The Molecular Evolution of the Small Heat-Shock Proteins in Plants

Elizabeth R. Waters

Department of Ecology and Evolutionary Biology, University of Arizona, Tucson, Arizona 85721

Manuscript received April 11, 1995
Accepted for publication June 28, 1995

ABSTRACT

The small heat-shock proteins have undergone a tremendous diversification in plants; whereas only a single small heat-shock protein is found in fungi and many animals, over 20 different small heat-shock proteins are found in higher plants. The small heat-shock proteins in plants have diversified in both sequence and cellular localization and are encoded by at least five gene families. In this study, 44 small heat-shock protein DNA and amino acid sequences were examined, using both phylogenetic analysis and analysis of nucleotide substitution patterns to elucidate the evolutionary history of the small heat-shock proteins. The phylogenetic relationships of the small heat-shock proteins, estimated using parsimony and distance methods, reveal that gene duplication, sequence divergence and gene conversion have all played a role in the evolution of the small heat-shock proteins. Analysis of nonsynonymous substitutions and conservative and radical replacement substitutions (in relation to hydrophobicity) indicates that the small heat-shock protein gene families are evolving at different rates. This suggests that the small heat-shock proteins may have diversified in function as well as in sequence and cellular localization.

THE small heat-shock proteins are those proteins produced in response to high temperature stress that are smaller than 30 kDa in size. Higher plants have at least 20 and some plant species may have as many as 40 different small heat-shock proteins (VIERLING 1991). In contrast, most other organisms have one or only a few small heat-shock proteins. *Saccharomyces cerevisiae* has one small heat-shock protein and Drosophila has four (ARRIGO and LANDRY 1994). The diversification of the plant small heat-shock proteins occurred after the split of the plant and animal lineages. This suggests that the tremendous diversification of small heat-shock proteins in plants may reflect adaptations to stresses unique to plants. The small heat-shock protein genes in plants comprise a large multigene family composed of at least five distinct gene families; all are nuclear encoded. The plant small heat-shock proteins have previously been divided into four classes based on sequence similarity and cellular localization (VIERLING 1991). One class of proteins localizes to the chloroplast (CP), one to the endoplasmic reticulum (ER), and two to the cytosol, classes I and II. Recently a fifth class of mitochondrial (MT)-localized proteins has been reported (LENNE and DOUCE 1994). The diversification of cellular localization of small heat-shock proteins is unique to plants; all of the nonplant small heat-shock proteins localize to the cytosol (ARRIGO and LANDRY 1994).

The plant small heat-shock proteins are related to the small heat-shock proteins in other organisms and to the vertebrate alpha-crystallin proteins (PLESOFSKY-VIG *et al.* 1992; JONG *et al.* 1993). All share a conserved heat-shock region in the carboxyl terminal domain. Comparisons of the amino acid sequences of the carboxyl terminal domain of some plant small heat-shock proteins and other small heat-shock proteins confirms that the plant proteins are related to but quite distinct from other small heat-shock proteins (PLESOFSKY-VIG *et al.* 1992; JONG *et al.* 1993). PLESOFSKY-VIG *et al.* (1992) concluded, based on branch lengths and tree topology, that the plant small heat-shock proteins have evolved more slowly than the animal small heat-shock proteins. They also concluded that the CP-localized protein originated from the chloroplast endosymbiotic event and is thus only distantly related to the other small heat-shock proteins (PLESOFSKY-VIG *et al.* 1992).

The *in vivo* function of the small heat-shock proteins is not known. Recent *in vitro* studies suggest that the small heat-shock proteins, like the large HSPs, may be molecular chaperones (JAKOB *et al.* 1993; MERCK *et al.* 1993; JAKOB and BUCHNER 1994; LEE *et al.* 1995). The biochemistry of the large heat-shock proteins (HSPs 70, 90 and 60) has been well studied (BECKMANN *et al.* 1990; GETHING and SAMBROOK 1992; BECKER and CRAIG 1994; CRAIG *et al.* 1994; SCHNEIDER *et al.* 1994). The evolution of HSP 70s has also been studied in some detail (BOORSTEIN *et al.* 1994; RENSING and MAIER 1994). These studies reveal that, in contrast to the small heat-shock proteins, the genes coding for the HSP 70 proteins duplicated very early in the evolution of eukaryotes. The selective constraints on the large HSPs and the small HSPs are very different. Amino acid sequences of HSP 70 are highly conserved; there is almost 50% amino acid identity from *Zea mays* to *Escherichia coli*

Author e-mail: ewaters@ccit.arizona.edu

TABLE 1

Gene and protein accession numbers

Species	Protein	DNA accession number	Protein accession number
Chloroplast-localized proteins			
Arabidopsis thaliana	HSP 21	X54102	P31170
Glycine max	HSP 22	X07188	P09887
Petunia hybrida	HSP 21	X54103	P30222
Pisum sativium	HSP 21	X07187	P09886
Triticum aestivum	HSP 26A	X58280	Q00445
Triticum aestivum	HSP 26B	X67328	S26581
Zea mays	HSP 26	L28712	
Mitochondrial-localized protein			
Chenopodium rubrum	HSP 23	X15333	
Endoplasmic reticulum-localized proteins			
Arabidopsis thaliana	HSP 22	U11501	
Glycine max	HSP 22	X63198	P30236
Pisum sativum	HSP 22	M33898	
Class I cytocolically localized proteins			
Arabidopsis thaliana	HSP 17.6	X16076	P13853
Arabidopsis thaliana	HSP 17.4	X17293	P19036
Arbidopsis thaliana	HSP 18.2	X17295	P10307
Chenopodium rubrum	HSP 18.3	X53870	S20803
Daucus carota	HSP 18.0	X53852	P27397
Daucus carota	HSP 17.8	X53851	P27396
Glycine max	HSP 17.5	M11318	P04793
Glycine max	HSP 17.6	M11317	P04795
Glycine max	HSP 18.5	X07160	P05478
Helianthus annuus	HSP 17.6	X59701	P30693
Lycopericoscon esculentum	HSP 17.8	X56138	P30221
Medicago sativa	HSP 18.1	X58710	P27879
Medicago sativa	HSP 18.2	X58711	P27880
Oryza sativa	HSP 16.9	X60820	P27777
Oryza sativa	HSP 17.4	D12635	P31673
Pisum sativum	HSP 18.1	M33899	P19243
Triticum aestivum	HSP 16.9A	X13431	P12810
Triticum aestivum	HSP 16.9B	X64618	S21600
Triticum aestivum	HSP 16.9C	L14444	
Zea mays	HSP 17.2	X65725	
Class II cytocolically localized proteins			
Arabidopsis thaliana	HSP 17.6	X63443	P29830
Glycine max	HSP 17.9	X07159	P05477
Ipomea nil (Pharbatis nil)	HSP 18.8	M99430	QO1545
Ipomea nil (Pharbatis nil)	HSP 17.2	M99429	QO1544
Lilium longiflorum	HSP 18.2	Bouchard (1990)	
Lilium longiflorum	HSP 17.6	D21816	
Lilium longiflorum	HSP 16.5	D21818	
Pisum sativum	HSP 17.7	M33901	S12720
Triticum aestivum	HSP 17.3	X58279	S16525
Zea mays	HSP 17.5	X54076	P24631
Zea mays	HSP 17.8	X54075	P24632

