动物科学

DOI:CNKI:50-1165/N.20110516.1013.174

黑腹果蝇热激蛋白基因家族的生物信息学分析*

牟国鹏,陈 斌,何正波
(重庆师范大学 生命科学学院 昆虫与分子生物学研究所 活性物质生物技术教育部工程研究中心
重庆市动物生物学重点实验室,重庆 400047)

摘要:热激蛋白是细胞或生物体受到热激后新合成的一类遗传上高度保守的蛋白,在生物体中普遍存在,在细胞生长、发育、分化、基因转录等功能方面发挥重要的作用。本研究利用生物信息学方法首次对黑腹果蝇(*Drosophila melanogaster*)全部热激蛋白进行分析。结果表明,黑腹果蝇热激蛋白共有 HSPC(HSP90)、HSPA(HSP70)、DNAJ(HSP40)、HSPB(小分子 HSP)、HSPD(HSP60)、CCT(TRiC)等 6 个基因家族,不存在 HSPE(HSP10)和 HSPH(HSP110)家族,共有 35 个基因,新发现 7 个基因,命名为 *HSPBA*5、*HSPBB*2、*HSPBB*3、*DNAJA*7、*DNAJA*8、*HSPD*5 和 *HSPCB*1。染色体定位表明,黑腹果蝇热激蛋白主要分布在 2L、3L、2R、3R 和 X 上。序列比对分析发现,黑腹果蝇热激蛋白家族序列高度保守。进化分析表明,他们有共同的祖先,可分为 6 个基因家族。

关键词:黑腹果蝇;热激蛋白;生物信息学;系统发育
中图分类号:Q811.4;Q963 文献标识码:A 文章编号:1672-6693(2011)03-0010-08

热激蛋白(Heat shock proteins,HSPs)是细胞或生物体在一定时间内受高温热激或其它应激原作用下诱导新合成或含量增加的一类蛋白质。HSPs 在分子进化上高度保守,并广泛分布于微生物、植物、动物等各种生物体内,HSPs 在细胞生长、发育、分化、基因转录等功能方面发挥重要的作用[1-2]。近年来的研究表明 HSPs 还参与某些病毒蛋白的成熟。自 1982 年在美国冷泉港召开第一届 HSPs 的国际会议后,对 HSPs 的研究蓬勃发展。中国对 HSPs 的研究始于 20 世纪 80 年代中期,研究对象以人类、哺乳动物和高等植物为主。

果蝇作为模式生物已广泛应用于发育生物学、分子生物学、人类疾病研究等领域。在利用果蝇模型研究的人类疾病中,较多的有神经退行性疾病,包括帕金森病、阿尔兹海默病、多聚谷氨酰胺病等[3-4]。果蝇基因组序列分析表明,超过 60% 的人类疾病基因在果蝇中有直系同源基因[5]。对于果蝇 HSPs 尚无明确的分类标准。王秋香等[6]根据 HSPs 同源程度及分子量大小,将其分为 HSP110、HSP90、HSP70、HSP60、HSP40、小分子 HSP 等 6 个家族。HSP90 家族的第一个成员是在果蝇中发现的,位于胞浆中,其基因位于染色体 3L。HSP70 家族包括诱导性成员 *HSP*68、*HSP*70,分布在果蝇的胞浆和内质网中,有两个主要的结构域,N-端高度保守的 ATPase 功能域和 C-端区域。HSP60 家族,包括 4 个基因 *HSP*60*A*、*HSP*60*B*、*HSP*60*C* 和 *HSP*60*D*。HSP40 家族在野生型果蝇中不存在[7]。小分子 HSPs 家族包括 *HSP*22、*HSP*23、*HSP*26 和 *HSP*27。*HSP*22 分布在线粒体基质内,*HSP*27 分布在核内,*HSP*26、*HSP*23 分布在基质内[8-9]。目前,对于黑腹果蝇(*Drosophila melanogaster*)HSPs 的研究主要集中在小分子 HSP 家族及 HSP70 家族[10-12],对于黑腹果蝇热激蛋白的家族成员的命名及分类比较混乱,尚未见有系统的对黑腹果蝇基因组中全部热激蛋白进行生物信息学的分析,而对人热激蛋白的研究较多并有系统的分类与命名[13-16]。本研究通过生物信息学方法,对该基因家族的成员进行系统分析,包括蛋白序列分析、系统发育分析和蛋白质的结构域分析、基因的染色体定位、基因结构分析等,希望对后续研究该热激蛋白各基因家族成员的基因功能提供参考资料。

* 收稿日期:2011-03-31 网络出版时间:2011-05-16 10:13:47
资助项目:国家自然科学基金(No. 30870340;No. 31071968);重庆市自然科学基金重点项目(No. CSTC2008BA5030);重庆市教委科学技术研究项目(No. KJ100620)
作者简介:牟国鹏,男,硕士研究生,研究方向为生物信息学;通讯作者:陈斌,E-mail:c_bin@hotmail.com

1 材料和方法

1.1 数据的获得及数据库检索

从 NCBI 数据库和已公布序列中检索获得黑腹果蝇热激蛋白各基因家族代表基因的氨基酸序列及相应的核苷酸序列作为询问序列(表 1),检索 NCBI 数据库种的相似性序列。为尽可能获得全部的热激蛋白家族成员,利用已获得氨基酸序列用 PSI-BLAST 程序在 NCBI 数据库中检索该家族成员氨基酸序列。利用已获得的核苷酸序列用 BLASTn 程序在 NCBI 数据库中检索该家族成员的核苷酸序列,检索均采用默认参数设置,选取序列长度不小于 200 bp,序列相似度不小于 95%,$E \leqslant 10^{-10}$的序列为候选序列。每次检索低于该参数设置的序列被排除,符合参数的序列用 Clustal X 程序进行一一比对,完整的具有代表性的序列数据将用于系统发育分析。同时,在 FlyBase 数据库中检索已注册的黑腹果蝇热激蛋白的相似序列。下载获得的全部黑腹果蝇热激蛋白的核苷酸序列、氨基酸序列及相应的基因组序列。

