110	956	MmKIF5C	3929110	956	
MmKIF6	2443284	165	MmKIF6	31581530	481	600
MmKIF7	2443286	168	MmKIF7	38086933	1328	
MmKIF8	2443288	185				
MmKIF9	5295882	790	MmKIF9	26325458	810	3205
MmKIFC2	1944330	792	MmKIFC2	1944330	792	
MmKIFC3a	2443294	157	MmKIFC3	12585614	709	
MmKIFC3b	12585614	709				
MmKIFC1	1944328	609				
MmKIFC5A	6979905	674	MmKIFC5A/MmKIFC1	13277705	674	
MmKIFC4	2558829	155				
RnKRP6	2674187	169	RnKIF11	34862680	1165	
			RnKIF12	34868461	617	
			RnKIF13A	34874048	1826	
			RnKIF13B	34874311	1903	
			RnKIF14	34880426	1659	
			RnKIF15	31335233	1385	
			RnKIF16A	62645648	4562	

续表 1

驱动蛋白主页			同源搜索之后			FKPP
蛋白名称	蛋白标示符	长度(aa)	驱动蛋白统一命名	蛋白标示符	长度(aa)	长度(aa)
			RnKIF16B	34859296	1569	2254
			RnKIF17^{e)}			983
			RnKIF18A	34856640	659	804
			RnKIF18B	34873908	991	
			RnKIF19A	34875045	1046	
			RnKIF19B	34871266	1882	
			RnKIF1A	34877667	1943	
RnKIF1B	3493139	689	RnKIF1B	29789307	1816	
RnKIF1D	2370435	1097	RnKIF1C	22024392	1097	
			RnKIF20A	34878647	888	
			RnKIF20B	34862643	2017	
			RnKIF21A	34867881	1767	
			RnKIF21B	34880431	1670	
			RnKIF22	34859268	609	
			RnKIF23	34864667	896	947
			RnKIF24	34867277	1277	1320
			RnKIF26A	34935518	1933	
			RnKIF26B	34881041	1813	
RnKRP5	2674185	167	RnKIF27A	38016129	1394	
			RnKIF2A	34854206	771	
			RnKIF2B	27675158	712	
RnKrp2	2772516	671	RnKIF2C	20279134	671	
			RnKIF3A	34870729	708	
			RnKIF3B	34859022	562	747
RnKIF3C	3913949	796	RnKIF3C	16758244	796	
			RnKIF4A	34881081	1243	
			RnKIF4B	27668154	1224	
			RnKIF5A	34865745	1066	
			RnKIF5B	34876212	1114	
RnKHC	3122309	238	RnKIF5C	34854278	1004	
RnKRP3	2674181	160	RnKIF6	34874329	631	
			RnKIF7	34857299	1334	
			RnKIF9	34866378	3304	
RnKRP1	2674179	162	RnKIFC1	2674179	163	
RnKRP1	5070666	247	RnKIFC5A	34852223	616	
			RnKIFC2	38454244	791	
RnKRP4	2674183	153	RnKIFC3	34851230	739	
			RnKLP-6	34881054	1057	
			RnKIF10	34860597	2726	

a) 总共 134 个驱动蛋白被鉴定, 而之前发表的工作只包含 94 个. 通过 FKPP 得到的完整的驱动蛋白标注为粗体. 本表采用的数据库主要是 nr 数据库; b) HsKIF6 的序列取自 Ensembl 数据库; c) RnKIF17 则是通过 FKPP 分析小鼠和人中的可能直向同源物得到

KIF19A 及 KIF19B 相似度很高的 KIF18A 和 KIF18B 的发现, 我们对这个亚家族进行了重新划分. KIF22 被归入 kinesin-10 亚家族, 而 KIF18A 和 KIF18B 以及 KIF19A 和 KIF19B 则被归入 kinesin-8 亚家族. 对 kinesin-8 亚家族的分析在后文给予了详细的报道.