(Lindquist and Craig 1988). The small heat-shock proteins evolve much more quickly; there is <40% amino acid identity between the small heat-shock protein in *S. cerevisiae* and the plant small heat-shock proteins. The different evolutionary histories of the large and small HSPs suggest that, even if both types of HSPs are molecular chaperones, the specific functions within the cell and the selective constraints on these groups of proteins are very different.

Patterns of DNA sequence divergence can be very useful indicators of differences in selective constraint and possible functional divergence (Hughes *et al.* 1990; Hughes 1993a,b; Karlin *et al.* 1992). In a study of the HSP 70 genes, Hughes demonstrated that rates of nucleotide substitutions reflect the known functional differences among the HSP 70s (Hughes 1993b). In this study of small heat-shock proteins, I examined both the complete DNA and amino acid sequences of 44 plant

```
                               10         20        30        40       50         60         70        80         90      100

T.aestivum 26a        MA........AANAPFALVSRLSPAARLPIRAWRAARPAPLST..GGRTRPLSVASAAQ ENRDNSVDVQ.V  SQAQNAGN.QQGNAVQRRPRRA.GFDISP
T.aestivum 26b        MA........AANAPFAL.SRLSPAARLPFRAWRAARPAPVWT...GRTRPLSVASAAQ ENRDNSVDVQ.V  SQAQNAGN.QQGNAVQRRPRRA.GFDISP
Z.mays 26             MA........AAPFAIAGRLSPVARLPVRA.....WRPAHGFASS.GRARSLAVASAAQ ENRDNSVDVQ.V  SQ..NGGNRQQGNAVQRRPRRATALDISP
P.sativum 21          MAQSVSLSTIASPILSQ...KPGSSVKSTPPCMASFPLRRQLPRLGLRNV......RAQ AGGDGDNKDNSV  EVHRVNKDD.QGTAVERKPRRS.SIDISP
G.max 21              ..........................................................G GDNKDNSVEVQH  VSKGD.....QGTAVEKKPRRT.AMDISP
A.thaliana 21         MA..STLSFAASALCSP..LAPSPSVSSKSA..TPFSVS....FPRKIPS....RIRAQ DQRENSIDVV..  ..QQGQQKGNQGSSVEKRPQQRLTMDVSP
P.hybrida 21          MA.CKTLTCSASPLVSNGVVSATSRTNNKKTTTAPFSVCFPYSKCSVRKPASRLVAQAT GDNKDTSVDVHV  SNNNQGGNNQGSAVE.RRPRRM.ALDVSP
C.rubrum 23           MA.SMALRRLASRNLVSGGIFR...........PLSVSRSFNTN...........AQMG RVDHDHELDDRS  NRAPISRRG........DFPASFFSDVFD
L.longiflorum 18.2    ........................................................... ............  .................MGSKLTREEYNT
L.longiflorum 17.6    ........................................................... ............  .................MGSKLTREEYDT
L.longiflorum 16.5    ........................................................... ............  .................MDSKFEVDHSLI
Z.mays 17.8           ........................................................... ............  .................MDAVMFGLET..
Z.mays 17.5           ........................................................... ............  .................MDGRMFGLET..
T.aestivum 17.3       ........................................................... ............  .................MAGMVFGLDA..
P.sativum 17.7        ........................................................... ............  .................MDFRLMDLDS..
G.max 17.9            ........................................................... ............  .................MDFRVMGLES..
I.nil 17.2            ........................................................... ............  .................MDLRLMGFDH..
I.nil 18.8            ........................................................... ............  .................MDLRNFGLSNFG
A.thaliana 17.6II     ........................................................... ............  .................MDLGRF......
D.carrota 18.0        ........................................................... ............  .................MSIIPS..FFGS
D.carrota 17.8        ........................................................... ............  .................MSIIPS..FFG.
M.sativus 18.1        ........................................................... ............  .............................
M.sativus 18.2        ........................................................... ............  .................MSLIPS..FFG.
P.sativum 18.1        ........................................................... ............  .................MSLIPS..FFS.
G.max 17.5            ........................................................... ............  .................MSLIPS..IFG.
G.max 17.3            ........................................................... ............  .................MSLIPS..FFG.
G.max 18.5            ........................................................... ............  .................MSLIPN..FFG.
G.max 17.6            ........................................................... ............  .................MSLIPS..IFG.
L.esculentum 17.8     ........................................................... ............  .................MSLIPR..IFG.
A.thaliana 17.6       ........................................................... ............  .................MSLIPS..IFG.
A.thaliana 17.4       ........................................................... ............  .................MSLVPS..FFG.
A.thaliana 18.2       ........................................................... ............  .................MSLIPS..IFG.
H.annuus 17.6         ........................................................... ............  .................MSIIPS..FFT.
P.sativum 17.9        ........................................................... ............  .................IIPRV.FGT.
T.aestivum 16.9b      ........................................................... ............  .................MSIV........
T.aestivum 16.9c      ........................................................... ............  .............................
T.aestivum 16.9a      ........................................................... ............  .................MSIV........
O.sativa 16.9         ........................................................... ............  .................MSLV........
Z.mays 17.2           ........................................................... ............  .................MSLV........
O.sativa 17.4         ........................................................... ............  .................MSMI........
C.rubrum 18.3         ........................................................... ............  .................MSLIPNNWFNT.
P.sativum 22          ........................................................... ............  MSLKPLNMLLVPFLLLILAADFPLKAKGS
G.max 22              ........................................................... ............  MRLQQLNLF...FLLLCVA.....KANGS
A.thaliana 22         ........................................................... ............  MM....KHLLSIFFIGALLLGNIKTSEGS
```

FIGURE 1.—Amino acid alignment. Boxes mark conserved regions. #, highly conserved residue; *, completely conserved residue.

small heat-shock proteins. Using both distance- and parsimony-based phylogenetic methods, I constructed gene trees to determine the evolutionary relationships among and within the plant small heat-shock protein gene families. In addition I examined the rates of nucleotide substitutions among the plant small heat-shock proteins. I have found evidence of differences in selective constraint among the small heat-shock proteins suggesting that functional differences may also exist among the plant small heat-shock proteins.