表 1　各基因家族代表性基因序列(询问序列)的登录号

Tab. 1　The NCBI Genbank accession numbers of query sequences

基因(基因家族)	核苷酸序列	氨基酸序列
*HSP*23(HSPB)	NM_079275.2	NP_523999.1
*HSP*40(DNAJ)	NM_079212.2	NP_523936.2
*HSP*60(HSPD)	NM_078560.2	NP_511115.2
*HSP*68(HSPA)	NM_079750.3	NP_524474.1
*CCT*1(CCT)	NM_078983.2	NP_523707.1
*HSP*83(HSPC)	NM_079175.2	NP_523899.1

1.2 序列分析

在获得黑腹果蝇热激蛋白家族的所有核苷酸序列和氨基酸序列后,利用 Clustal X 软件中默认的参数设置对该家族的氨基酸序列进行多序列联配。在多序列联配的保守区段,使用 BioEdit 对联配结果进行编辑。蛋白质序列的相似性、氨基酸长度、分子量大小计算使用 BioEdit 程序。利用 NCBI 的 Map Viewer 对检索到的序列进行染色体定位,对位于染色体同一位置的 mRNA 序列和编码序列被认为来自于同一个基因,而与检索序列和基因组数据最相配的序列被选择作为该基因的代表基因。利用在线程序 BLAT 和 Entrez 工具检索序列外显子、内含子位置,并记录外显子的数量。同时获得该基因家族成员的 CDS 序列和对应的全基因组序列,利用 Gene Structure Display Server (GSDS)绘出基因结构图。信号肽预测使用工具 Signal P,预测中同时采用可选神经网络法(Neural network,NN)和隐马可夫模型(Hidden Markov models,HMM),其他参数采用默认设置。蛋白质序列的结构域及功能基序(Motifs)检查和预测同时使用了两种在线工具,ScanProsite 预测蛋白质的基因家族和结构域,InterProScan 预测蛋白质结构域和功能基序[13]。经过检验,所得序列中含有热休克蛋白保守结构与或功能基序的认为属于该基因家族。用于生物信息学研究的数据库和软件网址见表 2。

表 2　用于生物信息学研究的数据库和软件网址

Tab. 2　The website and software used in the study

软件或数据库	网址	版本
NCBI	http://www.ncbi.nlm.nih.gov/	
FlyBase	http://flybase.org/	
BLAST	http://blast.ncbi.nlm.nih.gov/Blast.cgi	V2.2.24
BLAT	http://genome.ucsc.edu/cgi-bin/hgBlat	
Signal P	http://www.cbs.dtu.dk/services/SignalP/	V3.0
ScanProsite	http://www.expasy.ch/tools/scanprosite/	
InterProScan	http://www.ebi.ac.uk/Tools/InterProScan/	
GSDS	http://gsds.cbi.pku.edu.cn/	
BioEdit	http://www.mbio.ncsu.edu/bioedit/bioedit.html	V7.0
Clustal X	http://www.clustal.org/	V2.0
Phylip	http://cmgm.stanford.edu/phylip/	V3.68

1.3　系统发育分析

利用 Clustal X 软件对黑腹果蝇热激蛋白家族的氨基酸序列进行多序列比对并进行人工校对,用软件 Phylip 构建系统发育树。构建最大简约树(Maximum parsimony,MP),并采用 1 000 次可置信度分析(Bootstrap)。外群选择大肠杆菌 *HTPG*(登录号:AAA23460)。

2 结果

2.1 黑腹果蝇热激蛋白家族成员及染色体分布

利用已获得的询问序列,在 NCBI 数据库中用 PSI-BLAST 及 BLASTn 程序检索黑腹果蝇热激蛋白家族成员。分别利用 PSI-BLAST 和 BLASTn 程序在 NCBI 公布的黑腹果蝇基因组数据库中检索该家族成员的氨基酸序列和核酸序列。通过检索,得到蛋白质序列 410 个和核苷酸序列 406 个。在剔除冗余序列后,经过序列比对并进行序列保守结构域的比对检查,获得黑腹果蝇热激蛋白家族成员序列 94 个,mRNA 序列 94 个。在找到所有的黑腹果蝇热激蛋白家族成员后,进行了归类及重命名,命名采

用HUGO基因命名委员会及人基因组热激蛋白的命名及归类方法[16]。本研究发现新基因7个,其中HSPB家族发现3个,分别为*HSPBA*5、*HSPBB*2、*HSPBB*3。DNAJ家族发现新基因2个,为*DNAJA*7、*DNAJA*8。在HSPD及HSPC家族中分别发现1个新基因,命名为*HSPD*5及*HSPCB*1。将黑腹果蝇热激蛋白分为6个基因家族,分别为HSPB(分子量12.6~46.9 kDa,小分子热激蛋白家族)、DNAJ(分子量24.7~56.1 kDa,HSP40家族)、HSPD(分子量57.8~68.6 kDa,HSP60家族)、CCT(分子量55.7~59.5 kDa,TRiC)家族、HSPA(分子量61.3~74.2 kDa,HSP70家族)和HSPC(分子量81.8~91.2 kDa,HSP90家族),然而在本研究中没有发现HSPE(HSP10)和HSPH(HSP110)家族。其中HSPB分为两个亚族即HSPBA、HSPBB,分别有4个和3个家族成员。DNAJ家族分3个亚族即DNAJA、DNAJB和DNAJC,分别有8、1、1个家族成员。HSPD、CCT、HSPA均仅有1个亚族,分别有5、3、8个家族成员。而HSPC家族有两个亚族即HSPCA、HSPCB,每个亚族仅有1个家族成员。利用Signal P进行信号肽预测分析,在*DNAJA*5、*CCT*3、*HSPA*4、*HSPCB*1中分别发现长度为21、18、18、20个氨基酸长度的信号肽(表3)。