FKPP 发现了哺乳动物中含有一些物种特异性的驱动蛋白, 即只存在于该物种中, 而在其他物种中不存在直向同源物. 例如, 小鼠中的 MmKIFC5C, 大鼠

中的 RnKLP-6, 以及人中的 HsKIF25. 因此, 研究这些物种特异性的驱动蛋白将是一个有趣的工作. 基于系统发育的分析, 我们将 MmKIFC5C 和 HsKIF25 归类到 kinesin-14 亚家族中, 而 RnKLP-6 则属于 kinesin-3 亚家族. 进一步地, 本研究认为, 哺乳动物的 KIF24 蛋白应当是一个 I 型/M-亚家族的成员 (kinesin-13), 因为该蛋白与其他已经确证的 M-亚家族驱动蛋白之间的进化距离非常接近, 而远离传统

意义上的N型驱动蛋白。并且,大鼠中RnKIF24的分子马达位于N端下游约170个残基,离N端较远。而传统意义上N型驱动蛋白的分子马达结构域通常位于N端,同时在分子马达下游不超过50个残基处有一个颈部。M型的驱动蛋白则没有发现存在颈部结构。因此,一个重要而有意思的问题是,KIF24蛋白究竟能否运动,并且与没有颈部结构的M型亚家族的驱动蛋白相比,该蛋白是如何运动的?实验生物学研究将为我们提供一个肯定的答案。

尽管关于驱动蛋白的分类已经有报道^[12~14],然而之前的工作并没有考虑哺乳动物中的KIF7究竟应该属于哪一个亚家族及其与其他驱动蛋白的关系。基于本研究的系统发育分析,我们将哺乳动物中的KIF7蛋白归于kinesin-4亚家族。并且,我们还鉴定出属于kinesin-4亚家族的一个新成员KIF27。该驱动蛋白基因至少包括3个剪接异构体(A, B, C)。在之前的工作中^[12~14],哺乳动物中的KIF24和KIF25都被认为是N型-11亚家族(该亚家族还包括了KIF26A与KIF26B),然而本研究的分析表明,KIF24应该是一个M型的驱动蛋白,因此被列入kinesin-13亚家族(该亚家族还包括了KIF2A, KIF2B, KIF2C及其他蛋白)。并且,HsKIF25仅仅存在于人中;作为人的特异性驱动蛋白,被列入kinesin-14亚家族。尽管从整体上来说,哺乳动物中的KIF6与KIF9的相似性并不算高,但是这两个驱动蛋白分子马达结构域的序列具有高度的相似性,表明这两个驱动蛋白可能起源于同一个祖先。因此,我们将这两个驱动蛋白列入kinesin-9亚家族中。

从进化距离来看,哺乳动物中的KIF18A和KIF18B以及KIF19A和KIF19B(人中只有HsKIF19A)与裂殖酵母中的SpoKLP5和SpoKLP6以及芽殖酵母中的ScKIP3具有非常高的同源性。这些基因/蛋白都被划入kinesin-8亚家族中。有意思的是,ScKIP3, SpoKLP5以及SpoKLP6都在有丝分裂(Mitosis)期间发挥重要功能,且定位于动点(kinetochore)上^[21]。因此,我们推测这个kinesin-8亚家族的成员可能定位在动点,参与有丝分裂。

为了进一步验证通过FKPP方法得到的人驱动蛋白确实存在并且在人的组织中正常表达,我们对人(细胞)的驱动蛋白的表达情况进行了初步的分析。将人中总共44个驱动蛋白在UniGene^[22]以及人的ESTs数据库中分别进行同源比对的搜索,以此来考