MATERIALS AND METHODS

Sequence alignment: DNA and amino acid sequences of 44 small heat-shock proteins were obtained from the databases or the literature. Accession numbers or references are listed in Table 1. When amino acid sequences were not available, DNA sequences were translated using Translate in GCG (Genetics Computer Group 1991). The size of the HSPs (in kDa) were either taken from the literature or determined using the program PeptideSort in GCG. Amino acid sequences were aligned using PileUp in GCG. The alignment was further refined by hand in LineUp in GCG (Figure 1). The aligned protein sequences were imported into the program DNA Stacks (EERNISSE 1992). The unaligned coding regions of the DNA sequences were also imported. The DNA sequences were aligned by imposing the gaps in the amino acid alignment upon the DNA sequences (DNA alignment is available upon request from the author). Pairwise comparisons of overall sequence similarity were done using the program Gap in GCG.

Phylogenetic analysis: Phylogenetic analysis of the aligned DNA and amino acid sequences were conducted using parsimony in PAUP (SWOFFORD 1993) version 3.1.1 and distance (DNAdist, Protdist and NeighborJoining) in PHYLIP (FELSENSTEIN 1993) version 3.5c. PHYLIP is available by anonymous FTP at "evolution.genetics.washington.edu."

The parsimony analyses were conducted as follows: heuristic searches with 100 random addition replicates, with MULPARS and TBR branch swapping (steepest descent was not invoked), were conducted to find the most parsimonious trees. All trees were found in the first or second replicate, no additional trees were found in the next 98 replicates. The strict consensus of the most parsimonious trees was constructed. Support for branches was evaluated by bootstrap analysis: 100 Bootstrap replicates with the same conditions as above were conducted.

The tree presented in this paper is arbitrarily rooted with the sequences for the chloroplast proteins. At the present time it is also not possible to unequivocally choose a root for the small heat-shock proteins. Additions of other eukaryotic small heat-shock proteins (from yeast and humans) to the data matrix make alignment more difficult and, in addition, do not resolve the relationships among the plant small heat-shock protein gene families.