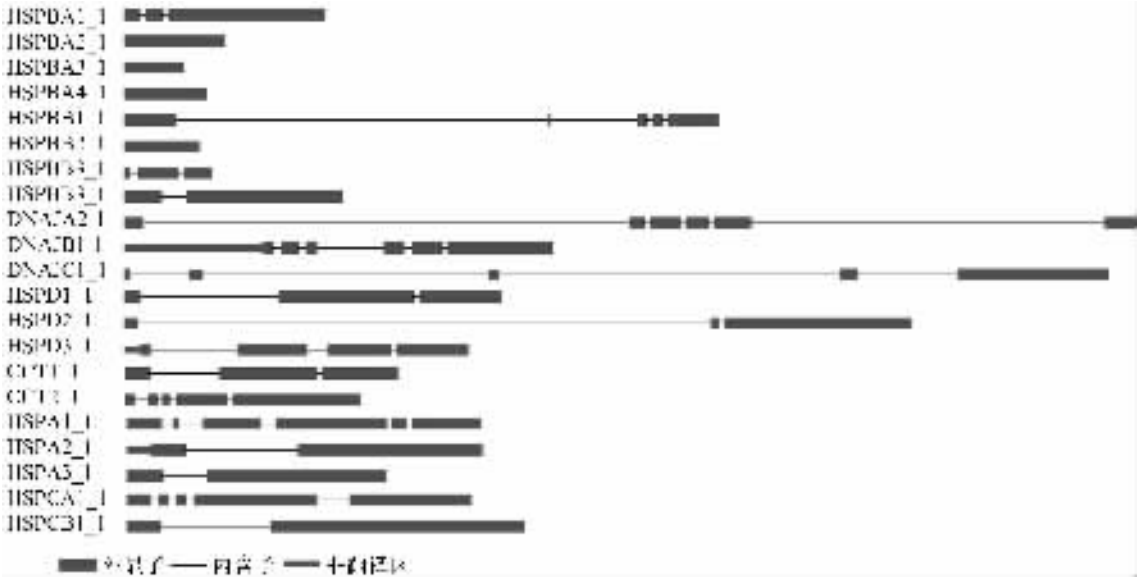

图1 黑腹果蝇热激蛋白家族基因结构图

Fig. 1 The gene structure of *Drosophila melanogaster* HSPs

从染色体分布来看,黑腹果蝇热激蛋白主要分布在染色体2L、3L、2R、3R和X上。其中,HSPB家族基因分布主要集中在染色体3L上,染色体2R及X上有少量分布,其中仅新发现基因*HSPBB*2位于染色体X上。DNAJ家族主要分布在染色体2L、3L、2R和3R。HSPD家族主要分布在染色体X、2L上。CCT家族主要分布在染色体2R、3R上。HSPA家族主要分布在染色体3L、2R和3R上,染色体X上仅有1个基因。HSPC家族分布在两条染色体上,分别为3R、3L上(表3)。

在黑腹果蝇热激蛋白HSPB家族中大部分成员均只有1个外显子,没有内含子存在,而在*HSPBA*1,*HSPBB*1、*HSPBB*2中分别存在多个外显子。在DNAJ、HSPD、CCT、HSPA、HSPC家族中,该家族基因均含有多个外显子,其中在*HSPA*6、*HSPA*7、*HSPA*8中仅有1个外显子,并且没有内含子存在(图1)。黑腹果蝇热激蛋白家族成员外显子具体数量见表3。

2.2 黑腹果蝇中热激蛋白家族成员结构域分析

将黑腹果蝇热激蛋白家族16个代表基因氨基酸序列整合为FASTA格式,利用Clustal X进行同源序列比对,序列比对的结果见封三彩图2。黑腹果蝇热激蛋白家族成员的蛋白质结构域及功能基序预测同时使用了两种在线工具ScanProsite和InterProScan。在HSPB家族中均含有约100个氨基酸的C-端保守结构域(ScanProsite中记录号为PS01031),该区域与χ-晶体蛋白有很高的同源性,也称为χ-晶体蛋白区域。而在HSPBA2中除HSP20保守结构域外,还预测到1个TonB-dependent receptor的保守序列标签(ScanProsite中记录号为

PS00430),由序列比对分析看出该结构域相对保守。DNAJ 家族也被称为 J 蛋白家族,DNAJ 家族成员与 HSPA/DnaK 蛋白协作,促进 ATP 向 ADP 的水解,从而帮助 HSP70 伴侣蛋白锁定其所结合的蛋白质底物。家族成员均具有 DNAJ 结构,主要包括 1 个N-端约 70 个氨基酸的保守区(J-结构域)、1 个富 G/F 结构域、1 个类锌指结构域和 1 个 C-端结构域。同时该家族成员均含有 1 个 20 个氨基酸的保守序列标签(ScanProsite 中记录号为 PS00636)。在 HSPD 及 CCT 家族中,HSPD 中均包括 1 个 12 个氨基酸的保守序列标签 CPN60(ScanProsite 中记录号为 PS00296)。在 CCT 中均包含 3 个保守序列标签,分别为 TCP1_1(ScanProsite 中记录号为 PS00750)、TCP1_2(ScanProsite 中记录号为 PS00751)和 TCP1_3(ScanProsite 中记录号为 PS00995),而在 CCT1 中缺少 TCP1_2。HSPA 家族成员均含有两个主要的结构域,N-端高度保守的 ATPase 功能域(ATP-binding domain)和 1 个 C-端区域组成,含有保守序列标签 HSP70_1(ScanProsite 中记录号为 PS00297)、HSP70_2(ScanProsite 中记录号为 PS00329)、HSP70_3(ScanProsite 中记录号为 PS01036),其中 HSP70_1 位于 N-端 ATPase 功能域。HSPC 家族有两个结构域,1 个 N-端保守区和 1 个 C-端结构域,而 C-端最末端 4 个氨基酸 EEVD 是所有真核细胞 HSP90 的特征标签,而在果蝇体内只有 1 种 HSPC 家族蛋白 HSP83。而 HSPC 家族成员 HSPCB 位于内质网上,N-端有 1 段信号肽,C-端为 KDEL。在 HSPC 中,含有保守序列标签 HSP90(ScanProsite 中记录号为 PS00298)。而在 HSPCA 中单独含有保守序列标签 ER_TARGET(ScanProsite 中记录号为 PS00014)。