察这些驱动蛋白是否在人的组织细胞中正常表达。同时,参照GeneCards^[23]数据库,以考察在正常人组织细胞中驱动蛋白的表达谱以及组织表达的特异性。结果(表2)表明,有9个驱动蛋白被鉴定为免疫系统特异性表达的蛋白,这些驱动蛋白在免疫系统的表达量比在其他组织中高出8~10倍。因为免疫组织是一个细胞有丝分裂旺盛的组织,这些免疫系统特异性表达的驱动蛋白,很有可能在细胞有丝分裂过程中有重要调控功能。本研究分析结果也初步印证了这个假设。在这9个驱动蛋白中,至少有4个已经被报道在细胞有丝分裂过程中起着重要的作用,它们是KIF10/CENP-E^[9,10], KIF2C/MCAK^[24], KIF11/EG5^[25]和KIF20A/Rabkinesin-6/MKLP2(有丝分裂末期/胞质分裂期间定位于中体/midbody,结果未显示)。而另外5个驱动蛋白究竟是否在有丝分裂中起调控作用,则有待进一步的实验生物学验证。

2.3 FKPP的实验验证

我们之前的研究发现,人源CENP-E蛋白是一个有丝分裂驱动蛋白,在有丝分裂前中期和中期定位于动点^[6]。我们近期的工作表明,作为一个分子马达,CENP-E蛋白调控染色体在赤道板处的聚集^[9,10]。令人惊讶的是,FKPP的预测结果表明,全长的人源CENP-E蛋白应当是包含2701个残基,与之前所报道的序列相比,多38个氨基酸残基^[6]。事实上,实验结果表明,在昆虫细胞中重组的全长(人)CENP-E大约比人细胞中内源CENP-E蛋白小5 kD,这说明通过表达型文库克隆得到的CENP-E可能是不完整的。图3(a)所示为通过计算得到的在之前的研究工作中不存在的人源CENP-E的部分序列^[6]。为了验证FKPP的准确性,我们制备了多肽抗体。如图3(b)所示,本研究新制备的多肽抗体以及之前CENP-E的抗体HpX^[7],可以将CENP-E免疫沉淀下来。然而,作为对照的兔IgG则不能免疫沉淀俘获内源性的CENP-E^[6]。本研究的三色免疫荧光技术结果表明,CENP-E蛋白的鼠抗177和多肽抗体都识别有丝分裂期的动点,且二者的染色完全共定位。此外,对从人的睾丸cDNA文库中得到的CENP-E cDNA进行序列测定,测序的结果与计算结果一致。因此,人中的CENP-E蛋白应包含2701个氨基酸残基。通过此3个不同的实验测试,验证了FKPP是一个有效的、准确的预测全长驱动蛋白的工具。