The analysis of the DNA sequences were first conducted with the complete sequences and then with the transit peptides and the third positions removed. Transit sequences were

```
                           110           120        130        140        150        160        170        180        190        200

T.aestivum 26a       ..FGLV DPMSPMRTMRQMLDTM DRLF DDAVGFP..TRRSPAA RAR..RRMPWDI MEDEKEVKMRF DMPGLSREEVRVMVEDDALVIRGEHKKE. .AGEGQ
T.aestivum 26b       ..FGLV DPMSPMRTMRQMLDTM DRLF DDAVGFP..TARSPAR RAKTP.RMPWDI MEDEKEVKMRF DMPGLSREEVRVMVEDDALVIRGEHKKE. .AGEGQ
Z.mays 26            SPFGLV DPMSPMRTMRQMLDTM DRLF DDAVGFPMGTRRSPAT TGDV..RLPWDI VEDEKEVKMRI DMPGLARDEVKVMVEDDTLVIRGEHKKEE GAEGGS
P.sativum 21         ..FGLL DPWSPMRSMRQMLDTM DRIF EDAITIPG.RNIGGGE I.....RVPWEI KDEEHEIRMRF DMPGVSKEDVKVSVEDDVLVIKSDHR... ..EENG
G.max 21             ..FGIL DPWSPMRSMRQILDTM DRVF EDTMTFPG.RNIGGGE I.....RAPWDI KDEEHEIRMRF DMPGLAKEDVKVSVEDDMLVIKGGHKSE. ..QEHG
A.thaliana 21        ..DPWS DPLSPMRTMRQMLDTM DRMF EDTMPVSG.RNRGGSG V..SEIRAPWDI KEEEHEIKMRF DMPGLSKEDVKISVEDNVLVIKGEQKKE. .....
P.hybrida 21         ..FGLL DPMSPMRTMRQMMDTM DRLF EDTMTFPGSRNRGTGE I.....RAPWDI KDDENEIKMRF DMPGLSKEEVKVSVEDDVLVIKGEHKKE. .....
C.rubrum 23          P. FRAT R..SVGQLMNLMDQLM ENPF ..........MAASR GSGRAMRRGWDV REDEEALELKV DMPGLAKEDVKVSVEDNTLIIKSEAEKE. .....
L.longiflorum 18.2   LL AAFH KLTVRLEVASVPKD.. .... ................ ......ATPADI KNLPDAYLYFI AMPRVRTGEIKVEVEDDSDLVVISGERKR ...EEE
L.longiflorum 17.6   LL AAFH KLTVRLEVASVPKD.. .... ................ ......ATPADI KNLPDAYLYFI DMPGVRTGEIKVEVEDDSALVIIGERKRE ...EEE
L.longiflorum 16.5   AK LNQL TEFL............ .... ...ANRNQPLRAPFVP DARAMPAAATDI KDMPGAYVFII DMPGVESEEIKIDVEEGNMLVISGERKRE ..EEEE
Z.mays 17.8          .....P LMAALQHLLDVPDGDA GAGG DNKTGSGGSATRTYVR DARAMAATPADV KELPGAYAFVV DMPGLGTGDIRVQVEDERVLVVSGERRRE ...ERE
Z.mays 17.5          .....P LMVALQHLLDVPDGDA GAGG DKA..GGGGPTRTYVA DARAMAVTPADV KELPGAYAFVV DMPGLGTGDIKVQVEDERVLVISGERRRE ...ERE
T.aestivum 17.3      .....P MMAALQHLLDIPDGEA EPPP EK.....QGPTRAYVR DARAMAATPADV KELPGAYAFVV DMPGLGSGDIKVQVEDERVLVISGERRRE ...EKE
P.sativum 17.7       .....P LFNTLHHIMDLTDD.T TEKN ......LNAPTRTYVR DAKAMAATPADV KEHPNSYVFMV DMPGVKSGDIKVQVEDENVLLISGER.KR ...EEE
G.max 17.9           .....P LFHTLQHMMDMSED.G AGDN K....THNAPTWSYVR DAKAMAATPADV KEYPNSYVFEI DMPGLKSGDIKVQVEDDNLLLICGER.KR ...DEE
I.nil 17.2           .....P LF...HHIMDYAGD.D KSSN S......SAPSRTFML DAKAMAATPADV KEYPNSYVFII DMPGLKSGDIKVQVDGDNVLSISGER.KR ..EAEE
I.nil 18.8           LE P..Q LLSTIQDMLDFADDHD RAGR A....PPEQPIRAYVR DAKAMAATPADV KEYPNSYVFIA DMPGVKAAEIKVQVEDDNVLVVSGERTER ..EKDE
A.thaliana 17.6II    .....P IISILEDMLEVPEDHN NEK. .....TRNNPSRVYMR DAKAMAATPADV IEHPNAYAFVV DMPGIKGDEIKVQVENDNVLVVSGERQRE ..NKEN
D.carrota 18.0       SR R... SNVLNPFSLDIWDPFQ DYPL ITSSGTSSEFG....K ETAAFANTHIDW KETPQAHVFKA DLPGLKKEEVKVEVEEGKVLQISGERNKE ...KEE
D.carrota 17.8       GR R... SNVFDPFSLDVWDPFK DFPL VTSSA..SEFG....K ETAAFVNTHIDW KETPQAHVFKA DLPGLKKEEVKVEVEEGKVLQISGERNKE ...KEE
M.sativus 18.1       ...... ....DPFSLDVWDPFK DFPF TNSALSASSFP....Q ENSAFVSTRIDW KETPEAHVFKA DLPGLKKEEVKVEIEDDRVLQISGERNVE ...KED
M.sativus 18.2       GR R... SNVFDPFSLDVWDPFK DFPF NNSALSA.SFP....R ENSAFVSTRVDW KETPEAHVFKA DLPGMKKEEVKVEIEDDRVLQISGERSVE ...KED
P.sativum 18.1       GR R... SNVFDPFSLDVWDPLK DFPF SNSSPSA.SFP....R ENPAFVSTRVDW KETPEAHVFKA DLPGLKKEEVKVEVEDDRVLQISGERSVE ...KED
G.max 17.5           GR R... SNVFDPFSLDVWDPFK DFHF PTSLSA.......... ENSAFVNTRVDW KETPEAHVFEA DIPGLKKEEVKVQIEDDRVLQISGERNLE ...KED
G.max 17.3           GR R... SSVFDPFSLDVWDPFK DFPF PSSLSA.......... ENSAFVSTRVDW KETPEAHVFKA DIPGLKKEEVKLEIQDGRVLQISGERNVE ...KED
G.max 18.5           GR R... NNVFDPFSLDVWDPFK DFPF PNTLSSASFPEFSR.. ENSAFVSTRVDW KETPEAHVFKA DIPGLKKEEVKVQIEDDKVLQISGERNVE ...KED
G.max 17.6           GP R... SNVFDPFSLDMWDPFK DFHV PTSSVSA......... ENSAFVNTRVDW KETQEAHVLKA DIPGLKKEEVKVQIEDDRVLQISGERNVE ...KED
L.esculentum 17.8    DR R..S SSMFDPFSIDVFDPFR ELGF PSTNSG.......... ESSAFANTRIDW KETPEPHVFKV DLPGLKKEEVKVEVEEDRVLQISGERNVE ...KED
A.thaliana 17.6      GR R... TNVFDPFSLDVFDPFE GFLT P.SGLANAP.A....M DVAAFTNAKVDW RETPEAHVFKA DLPGLRKEEVKVEVEDGNILQISGERSNE ...NEE
A.thaliana 17.4      GR R... TNVFDPFSLDVWDPFE GFLT P..GLTNAP.A....K DVAAFTNAKVDW RETPEAHVFKA DVPGLKKEEVKVEVEDGNILQISGERSSE ...NEE
A.thaliana 18.2      GR R... SNVFDPFSQDLWDPFE GFFT PSSALANASTA....R DVAAFTNARVDW KETPEAHVFKA DLPGLKKEEVKVEVEDKNVLQISGERSKE ...NEE
H.annuus 17.6        SK R... SNIFDPFSLDTWDPFQ GII. ......STEPA....R ETAAIVNARIDW KETPEAHVLKA DLPGMKKEEVKVEVEDGRVLQISGERCRE ...QEE
P.sativum 17.9       GR R... TNAFDPFSLDLWDPFQ NFQL ARSATGTTN....... ETAAFANAHIDW KETPEAHVFKA DLPGVKKEEVKVEIEEDRVLKISGERKTE ...KED
T.aestivum 16.9b     .R R... SNVFDPFADLWADPFD T..F R.SIVPAI...SGGSS ETAAFANARVDW KETPEAHVFKV DLPGVKKEEVKVEVEDGNVLVVSGERSRE ...KED
T.aestivum 16.9c     ...... ..............DT ..FR .SIVPAI....SGGTS ETAAFANARVDW KETPEAHVFKA DLPGVKKEEVKVEVEDGNVLVVSGERTKE ...KED
T.aestivum 16.9a     .R R... TNVFDPFADLWADPFD T..F R.SIVPAI...SGGGS ETAAFANAEMDW KETPEAHVFKA DLPGVKKEEVKVEVEDGNVLVVSGERTKE ...KED
O.sativa 16.9        .R R... SNVFDPFSLDLWDPFD SV.F R.SVVPA....TSDN. DTAAFANARIDW KETPESHVFKA DLPGVKKEEVKVEVEEGNVLVISGQRSKE ...KED
Z.mays 17.2          .R R... SNVFDPFSMDLWDPFD TM.F RSIVPSA...TSTN.S ETAAFASARIDW KETPEAHVFKA DLPGVKKEEVKVEVEDGNVLVISGQRSRE ...KED
O.sativa 17.4        .R R... SNVFDPFSLDLWDPFD GFPF G..SGSGSL.FPRANS DAAAFAGARIDW KETPEAHVFKA DVPGLKKEEVKVEVEDGNVLQISGERIKE ...QEE
C.rubrum 18.3        GR R... SNIFDPFSLDEIWDPF FGLP ...STLSTVPRSETAA ETAAFANARIDW KETPEAHVFKA DLPGVKKEEVKVEVEDGNVLRISGQRARE ...KEE
P.sativum 22         LL PFID SPNTLL.SDLWSDRFP DPFR VLEQIPYGVEKHEPSI TLSHA...RVDW KETPEGHVIMV DVPGLKKDDIKIEVEENRVLRVSGERKKE ...EDK
G.max 22             LL PFMD PPITLL.ADLWSDRFP DPFR VLEHIPFGVDKDEASM AMSPA...RVDW KETPEGHVIML DVPGLKREEIKVEVEENRVLRVSGERKKE ...EEK
A.thaliana 22        LS SALE TTPGSLLSDLWLDRFP DPFK ILERIPLGLERDT.SV ALSPA...RVDW KETAEGHEIML DIPGLKKDEVKIEVEENGVLRVSGERKRE ...EEK
                                                                           #  ##             #  *#        #  ##     #
```