表 3　黑腹果蝇热激蛋白家族:GenBank 登录号、染色体位置、外显子、氨基酸长度、分子量、信号肽及基因功能

Tab. 3　The *D. melanogaster* HSP families of genes: Genbank accession number, chromosomal location, exon number, amino acid length, mass, signal peptide and gene function

基因家族	基因	基因变体		登录号		染色体定位[a]	外显子数	氨基酸长度	分子量/kDa	信号肽[b]	基因功能[c]
		基因剪切(新命名)	旧名称	核苷酸	氨基酸						
HSPC	**HSPCA**										
(HSP90)	*HSPCA*1	*HSPCA*1-1	heat shock protein 83	NM_079175.2	NP_523899.1	3L+(63B)	2	717	81.861	N	F
		*HSPCA*1-2	hsp83	X00065.1	CAA24938.1	3L+(63B)	2	375	42.931	N	P
		*HSPCA*1-3	hsp 82	X03810.1	CAA27435.1	3L+(63B)	2	717	81.861	N	F
		*HSPCA*1-4	AT20544p	AY122080.1	AAM52592.1	3L+(63B)	2	717	81.861	N	F
	HSPCB										
	*HSPCB*1										
		*HSPCB*1-1	glycoprotein 93	NM_143344.2	NP_651601.1	3R+(98B)	5	787	90.233	Y(20)	F
		*HSPCB*1-2	IP13374p	BT025090.1	ABE73261.1	3R+(98B)		787	90.233	Y(20)	F
		*HSPCB*1-3	LD23641p	AY075384.1	AAL68222.1	3R+(98B)	5	787	90.233	Y(20)	F
HSPA	**HSPA**										
(HSP70)	*HSPA*1	*HSPA*1-1	heat shock protein cognate 5	NM_079017.2	NP_523741.2	2r-(50E)	6	686	74.062	N	F
		*HSPA*1-2	heat shock protein cognate 71	L01502.1	AAA28628.1	2r-(50E)	6	687	74.204	N	F
	*HSPA*2	*HSPA*2-1	heat shock protein cognate 1, isoform A	NM_079339.2	NP_524063.1	3L+(70C)	2	641	70.683	N	F
		*HSPA*2-2	heat shock protein cognate 1, isoform C	NM_168568.2	NP_729940.1	3L+(70C)	3	555	61.344	N	F
		*HSPA*2-3	heat shock protein cognate 1, isoform D	NM_168569.1	NP_729941.1	3L+(70C)	2	555	61.344	N	F
		*HSPA*2-4	heat shock protein cognate 70	L01501.1	AAA28625.1	3L+(70C)	6	641	70.682	N	F
		*HSPA*2-5	AT28834p	BT001361.1	AAN71116.1	3L+(70C)	6	680	75.182	N	F
	*HSPA*3	*HSPA*3-1	heat shock protein cognate 2	NM_079615.3	NP_524339.1	3R+(87D)	2	633	69.718	N	F
	*HSPA*4	*HSPA*4-1	heat shock protein cognate 3, isoform A	NM_167306.1	NP_727563.1	X-(10E)	3	656	72.257	Y(18)	F
		*HSPA*4-2	heat shock protein cognate 3, isoform B	NM_078577.2	NP_511132.2	X-(10E)	3	656	72.257	Y(18)	F
		*HSPA*4-3	heat shock protein cognate 3, isoform C	NM_167307.1	NP_727564.1	X-(10E)	3	656	72.257	Y(18)	F
		*HSPA*4-4	heat shock protein cognate 3, isoform D	NM_167308.1	NP_727565.1	X-(10E)	3	656	72.257	Y(18)	F
		*HSPA*4-5	RH21402p	BT004838.1	AAO45194.1	X-(10E)	6	655	72.109	Y(18)	F