表 2 人驱动蛋白的表达谱分析^{a)}

驱动蛋白名称	蛋白标示符	蛋白长度 (aa)	组织表达特异性 (GeneCard)	基因表达谱(UniGene/dbEST)
KIF1A	GI: 2497523 ^{b)}	1690	神经特异性(>8~10 倍)	神经, 胰腺, 骨髓
KIF1B/KLP	O60333	1816	广泛表达	广泛表达
KIF1C	O43896	1103	广泛表达	广泛表达
KIF10/CENP-E	GI: 399227	2701	免疫特异性(>8~10 倍)	免疫, 肝脏, 肾脏, 肺
KIF14	Q15058	1648	免疫特异性(>8~10 倍)	免疫, 肝脏, 肾脏
KIFC1	Q9BW19	673	免疫特异性(>8~10 倍)	免疫, 肌肉, 肝脏, 胰腺
KIFC2	Q96AC6	838	神经特异性(>8~10 倍)	神经
KIFC3	Q9BVG8	694	广泛表达	广泛表达
KIF3B	O15066	747	广泛表达	广泛表达
KIF3A	Q9Y496	702	广泛表达	广泛表达
KIF3C	O14782	793	神经特异性(>8~10 倍)	神经特异性(>8~10 倍)
KIF4A	O95239	1232	广泛表达	广泛表达
KIF4B	GI: 41147002	1234	/	睾丸(小鼠中)/Hs.529460
KIF13A	Q9H1H9	1805	广泛表达	广泛表达
KIF13B	Q9NQT8	1826	广泛表达	广泛表达
KIF27A	Q86VH2	1401	广泛表达	肌肉, 胰腺, 肾脏
KIF5A	Q12840	1032	神经特异性(>8~10 倍)	脑, 肌肉, 肺
KIF5B	P33176	963	/	神经特异性(>8~10 倍)
KIF5C	O60282	957	神经和前列腺特异性(>8~10 倍)	广泛表达
KIF17	Q9P2E2	1029	广泛表达	脾, 脑
KIF11/EG5	P52732	1057	免疫特异性(>8~10 倍)	肌肉, 肝脏, 肺
KIF9	Q9HAQ2	790	广泛表达	广泛表达
KIF22	Q14807	665	广泛表达	广泛表达
KIF25	Q9UIL4	384	广泛表达	胎盘, 神经/Hs.150013
KIF20A/MKLP2	O95235	890	免疫特异性(>8~10 倍)	广泛表达
KIF2A/KIF2	O00139	679	免疫和神经特异性(>5~8 倍)	免疫特异性(>8~10 倍)
KIF2B	Q8N4N8	673	广泛表达	骨髓, 睾丸/Hs.226805
KIF2C/MCAK	Q99661	725	免疫特异性	免疫, 脑, 肌肉, 肝脏
KIF23/MKLP1	Q02241	856	广泛表达	广泛表达
KIF18A	Q8NI77	898	免疫特异性(>8~10 倍)	睾丸, 胃/Hs.301052
KIF18B	37544008	870	免疫特异性(>8~10 倍)	卵巢, 脑, 膀胱 /Hs.406639
KIF12	Q96FN5	618	肌肉、分泌组织和肾脏特异性(>5~8 倍)	广泛表达
KIF20B	预测结果 ^{c)}	1897	免疫特异性(>8~10 倍)	睾丸, 肝脏/Hs.240
KIF15/HKLP2	Q9NS87	1388		睾丸, 肝脏/Hs.315051
KIF16A	GI: 41204881	846		
KIF16B	Q9HCI2	1393	广泛表达	海马区, 肾脏, 前列腺/Hs.101774
KIF21A	AAR04774	1674		肺, 肾脏, 肝脏/Hs.374201
KIF21B	GI: 41112866	1635	免疫和神经特异性(>3~5 倍)	胃, 肾脏/Hs.169182
KIF19A	Q8N1X8	894	广泛表达	肝脏/Hs.372773
KIF26A	Q9ULI4	1840	广泛表达	胃/Hs.134970
KIF26B	GI: 41114119	1726	广泛表达	免疫特异性(>8~10 倍) /Hs.125020
KIF24	预测结果	1256		肺, 肝脏 /Hs.436169
KIF6	ENSP00000287152 ^{d)}	484		
KIF7	预测结果	1343		肝脏, 胃 /Hs.528406

a) 用 GeneCard 数据库中的表达数据, 分析人驱动蛋白的组织表达特异性, 并通过 UniGene 数据库以及相应的 ESTs 数据对基因表达情况进行了分析. 蛋白标示主要采用了 Swiss-Prot 的标示规则; b) 在无 Swiss-Prot 蛋白标示符时, 则采用 GenBank 中的 GI 标示符; c) 预测得到的驱动蛋白, 无标示符; d) 从 Ensembl 数据中得到的预测结果