FIGURE 1.—*Continued*

removed because they are under very different selective pressures than the rest of the proteins and evolve very quickly. The third codon positions were removed after it was determined that, in most pairwise comparisons, synonymous substitutions were saturated *i.e.*, greater than two substitutions per site. The topology of the trees generated using complete sequences and without the transit sequences and third positions were almost identical. Removal of the transit sequences and third positions decreased resolution for some closely related sequences but significantly increased the overall consistency index. The tree presented in this paper was constructed from data matrices in which the transit sequences and third positions were removed. There were 311 informative sites in the DNA data matrix. A 5:1 transitions:transversions weighting was used because this ratio was found to be the empirical values for these substitutions among the plant small heat-shock protein data.

Amino acid distances were generated with Protdist in PHYLIP using the categories option. The distance matrices were then used to construct trees with the neighbor joining (NJ) method. One hundred bootstrap replicates were generated using Seqboot and the consensus trees generated in Consense.

Rate analysis: Estimates of synonymous (Ks) and nonsynonymous (Ka) substitutions were generated by the program Li93 (LI 1993). Positions that included gaps were removed from the analysis. Estimates of the number of conservative and radical amino acid replacement substitutions per site were generated by the program SCR-PC (HUGHES *et al.* 1990). Statistical significance of pairwise comparisons were estimated with T tests.

RESULTS

Sequence conservation and divergence among small heat-shock proteins: The small heat-shock proteins are more conserved, across protein families, in the carboxyl-terminal (C-terminal) domain than in the amino-terminal (N-terminal) domain. In the N-terminal domain (amino-acids 1–152) there are family specific conserved regions (Figure 1). The chloroplast (CP)-, mitochondrial (MT)- and endoplasmic reticulum (ER)-localized proteins all have transit sequences that are specific for each organelle (Figure 1). The CP-localized proteins also have a Met-rich region (amino acids 103–124) in the N-terminal domain (Figure 1 and VIERLING 1991). The class I cytosolic proteins have a consensus region in the N-terminal region (amino acids 107–120). The class II cytosolic proteins also have a small conserved region (amino acids 143–154) not present in the other protein classes at the very end of the N-terminal region.