续表3

基因家族	基因	基因变体		登录号		染色体定位[a]	外显子数	氨基酸长度	分子量/kDa	信号肽[b]	基因功能[c]
		基因剪切(新命名)	旧名称	核苷酸	氨基酸						
	HSPA5	*HSPA5*-1	heat shock protein cognate 4, isoform A	NM_079632.5	NP_524356.1	3R+(88E)	2	651	71.128	N	F
		HSPA5-2	heat shock protein cognate 70	L01500.1	AAB59186.1	3R+(88E)	7	651	71.128	N	F
		HSPA5-3	RH04426p	AY084193.1	AAL89931.1	3R+(88E)	7	651	71.1	N	F
		HSPA5-4	heat shock protein cognate 4, isoform B	NM_169625.2	NP_731987.1	3R+(88E)	2	651	71.128	N	F
		HSPA5-5	heat shock protein cognate 4, isoform C	NM_169626.2	NP_731988.1	3R+(88E)	2	651	71.128	N	F
		HSPA5-6	heat shock protein cognate 4, isoform D	NM_169627.2	NP_731989.1	3R+(88E)	2	651	71.128	N	F
		HSPA5-7	heat shock protein cognate 4, isoform E	NM_176502.2	NP_788679.1	3R+(88E)	2	651	71.128	N	F
		HSPA5-8	heat shock protein cognate 4, isoform F	NM_176503.2	NP_788680.1	3R+(88E)	1	651	71.128	N	F
	HSPA6	*HSPA6*-1	heat shock protein 68	NM_079750.3	NP_524474.1	3R-(95D)	1	635	69.74	N	F
	HSPA7	*HSPA7*-1	Heat-shock-protein-70Aa	NM_169441.1	NP_731651.1	3R+(87A)	1	642	70.156	N	F
		HSPA7-2	Heat-shock-protein-70Ab	NM_080059.2	NP_524798.2	3R+(87A)	1	642	70.156	N	F
	HSPA8	*HSPA8*-1	Heat-shock-protein-70Ba	NM_169469.1	NP_731716.1	3R+(87B)	1	641	70.177	N	F
		HSPA8-2	Heat-shock-protein-70Bb	NM_080188.2	NP_524927.2	3R+(87B)	1	641	70.191	N	F
		HSPA8-3	Heat-shock-protein-70Bc	NM_141952.1	NP_650209.1	3R+(87B)	1	641	70.191	N	F
DNAJ	**DNAJA**										
(HSP40)	*DNAJA1*	*DNAJA1*-1	DnaJ-like-1 isoform A	NM_079212.2	NP_523936.2	3L-(64E)	2	334	37.026	N	F
		DNAJA1-1	DnaJ-like-1 isoform B	NM_168128.1	NP_729086.1	3L-(64E)	1	334	37.026	N	F
	DNAJA2	*DNAJA2*-1	cysteine string protein isoform A	NM_079489.2	NP_524213.1	3L+(79D)	5	223	23.847	N	F
		DNAJA2-2	cysteine string protein isoform B	NM_168949.1	NP_730713.1	3L+(79D)	7	249	26.895	N	F
		DNAJA2-3	cysteine string protein isoform C	NM_168950.2	NP_730714.2	3L+(79D)	6	228	24.556	N	F
	DNAJA3	*DNAJA3*-1	DnaJ homolog, isoform A	NM_135761.2	NP_609605.1	2L-(34A)	3	389	43.935	N	F
		DNAJA3-2	DnaJ homolog, isoform B	NM_165025.1	NP_723785.1	2L-(34A)	4	389	43.935	N	F
	DNAJA4	*DNAJA4*-1	DnaJ-like-2, isoform A	NM_142026.2	NP_650283.1	3R+(87E)	3	403	45.228	N	F
		DNAJA4-2	DnaJ-like-2, isoform B	NM_169519.1	NP_731804.1	3R+(87E)	3	403	45.228	N	F
		DNAJA4-3	DnaJ-like-2, isoform C	NM_169520.1	NP_731805.1	3R+(87E)	3	403	45.228	N	F
		DNAJA4-4	DnaJ-like-2, isoform D	NM_169521.1	NP_731806.1	3R+(87E)	2	403	45.228	N	F
		DNAJA4-5	DnaJ-like-2, isoform E	NM_169522.1	NP_731807.1	3R+(87E)	4	403	45.228	N	F
	DNAJA5	*DNAJA5*-1	lethal (2) tumorous imaginal discs isoform A	NM_080193.3	NP_524932.2	2R+(59F)	2	447	49.398	Y(21)	F
		DNAJA5-2	lethal (2) tumorous imaginal discs isoform B	NM_206210.1	NP_995932.1	2R+(59F)	2	520	56.134	Y(21)	F
		DNAJA5-3	lethal (2) tumorous imaginal discs isoform C	NM_206209.1	NP_995931.1	2R+(59F)	2	507	56.198	Y(21)	F
		DNAJA5-4	Tid56 protein	X95242.1	CAA64531.1	2R+(59F)	5	518	55.961	Y(21)	F
		DNAJA5-5	Tid56 protein	X95247.1	CAA64536.1	2R+(59F)	5	520	56.134	Y(21)	F
		DNAJA5-6	SD10289p	AY069853.1	AAL39998.1	2R+(59F)	5	447	49.429	Y(21)	F
	DNAJA6	*DNAJA6*-1	DnaJ-like-60, isoform A	NM_079116.3	NP_523840.1	2R-(60C)	4	217	24.778	N	F
		DNAJA6-2	DnaJ-like-60, isoform C	NM_138048.2	NP_611892.2	2R-(60C)	4	217	24.778	N	F
		DNAJA6-3	dnaJ-like protein	Y11900.1	CAA72641.1	2R-(60C)	4	217	24.778	N	F
	DNAJA7	*DNAJA7*-1	CG30156	NM_165500.1	NP_724520.1	2R-(42E)	3	358	41.205	N	F
	DNAJA8	*DNAJA8*-1	CG5001	NM_134742.2	NP_608586.2	2L+(21F)	6	350	39.524	N	F
	DNAJB										
	DNAJB1	*DNAJB1*-1	mrj isoform A	NM_166152.1	NP_725541.1	2R-(52F)	6	259	29.026	N	F
		DNAJB1-2	mrj isoform B	NM_166153.1	NP_725542.1	2R-(52F)	6	259	29.026	N	F
		DNAJB1-3	mrj isoform C	NM_166154.1	NP_725543.1	2R-(52F)	6	259	29.026	N	F
		DNAJB1-4	mrj isoform D	NM_166155.1	NP_725544.1	2R-(52F)	6	259	29.026	N	F
		DNAJB1-5	mrj isoform E	NM_166156.2	NP_725545.2	2R-(52F)	7	346	38.735	N	F
	DNAJC										
	DNAJC1	*DNAJC1*-1	jdp isoform A	NM_143550.2	NP_651807.1	3R-(99F)	5	195	21.649	N	F
		DNAJC1-2	dp isoform B	NM_170501.1	NP_733380.1	3R-(99F)	5	182	20.061	N	F
		DNAJC1-3	jdp isoform C	NM_170502.1	NP_733381.1	3R-(99F)	4	170	19.42	N	F