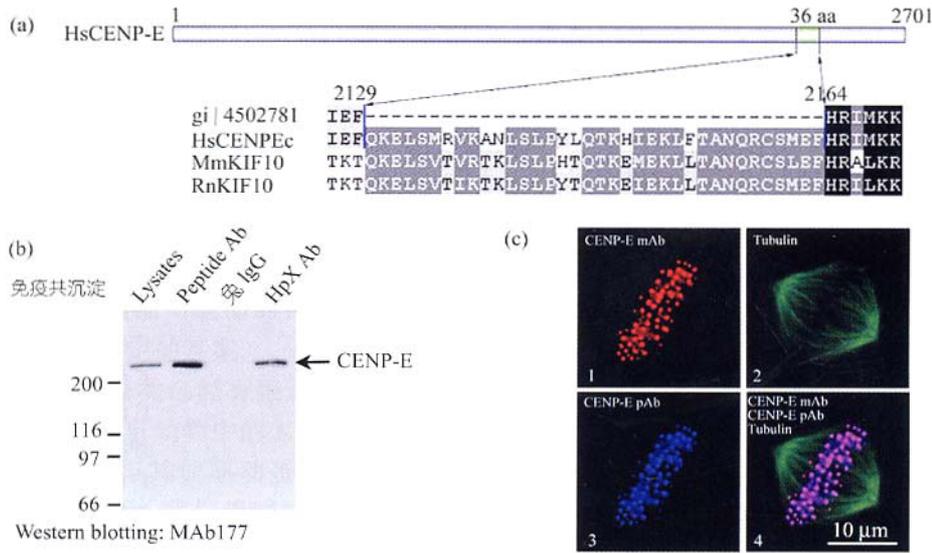


图3 FKPP发现最初克隆的人源CENP-E序列缺失了一个含36个氨基酸的片段

(a) 用FKPP方法发现最初克隆出的人CENP-E缺失了36个氨基酸残基的一段。本研究给出了这段缺失片段(2129~2164)的序列比对;(b)通过FKPP得到的人CENP-E蛋白全长序列。分别将预先结合了HpX抗体(HpX Ab)、多肽抗体(根据FKPP预测到的CENP-E片段序列而制备)和兔IgG蛋白的蛋白质A/G小球与有丝分裂期的HeLa细胞的裂解液进行孵育。洗涤之后,用小鼠的单抗177(mAb177)进行免疫印迹分析来鉴定免疫沉淀俘获物。小鼠的mAb177是最初鉴定CENP-E蛋白的抗体。(c)计算生物学预测的人CENP-E蛋白是一个动点结合的驱动蛋白。1,用mAb177标记的CENP-E在有丝分裂期的HeLa细胞中的定位;2,用大鼠的微管抗体标记的纺锤体;3,用本研究制备的多肽抗体标记的CENP-E的定位;4,1~3的叠加。图片采用激光共聚焦显微镜采集

3 讨论

相对于从头预测来说,比较基因组学是同源基因预测、全基因组比对以及调控区域预测的一个有效的工具^[26]。尽管在目前的公共数据库中许多物种的基因组已经得到了测序,但是单个基因的结构与功能的关系仍然需要实验验证。许多蛋白仍然还是只有片段或者部分的序列。这里,我们对Swiss-Prot/TrEMBL数据库中的蛋白片段进行了粗略的统计。首先,用物种的名称作为关键词,如“Homo sapiens”或者“Mus musculus”等,以查找每个物种究竟有多少条蛋白序列。然后,再在物种名称后加上“fragment”,成为一个复合的关键词(如“Homo sapiens fragment”),查找每个物种中所有的片段序列。结果见表3。在人、小鼠和大鼠中,分别大约有32.2%,24.6%和20.6%的蛋白是片段序列。即使是在芽殖酵母(budding yeast)中,仍然有2.8%的蛋白是片段序列。

因此,准确的计算生物学方法提供了一种高通量的策略,对于实验学家开展下一步的实验工作具有重要的意义。FKPP是一种新颖的、简单有效的方法,可以将数据库中的信息尽可能地抽提出来,便于预测和鉴定新的哺乳动物的驱动蛋白。另外,这个方