The alignment of the small heat-shock proteins clearly shows the higher conservation in the C-terminal

```
                              210       220       230       240        250           260       270       280

T. aestivum 26a      GEGGDGWWKERSVSS YDMRLAL.PDECDKSQVRAELKNGVLLVSV PKR..  ...............ETERKVIDVQVQ......
T. aestivum 26b      GEGGDGWWKERSLSS YDMRLAL.PDECDKSQVRAELKNGVLLVSV KPR..  ...............ETERKVIDVQVQ......
Z. mays 26           GGDGDGWWKQRSVSS YDMRLAL.PDECDKSKVRAELKNGVLLVTV PKT..  ...............EVERKVIDVQVQ......
P. sativum 21        GED...CWSRKSYSC YDTRLKL.PDNCEKEKVKAELKDGVLYITI PKT..  ...............KIERTVIDVQIQ......
G. max 21            GDD...SWSSRTYSS YDTRLKL.PDNCEKDKVKAELKNGVLYITI PKT..  ...............KVERKVIDVQVQ......
A. thaliana 21       ..DSDDSWSGRSVSS YGTRLQL.PDNCEKDKIKAELKNGVLFITI PKT..  ...............KVERKVIDVQIQ......
P. hybrida 21        .SGKDDSWGRN.YSS YDTRLSL.PDNVDKDKVKAELKNGVLLISI PKT..  ...............KVEKKVTDVEI.......
C. rubrum 23         ......TEEEEQRRR YSSRIELTPNLYKIDGIKAEMKNGVLKVTV PKI..  ...............KEEEKKDVFQVMVD....
L. longiflorum 18.2  ....KYQIMERWTGR RMRKFER.PKNRDTKAVSAVWKNGVLAVTV GKLLA  WEVAGLFFNIERLPVPLPTKTKSIEVKIEVKIA
L. longiflorum 17.6  ....KYQMMERWTGK RMRKFEL.PENADTKAVSAVWKNGVLAVTV RKLPA  WEVAGISFNIERLPVPLPTKTKSIEVKIA....
L. longiflorum 16.5  ....RYLEMQRRMGK MMRKFKL.LENANSGAISAVCKNGVLTVTV EKLPS  ...............QEPK...AIEIKIA....
Z. mays 17.8         .DDAKYLRMERRMGK FMRKFVL.PDNADVDKVAAVCRDGVLTVTV EKLPP  ...............PEPKKPKTIEVKVA....
Z. mays 17.5         ..DAKYLRMERRMGK FMRKFVL.PDNADMDKISAVCRDGVLTVTV EKLPP  ...............PEPKKPKTIEVKVA....
T. aestivum 17.3     ..DAKYLRMERRMGK LMRKFVL.PENADMEKISP.CRDGVLTVTV DKLPP  ...............PEPKKPKTIQVQVA....
P. sativum 17.7      KEGVKYLKMERRIGK LMRKFVL.PENANIEAISAISQDGVLTVTV NKLPP  ...............PEPKKPKTIQVKVA....
G. max 17.9          KEGAKYLRMERRVGK LMRKFVL.PENANTDAISAVCQDGVLSVTV QKLPP  ...............PEPKKPRTIQVKVA....
I. nil 172           KEGAKYVRMERRVGK LMRKFVL.PENANKEKITAVCQDGVLTVTV ENVPP  ...............PEPKKPRTIEVKIG....
I. nil 18. 8         KDGVKYLRMERRVGK FMRKFVL.PENANVEAINAVYQDGVLQVTV EKLPP  ...............PEPKKPKTVEVKVA....
A. thaliana 17.6II   .EGVKYVRMERRMGK FMRKFQL.PENADLDKISAVCHDGVLKVTV QKLPP  ...............PEPKKPKTIQVQVA....
D. carrota           .KNDKWHPLEVSSGK FLRRFRL.PENANVDEVKAGMENGVLTVTV PKVE.  ...............MKKPEVKSIHISG.....
D. carrota 17.8      .KNDKWHRVERSSGK FLRRFRL.PENAKVDEVKAAMANGVVTVTV PKVE.  ...............IKKPEVKAIDISG.....
M. sativus 18.1      .KNDQWHRVERSSGK FMRRFRL.PENAKMDQVKAAMENGVLTVTV PKEE.  ...............IKKPEVKSIEISS.....
M. sativus 18.2      .KNDQWHRLERSSGK FMRRFRL.PENAKMDQVKAAMENGVLTVTV PKEE.  ...............VKKPEVKTIDISG.....
P. sativum 18.1      .KNDQWHRVERSSGK FMRRFRL.PENAKMDQVKAAMENGVLTVTV PKEE.  ...............IKKAEVKSIEISG.....
G. max 17.5          .KNDTWHRVERSSGN FMRRFRL.PENAKVEQVKASMENGVLTVTV PKEE.  ...............VKKPDVKAIEISG.....
G. max 17.3          .KNDTWHRVERSSGK LVRRFRL.PENAKVDQVKASMENGVLTVTV PKEE.  ...............IKKPDVKAIDISG.....
G. max 18.5          .KNDTWHRVERSSGK FMRRFRL.PENAKVEQVKASMENGVLTVTV PKEE.  ...............VKKPDVKAIEISG.....
G. max 17.6          .KNDTWHRVDRSSGK FMRRFRL.PENAKVEQVKACMENGVLTVTI PKEE.  ...............VKKSDVKPIEISG.....
L. esculentum 17.8   .KNDKWHRMERSSGK FMRRFRL.PENAKMDQVKASMENGVLTVTV PKEE.  ...............VKKPEVKSIEISG.....
A. thaliana 17.6     .KNDKWHRVERSSGK FTRRFRL.PENAKMEEIKASMENGVLSVTV PKVP.  ...............EKKPEVKSIDISG.....
A. thaliana 17.4     .KSDTWHRVERSSGK FMRRFRL.PENAKVEEVKASMENGVLSVTV PKVQ.  ...............ESKPEVKSIDISG.....
A. thaliana 18.2     .KNDKWHRVERASGK FMRRFRL.PENAKMEEVKATMENGVLTVVV PKAP.  ...............EKKPQVKSIDISGAN...
H. annuus 17.6       .KDDTWHRVERSSGK FIRRFRL.PENAKMDEVKAMMENGVLTVVV PKEE.  ...............EKKPMVKAIDISG.....
P. sativum 17.9      .KNDTWHRVERSQGS FLRRFRL.PENAKVDQVKAAMENGVLTVTV PKEE.  ...............VKKPEAKPIQITG.....
T. aestivum 16.9b    .KNDKWHRVERSSGK FVRRFRL.PEDAKVEEVKAGLENGVLTVTV PKAE.  ...............VKKPEVKAIEISG.....
T. aestivum 16.9c    .KNDKWHRVERSSGK FVRRFRL.PEDAKVEEVKAGLENGVLTVTV PKAE.  ...............VKKPEVKAIEISG.....
T. aestivum 16.9a    .KNDKWHRVERSSGK FVRRFRL.LEDAKVEEVKAGLENGVLTVTV PKAE.  ...............VKKPEVKAIQISG.....
O. sativa 16.9       .KNDKWHRVERSSGQ FMRRFRL.PENAKVDQVKAGLENGVLTVTV PKAE   ...............VKKPEVKAIEISG.....
Z. mays 17.2         .KDDKWHRVERSSGQ FIRRFRL.PDDAKVDQVKAGLENGVLTVTV PKAE.  ...............EKKPEVKAIEISG.....
O. sativa 17.4       .KTDKWHRVERSSGK FLRRFRL.PEDTKPEQIKASMENGVLTVTV PKEE.  ...............PKKPDVKSIQITG.....
C. rubrum 18.3       .KNDTWHRVERSSGQ FMRKFRL.PENAKVDQVKAGMENGVLTVTV PKNE.  ...............APKPQVKAINVY......
P. sativum 22        .KGDHWHRVERSYGK FWRQFKL.PQNVDLDSVKAKMENGVLTLTL HKLSH  DKIKGPRMVSIVEEDDKPSKIVNDELK......
G. max 22            .KGDHWHRVERSYGK FWRQFRL.PQNVDLDSVKAKLENGVLTLTL DKLSP  GKIKGPRVVSIAGEDHQQGNLNNDGAKQEL...
A. thaliana 22       .KGDQWHRVERSYGK FWRQFKL.PDNVDMESVKAKLENGVLTINL TKLSP  EKVKGPRVVNIAAEEDQTAKISSSESKEL....
                              #            # #         #    **# ###
```

FIGURE 1.—*Continued*

domain (amino acids 152–282) (Figure 1). This domain contains four completely conserved and 15 highly conserved amino acids. The plant small heat-shock proteins share a consensus region (amino acids 166–193) (Figure 1 and VIERLING 1991) not present in other eukaryotic small heat-shock proteins. All plant small heat-shock proteins also share a eukaryotic HS region (amino acids 214–250). The proline . . . glycine, valine, leucine amino-acid motif (amino acids 224, 239, 240, 241) in the HS domain is highly conserved among all eukaryotic small heat-shock proteins. This motif is highly conserved in the plant small heat-shock proteins. In the class II *Lilium longiflorum* HSP 16.5 and in *Triticum aestivum* HSP 16.9b the proline has been replaced by a leucine. The leucine at position 241 has been replaced by a valine in *Daucus carota* HSP 17.8.