续表 3

基因家族	基因	基因变体		登录号		染色体定位[a]	外显子数	氨基酸长度	分子量/kDa	信号肽[b]	基因功能[c]
		基因剪切(新命名)	旧名称	核苷酸	氨基酸						
HSPB	**HSPBA**										
(sHSP)	*HSPBA*1	*HSPBA*1 - 1	HSP22 isoform A	NM_001031944.1	NP_001027115.1	3L + (67B)	3	174	19.762	N	F
		*HSPBA*1 - 2	HSP22 isoform B	NM_001031943.1	NP_001027114.1	3L + (67B)	1	174	19.762	N	F
	*HSPBA*2	*HSPBA*2 - 1	HSP26	NM_079273.2	NP_523997.1	3L + (67B)	1	208	22.993	N	F
		*HSPBA*2 - 2	Hsp23	NM_079275.2	NP_523999.1	3L + (67B)	1	186	20.628	N	F
		*HSPBA*2 - 3	HSP27	NM_079276.2	NP_524000.1	3L + (67B)	1	213	23.615	N	F
	*HSPBA*3	*HSPBA*3 - 1	heat shock gene 67 Bc	NM_079270.1	NP_523994.1	3L + (67B)	1	199	22.178	N	F
		*HSPBA*3 - 2	heat shock gene 67 Ba	NM_079274.3	NP_523998.1	3L + (67B)	1	445	46.928	N	F
		*HSPBA*3 - 3	heat shock gene 67 Bb, isoform A	NM_001031946.1	NP_001027117.1	3L + (67B)	3	111	12.613	N	P
		*HSPBA*3 - 4	heat shock gene 67 Bb, isoform B	NM_001031945.1	NP_001027116.1	3L + (67B)	3	111	12.613	N	P
	*HSPBA*4	*HSPBA*5 - 1	CG4461	NM_140047.1	NP_648304.1	3L + (67B)	1	200	23.776	N	F
	HSPBB										
	*HSPBB*1	*HSPBB*1 - 1	lethal (2) essential for life isoform A	NM_079103.2	NP_523827.1	2R - (59F)	3	187	21.307	N	F
		*HSPBB*1 - 2	lethal (2) essential for life isoform B	NM_166620.2	NP_726354.1	2R - (59F)	2	145	17.233	N	F
	*HSPBB*2	*HSPBB*2 - 1	CG14207 isoform B	NM_167669.1	NP_728275.1	X + (18D)	5	192	21.776	N	F
		*HSPBB*2 - 2	CG14207 isoform A	NM_134482.2	NP_608326.1	X + (18D)	4	183	20.841	N	F
	*HSPBB*3	*HSPBB*3	CG7409	NM_139898.1	NP_648155.1	3L + (66A)	1	154	17.965	N	F
HSPD	**HSPD**										
(HSP60)	*HSPD*1	*HSPD*1 - 1	heat shock protein 60 isoform A	NM_078560.2	NP_511115.2	X + (10A)	3	573	60.805	N	F
		*HSPD*1 - 2	heat shock protein 60 isoform B	NM_167266.1	NP_727489.1	X + (10A)	3	573	60.805	N	F
	*HSPD*2	*HSPD*2 - 1	heat shock protein 60 isoform C	NM_135104.4	NP_608948.2	2L + (25F)	3	576	61.583	N	F
	*HSPD*3	*HSPD*3 - 1	heat shock protein 60 related	NM_080186.3	NP_524925.1	2L + (21E)	1	648	68.633	N	F
	*HSPD*4	*HSPD*4 - 1	Hsp60D, isoform A	NM_135812.2	NP_609656.1	2L - (34C)	2	558	60.28	N	F
		*HSPD*4 - 2	Hsp60D, isoform B	NM_165043.1	NP_723818.1	2L - (34C)	1	558	60.28	N	F
	*HSPD*5	*HSPD*5 - 1	CG7033 isoform A	NM_132296.3	NP_572524.1	X + (8C)	4	535	58.06	N	F
		*HSPD*5 - 2	CG7033 isoform C	NM_176715.1	NP_788888.1	X + (8C)	4	533	57.8	N	F
CCT	**CCT**										
(TRiC)	*CCT*1	*CCT*1 - 1	T - complex chaperonin 5, isoform A	NM_078983.2	NP_523707.1	2R + (48E)	3	542	59.275	N	F
		*CCT*1 - 2	T - complex chaperonin 5, isoform B	NM_165867.1	NP_725107.1	2R + (48E)	3	512	55.788	N	F
	*CCT*2	*CCT*2 - 1	cctgamma, isoform A	NM_142315.2	NP_650572.2	3R + (89D)	5	544	59391	N	F
		*CCT*2 - 2	cctgamma, isoform B	NM_169727.1	NP_732167.1	3R + (89D)	4	544	59391	N	F
	*CCT*3	*CCT*3 - 1	Tcp1 - like, isoform A	NM_079726.3	NP_524450.2	3R - (94B)	4	557	59.553	Y(18)	F
		*CCT*3 - 2	Tcp1 - like, isoform B	NM_170016.1	NP_732748.1	3R - (94B)	5	557	59.553	Y(18)	F