表3 目前数据库中各个物种的片段序列情况

Swiss-Prot & TrEMBL	片段序列	总数	百分比
<i>Homo sapiens</i>	22779	70849	32.2%
<i>Mus musculus</i>	12136	49292	24.6%
<i>Rattus norvegicus</i>	2830	13716	20.6%
<i>Xenopus laevis</i>	2124	12182	17.4%
<i>Arabidopsis thaliana</i>	3017	43370	7.0%
<i>Drosophila melanogaster</i>	3119	28204	11.1%
<i>Caenorhabditis elegans</i>	352	22910	1.5%
<i>Schizosaccharomyces pombe</i>	273	5577	4.9%
<i>Saccharomyces cerevisiae</i>	414	14551	2.8%

法也可以很容易地与其他基因预测方法相结合,从而提高基因预测的准确性。由于所使用的工具和数据库的限制,我们的预测分析仍然可能有一些错误或者不准确,这需要进一步的优化。并且,BLAT使用的是标准的剪切子序列(standard splice site),如果一个基因使用的是非标准的剪切子序列,则基因的结构在预测过程中会被破坏。所幸的是,这样的事件在我们的驱动蛋白的鉴定工作中并未发现。

FKPP分析也提示可能存在着物种特异性的驱动蛋白。例如,小鼠的MmKIFC5C,大鼠的RnKLP-6和人的HsKIF25,这些基因/蛋白在其他物种中都没有找到对应的直向同源物。这些物种特异性的驱动

蛋白对于这 3 个物种在进化中的区别是否有所贡献? 这个问题仍然需要实验生物学的验证, 以观察这些驱动蛋白是否具有特定的生物学功能。

综上所述, FKPP 分析法为研究哺乳动物驱动蛋白超家族在细胞动力学过程中的机理与功能提供了一个坚实的理论基础. 驱动蛋白亚家族中的每一个成员都具有一些独特的结构特性, 而这些特性表明在微观尺度上驱动蛋白亚家族的分子运动机理及能量决定因素都有所不同. 尽管如此, 通过对驱动蛋白进行亚家族的分类, 本研究提供了一个统一的关于驱动蛋白的马达端在分子水平上的工作机制. 另外, FKPP 也提供了一个通用且简单的方法, 其意义并不局限于驱动蛋白的鉴定. 由于目前很多蛋白仍然只存在片段序列, FKPP 对于通过在几个相似的物种间(如人、小鼠、大鼠之间)作计算比较基因组学分析, 可以获得可能的全长序列. 显然, FKPP 分析与实验生物学结合将有助于阐明马达蛋白家族的系统生物学功能及其功能的结构基础.

致谢 本工作为国家自然科学基金(批准号: 39925018, 30270293 和 30121001)、中国科学院知识创新工程(批准号: KSCX2-2-01)、国家重点基础研究发展计划(批准号: 2002CB713700)、国家高新技术研究发展规划(批准号: 2001AA215331)以及教育部博士点基金(批准号: 20020358051)资助项目。

参 考 文 献

- 1 Brenda R P, Serbus L R, Duffy J B, et al. A function for kinesin I in the posterior transport of oskar mRNA and Staufen protein. *Science*, 2000, 289(5487): 2120—2122
- 2 Cleveland D W, Mao Y, Sullivan K F. Centromeres and kinetochores: From epigenetics to mitotic checkpoint signaling. *Cell*, 2003, 112(4): 407—421
- 3 Guzik B W, Goldstein L S. Microtubule-dependent transport in neurons: Steps towards an understanding of regulation, function and dysfunction. *Curr Opin Cell Biol*, 2004, 16: 443—450
- 4 Hirokawa N, Takemura R. Molecular motors and mechanisms of directional transport in neurons. *Nat Rev Neurosci*, 2005, 6: 201—214
- 5 McIntosh J R, Grishchuk E L, West R R. Chromosome-microtubule interactions during mitosis. *Annu Rev Cell Dev Biol*, 2002, 18: 193—219
- 6 Yen T J, Li G, Schaar B T, et al. CENP-E is a putative kinetochore motor that accumulates just before mitosis. *Nature*, 1992, 359: 536—539
- 7 Abrieu A, Kahana J A, Wood K W, et al. CENP-E as an essential component of the mitotic checkpoint *in vitro*. *Cell*, 2000, 102: 817—826
- 8 Hirokawa N, Takemura R. Kinesin superfamily proteins and their various functions and dynamics. *Exp Cell Res*, 2004, 301: 50—59
- 9 Yao X, Abrieu A, Zheng Y, et al. CENP-E forms a link between