Phylogenetic relationships of the small heat-shock proteins: To determine paralogous and orthologous relationships among the small heat-shock proteins, aligned amino acid and DNA sequences were analyzed using both distance (NJ)- and parsimony-based phylogenetic programs. Results from all of the analyses support the conclusion that the five major gene families form monophyletic groups and are most likely the result of gene duplications that occurred before the diversification of the angiosperms (Figures 2 and 3). The NJ tree generated from DNA distance matrices and the parsimony trees generated from amino acid data matrices are not shown but are highly congruent with the trees presented. In the NJ and parsimony trees the branches for individual gene families are highly supported by bootstrap analysis (Figures 2 and 3). It is not possible to deduce from this analysis the order of gene duplication events that gave rise to the five families, although the presence of both monocot and dicot sequences within each family indicates that the duplications occurred before the divergence of these two groups.

The class I cytosolic gene family contains paralogous genes. The phylogenetic relationships among the class I sequences are not always congruent with organismal relationships. The dicot sequences *H. annuus* HSP 17.6, *C. rubrum* HSP 18.3 and the *P. sativum* HSP 17.9 are consistently more closely related to the monocot (*T. aestivum, Z. mays* and *O. sativa*) sequences than to the other dicot sequences (Figures 2 and 3). This indicates that there have been duplications within the class I family.

There is evidence of gene conversion within the class I gene family: With the exception of the *P. sativum* HSP 17.9 and 18.1, and the *O. sativa* HSP 17.4 and 16.9 sequences, class I sequences from a single species are

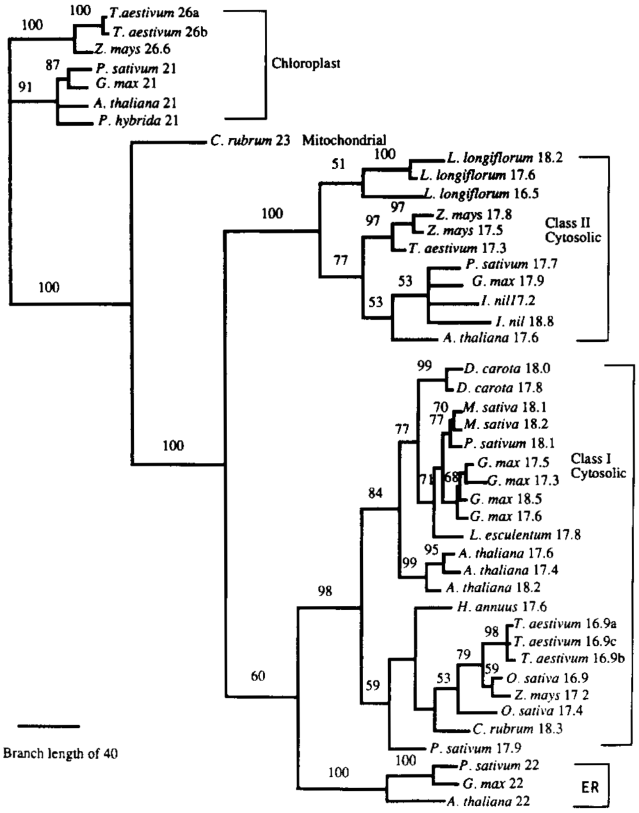

FIGURE 2.—Parsimony tree based on DNA sequences. Strict consensus of the six most parsimonious trees. Tree length, 1619; consistency index, 0.456. Branch lengths are proportional to changes found along the branches. The tree is rooted with the sequences for the CP-localized proteins. The number of times out of the 100 bootstrap replicates that a branch was present is noted above the branch; values below 50 are not noted.

each other's closest relatives (Figures 2 and 3). This pattern suggests that gene conversion is homogenizing some of the class I sequences. Separate parsimony analysis of the DNA sequences coding for the N-terminal and C-terminal domains have the same topology (data not shown), suggesting that if gene conversion is occurring it is not localized to one part of the genes.

Duplication and divergence of class II sequences: The class II genes from *L. longiflorum, Z. mays* and *I. nil* are developmentally and differentially expressed (BOUCHARD 1990; KRISHNA *et al.* 1992; KOBAYASHI *et al.* 1994). However, nothing is known about the function of these proteins. I examined the rates of nucleotide substitution and amino acid replacements for evidence of functional divergence among the class II proteins.

Sequences from *L. longiflorum* were isolated from meiotic cDNA libraries generated from microgametophyte tissue (BOUCHARD 1990; KOBAYASHI *et al.* 1994). *L. longiflorum* HSP 18.2 is induced by both meiosis and heat (BOUCHARD 1990); *L. longiflorum* HSP 17.6 and 16.5 are expressed during meiosis and it is not known if they are also expressed during heat shock (KOBAYASHI 1994). All three *L. longiflorum* proteins are clearly class II small heat-shock proteins although 18.2 and 17.6 have lost part (six amino acids) of the class II consensus region. Pairwise comparisons of the class II *L. longiflorum* sequences show an interesting pattern of sequence divergence, in that the DNA sequences are more similar than the corresponding amino acid sequences (Table 2). This pattern of similarity was not found in any of the other pairwise comparisons of the other plant small heat-shock proteins. On closer inspection the DNA alignments revealed that many of the third codon positions were conserved among these sequences while first and second codon positions were not. There are no significant differences in percentage G + C content or codon usage among the Lilium genes.

To explore this pattern of sequence divergence in more detail synonymous and nonsynonymous substitutions among the *L. longiflorum* genes were examined. Comparisons were made with complete sequences (Table 3). In addition class II sequences from *I. nil* and *Z. mays* were examined. The *I. nil* HSP 18.8 gene is induced by both heat-shock and the photoperiod changes that induce flowering, whereas 17.2 is induced by heat shock alone (KRISHNA *et al.* 1992). *Z. mays* HSP 17.5 is induced by heat shock and during pollen development (meiosis); while *Z. mays* HSP 17.8 is induced only by heat shock (ATKINSON *et al.* 1993).

When protein sequences are constrained by function, synonymous substitutions (Ks) are expected to be significantly higher than the nonsynonymous substitutions (Ka). In most, but not all, of the pairwise comparisons of the class II gene sequences the number of synonymous substitutions were higher than the number of nonsynonymous substitutions. The Ks between both *L. longiflorum* HSP 18.2 and 16.5, and *L. longiflorum* HSP 17.6 and 16.5 is not significantly greater than Ka (Table 3A).