注:a 处"+"为正义链;"-"为反义链;b 处 Y 表示有信号肽,括号内数字表示信号肽长度(氨基酸数目);c 处 F 表示基因有功能,P 表示假基因。

2.3 黑腹果蝇中热激蛋白家族系统发育分析

将黑腹果蝇中 35 个家族基因氨基酸序列整合为 FASTA 格式,利用 Clustal X 软件对氨基酸序列进行多序列比对并进行人工校对,用 Phylip 软件构建最大简约树(MP 树),采用 1 000 次可置信度分析。结果如图 3 显示,图中数字代表 Bootstrap 值。采用最大简约树法建立的黑腹果蝇中热激蛋白系统发育树将内群分为 4 个聚类群且具有较高的置信度。从发育树中可看出,DNAJ 家族和 HSPB 家族、HSPD 家族和 CCT 家族聚集在 1 个同一聚类分支上,这说明他们序列相似性较高并具有较近的亲缘关系。而整体进化趋势是一种趋异性的进化,不同的家族由同一个祖先进化而来。其中 HSPC 家族分野最早,HSPD 家族、CCT 家族分野最晚。可以推断,黑腹果蝇热激蛋白家族是由一个共同的祖先经过倍增形成同源基因,然后演变为不同的基因家族,其中一个演变为 HSPC 家族,另一个演变为其他家族的共同祖先。

3 讨论

热激蛋白首先在果蝇中发现,热激蛋白的研究对于预防病虫害以及研究人类疾病具有重要意义。之前对于人类及其他物种热激蛋白的研究较为集中与系统,而对于黑腹果蝇中热激蛋白的研究主要集中在小分子HSP,HSP70等家族中单个基因上,对于黑腹果蝇基因组中全部热激蛋白尚未有系统的研究,对基因的分类、命名等相当混乱。本文以黑腹果蝇热激蛋白家族代表性基因搜索基因组公共数据库,对黑腹果蝇全基因组中热激蛋白进行系统的分析,研究共获得35个基因,其中新发现7个,经过基因功能分析表明这7个基因是具有功能的。并将这些获得的基因按照HUGO基因命名委员会标准和人类基因组中热激蛋白命名规则进行重新命名和分类,发现在黑腹果蝇基因组中不存在HSPE(HSP10)家族和HSPH(HSP110)家族,这两个家族存在于人类及小鼠基因组中,这与王秋香等对黑腹果蝇热激蛋白的分类中存在HSPH结果不一致,这可能与早期黑腹果蝇命名的混乱或采用数据的命名不一致有关。

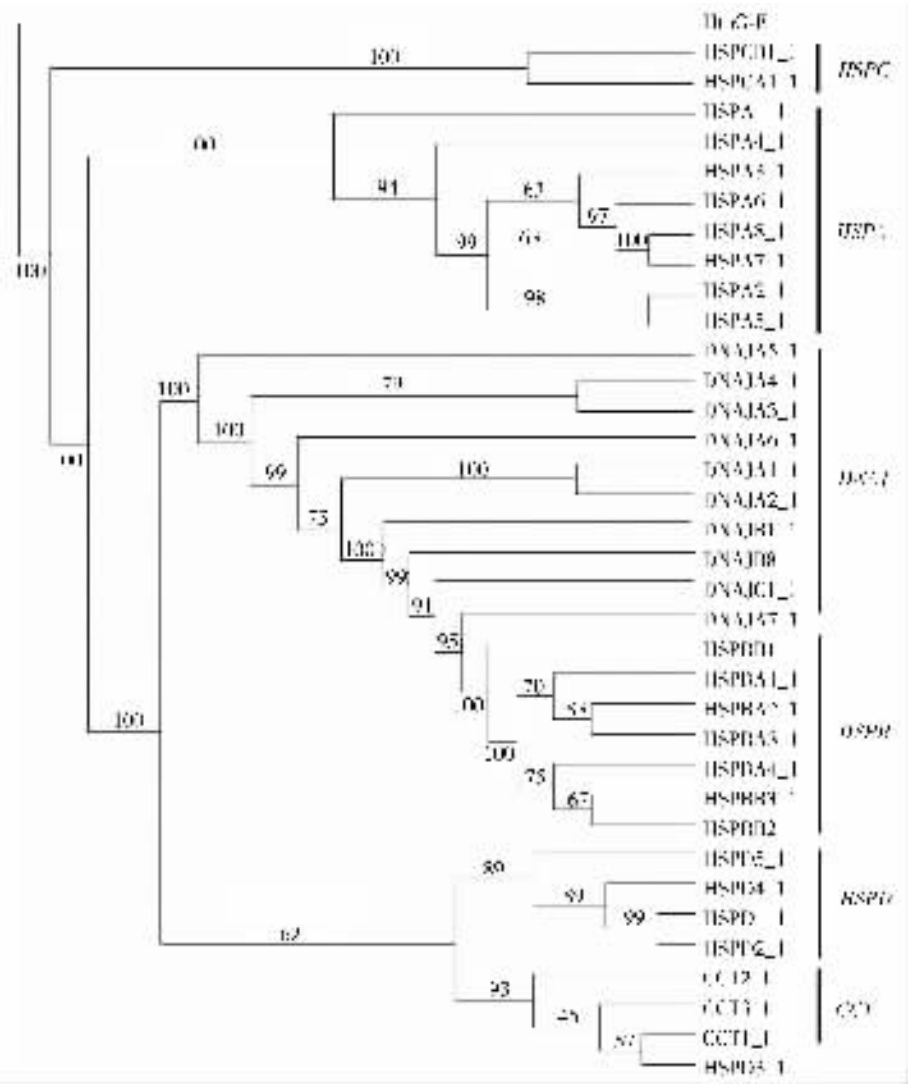

图3 黑腹果蝇热激蛋白序列系统发育树

Fig. 3 The molecular phylogenetic tree of the *D. melanogaster* HSP families of amino acids