- attachment of spindle microtubules to kinetochores and the mitotic checkpoint. *Nat Cell Biol*, 2000, 2: 484—491
- 10 Yao X, Anderson K L, Cleveland D W. The microtubule-dependent motor centromere-associated protein E (CENP-E) is an integral component of kinetochore corona fibers that link centromeres to spindle microtubules. *J Cell Biol*, 1997, 139: 435—447
- 11 Vale R D, Fletterick R J. The design plan of kinesin motors. *Annu Rev Cell Dev Biol*, 1997, 13, 745—777
- 12 Miki H, Setou M, Hirokawa N. Kinesin superfamily proteins (KIFs) in the mouse transcriptome. *Genome Res*, 2003, 13: 1455—1465
- 13 Miki H, Setou M, Kaneshiro K, et al. All kinesin superfamily protein, KIF, genes in mouse and human. *Proc Natl Acad Sci USA*, 2001, 98: 7004—7011
- 14 Nakagawa T, Tanaka Y, Matsuoka E, et al. Identification and classification of 16 new kinesin superfamily (KIF) proteins in mouse genome. *Proc Natl Acad Sci USA*, 1997, 94: 9654—9659
- 15 Kent W J. BLAT: The BLAST-like alignment tool. *Genome Res*, 2002, 12: 656—664
- 16 Thompson J D, Gibson T J, Plewniak F, et al. The CLUSTAL_X windows interface: Flexible strategies for multiple sequence alignment aided by quality analysis tools. *Nucleic Acids Res*, 1997, 25: 4876—4882
- 17 Taylor M S, Ponting C P, Copley R R. Occurrence and consequences of coding sequence insertions and deletions in mammalian genomes. *Genome Res*, 2004, 14: 555—566
- 18 Kumar S, Tamura K, Jakobsen I B, et al. MEGA2: Molecular evolutionary genetics analysis software. *Bioinformatics*, 2001, 17: 1244—1245
- 19 Lawrence C J, Dawe R K, Christie K R, et al. A standardized kinesin nomenclature. *J Cell Biol*, 2004, 167: 19—22
- 20 Mulder N J, Apweiler R, Attwood T K, et al. The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res*, 2003, 31: 315—318
- 21 Garcia M A, Koonrugsa N, Toda T. Spindle-kinetochore attachment requires the combined action of Kin I-like Klp5/6 and Alp14/Dis1-MAPs in fission yeast. *Embo J*, 2002, 21: 6015—6024
- 22 Schuler G D. Pieces of the puzzle: Expressed sequence tags and the catalog of human genes. *J Mol Med*, 1997, 75: 694—698
- 23 Safran M, Solomon I, Shmueli O, et al. GeneCards 2002: Towards a complete, object-oriented, human gene compendium. *Bioinformatics*, 2002, 18: 1542—1543
- 24 Wordeman L, Mitchison T J. Identification and partial characterization of mitotic centromere-associated kinesin, a kinesin-related protein that associates with centromeres during mitosis. *J Cell Biol*, 1995, 128: 95—104
- 25 Blangy A, Lane H A, d'Herin P, et al. Phosphorylation by p34cdc2 regulates spindle association of human Eg5, a kinesin-related motor essential for bipolar spindle formation *in vivo*. *Cell*, 1995, 83: 1159—1169
- 26 Ureta-Vidal A, Ettwiller L, Birney E. Comparative genomics: genome-wide analysis in metazoan eukaryotes. *Nat Rev Genet*, 2003, 4: 251—262

(2006-02-20 收稿, 2006-06-02 接受)