从染色体分布来看,主要分布在染色体2L、3L、2R、3R和X上。HSPB家族中大部分成员均只有1个外显子,没有内含子存在,在DNAJ、HSPD、CCT、HSPA和HSPC家族中,该家族基因均含有多个外显子。通过对热激蛋白的保守结构域、功能基序分析,发现黑腹果蝇热激蛋白家族高度保守。通过分析,将黑腹果蝇热激蛋白分为6个基因家族,分别为HSPB家族、HSPD家族、HSPA家族、HSPC家族、DNAJ家族及CCT家族。分子进化表明,黑腹果蝇中热激蛋白家族可能来源于同一个祖先,由一个共同的祖先经过倍增形成6个基因家族,并在不同时期分野,进化方式属于趋异性进化。这些工作,对于进一步研究黑腹果蝇热激蛋白家族提供了相应的生物信息学数据。

参考文献:

[1] 李冰祥,蔡惠罗,陈永林. 昆虫的热休克反应和热休克蛋白[J]. 昆虫学报,1997,40(4):417-423.

[2] Lindguist S,Craig E A. The heat shock proteins [J]. Annual Review of Genetics,1988,22:631-677.

[3] Bilen J,Bonini N M. *Drosophila* as a model for human neurodegenerative disease[J]. Annu Rev Genet,2005,39:153-171.

[4] Sang T K,Jackson G R. *Drosophila* a models of neurodegenerative disease [J]. Neurotherapeutics,2005, 2(3):438-446.

[5] 万永奇,谢维. 生命科学与人类疾病研究的重要模

型——果蝇[J]. 生命科学,2006,18(5):425-429.
[6] 王秋香,李宗芸,谢艳,等. 果蝇热激蛋白的研究进展[J]. 昆虫知识,2008,45(5):698-702.
[7] 王海鸿,雷仲仁. 昆虫热休克蛋白的研究进展[J]. 中国农业科学,2005,38(10):2023-2034.
[8] Corces V,Holmgren R,Freund R. Four heat shock proteins of *Drosophila melanogaster* coded within a 12-kilobase region in chromosome subdivision 67B[J]. Genetics,1980,77(9):5390-5393.
[9] Jakobt U,GaestelP M,Engel K. Small heat shock proteins are molecular chaperones[J]. The Journal of Biological Chemistry,1993,268(3):1517-1520.
[10] Dahlgaard J,Loeschcke V,Michalak P,et al. Induced thermotolerance and associated expression of the heat-shock protein Hsp70 in adult *Drosophila melanogaster*[J]. Function Ecol,1998,12:786-793.
[11] Robert A K,Martin E F. Hsp70 and larval thermo- tolerance in *Drosophila melanogaster*:how much is enough and when is more too much[J]. Journal of Insect Physiology,1998,44(11):1091-1101.
[12] Lakhotia S C,Prasanth K V. Tissue-and development-specific induction and turnover of hsp70 transcripts from loci 87A and 87C after heat shock and during recovery in *Drosophila melanogaster*[J]. The Journal of Experimental Biology,2002,205:345-358.
[13] Chen B,William H P,Gui L. The HSP90 family of genes in the human genome:Insights into their divergence and evolution[J]. Genomics,2005,86:627- 637.
[14] Chen B,Zhong D B,Monteiro A. Comparative genomics and evolution of the HSP90 family of genes across all kingdoms of organisms[J]. BMC Genomics,2006,7:156.
[15] Csermely P,Schnaider T,Soti C,et al. The 90-kDa molecular chaperone family:structure,function,and clinical applications. A comprehensive review[J]. Pharmacol Ther,1998,79:129-168.
[16] Kampinga H H,Hageman J,Vos M. Guidelines for the nomenclature of the human heat shock proteins[J]. Cell Stress and Chaperones,2006,14:105-111.

Animal Sciences

Bioinformatic Analysis of Gene Families of Heat Shock Proteins in *Drosophila melanogaster*

MOU Guo-peng, CHEN Bin, HE Zheng-bo

(Institute of Entomology and Molecular Biology, Engineering Research Center for Bioactive Substance Biotechnology of Ministry of Education, Chongqing Key Laboratory of Animal Biology, College of Life Sciences, Chongqing Normal University, Chongqing 400047, China)

Abstract: Heat shock proteins (HSPs) are a group of highly conserved proteins synthesized in cultured cells or whole organisms, which can generally be induced by stress stimulations and exist throughout all kinds of organisms. HSPs have key roles in cell growth, development, differentiation and gene transcription. In this study, it was performed bioinformatic analysis of gene families of heat shock proteins in *Drosophila melanogaster*. The results showed that there are 6 HSPs families: HSPC (HSP90), HSPA (HSP70), DNAJ (HSP40), HSPB (small HSP) as well as the chaperonin families HSPD (HSP60) and CCT (TRiC) in *Drosophila melanogaster*; however, there is no HSPE (HSP10) and HSPH (SHP110) families found, which exist in human and mouse. HSPC, HSPA, DNAJ, HSPB, HSPD and CCT contain 2,8,10,7, 5 and 8 genes, respectively. And 7 genes are newly found in this study, which has been classified and named as *HSPBA*5, *HSPBB*2, *HSPBB*3, *DNAJA*7, *DNAJA*8, *HSPD*5 and *HSPCB*1. All these gene mapped onto chromosomes 2L, 3L, 2R, 3R and X. The gene structures, and the functional motifs and polymorphic variants of their proteins were documented and discussed. Phylogenetic analyses based on protein data demonstrated that they had the common ancestor and clustered as 6 families.

Key words: *Drosophila melanogaster*; HSPs; gene family; bioinformatics; phylogenetics

(责任编辑 方 兴)