The pattern of nonsynonymous substitutions was examined using the program of HUGHES *et al.* (1990), which distinguishes between conservative and radical amino acid replacements. Proteins under strong selection to maintain function are expected to have more conservative (within the same amino acid chemical group) than radical replacements (across chemical groups). In comparisons of the class II sequences, I used the category of hydrophobicity, since hydrophobicity is conserved in the C-terminal domain among all the eukaryotic small heat-shock proteins (NOVER 1990). It is hypothesized (NOVER 1990) that the conserved hydropathy profiles of these proteins reflect strong selective constraints related to the ability of the small heat-shock proteins to form oligomers.

Comparisons of the *L. longiflorum* HSP 18.2 and 17.6 genes reveal that although Ks is higher than Ka, conservative replacements are not significantly more frequent than radical replacements (Table 3). Between *L. longiflorum* HSP 18.2 and 16.5, Ks is not significantly greater than Ka. However, conservative replacements are significantly more frequent than radical replace-

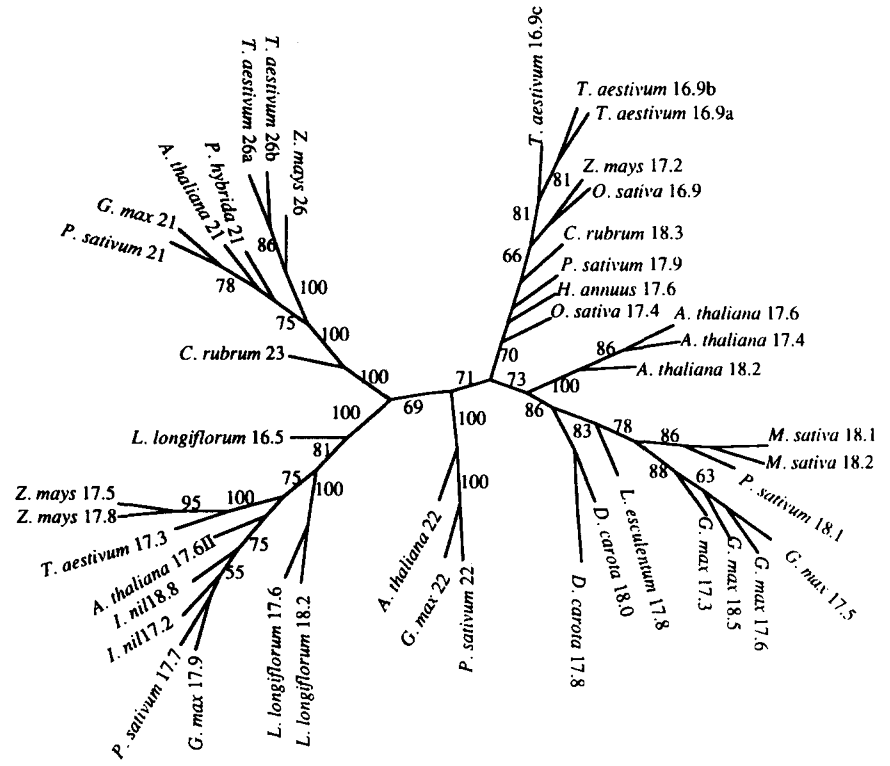

FIGURE 3.—NJ tree based on amino acid sequences. The number of times out of the 100 bootstrap replicates that a branch was present is noted above the branch; values below 50 are not noted.

ments (Table 3). The *I. nil* sequences that are differential expressed do not have significantly more conservative than radical replacement substitutions (Table 3).

Small heat-shock proteins do not evolve at equal rates: Relative rate tests (WU and LI 1985) were conducted within the gene families *i.e.*, CP, II, I and ER to see if there are any differences in evolutionary rates within gene families. No evidence of differences in evolutionary rates within families (data not shown) were found.

It was then determined if rates of substitution were variable among gene families. To examine this, rates of evolution were compared among species pairs from which sequences of at least three gene families are available. Taxa were examined in pairs to control for organismal divergence time. For example, the divergence time should be the same for all of the genes in *Z. mays* and *T. aestivum*. If all of the small heat-shock protein genes are evolving at the same rate, then the number of substitutions per site between each family of orthologous genes (*e.g.*, between the CP and ER genes) of *Z. mays* and *T. aestivum* should be the same.

TABLE 2

Pairwise comparisons of *Lilium longiflorum* sequences

Comparison	DNA identity (%)	Amino acid identity (%)
18.2 *vs* 17.6	89.4	85.8
18.2 *vs.* 16.5	63.0	46.3
17.6 *vs.* 16.5	73.6	53.3

Percentage identity was estimated with GAP in GCG.

Rates of nonsynonymous substitution: I examined the total number of Ka of the complete gene sequences and of the portion of the genes coding for the N and C terminal domains (data not shown). The class II and ER genes are evolving more quickly than the CP and class I genes (Table 4). The genes for the ER proteins had a consistently higher Ka than the CP and class I genes. The class II genes also had a higher Ka than the CP and class I genes, but this difference in rate was not statistically significant in the *Z. mays vs. T. aestivum* comparison (Table 4). Compared to the other gene families the class II genes had significantly higher Ka values in the portion of the genes coding for the N-terminal domain (data not shown). The gene families had more similar Ka values in the portion of the genes coding for the C-terminal domain (data not shown).

Rates of conservative and radical amino acid replacements: The nonsynonymous substitutions were examined in more detail and designated as conservative and radical according to hydrophobicity. The CP, class II and ER proteins had significantly higher conservative than radical replacements (Table 5). This pattern is expected under strong selection if hydrophobicity is important for function. None of the class I gene comparisons had significantly higher conservative than radical replacements. However the class I genes had significantly more conservative substitutions than radical replacement substitutions in the portion of the gene coding for the C-terminal domain (data not shown).

DISCUSSION

Increasing complexity of gene families reflects the inceasing complexity of organisms and functional di-

TABLE 3

Pairwise comparisons of class II sequences

	Ks	Ka
A. Pairwise estimates of synonymous (Ks) and nonsynonymous (Ka) substitutions per site for the Class II sequences		
L. longiflorum 18.2 *vs.* 17.6	0.233 ± 0.096	0.104 ± 0.027***
L. longiflorum 18.2 *vs.* 16.5	0.512 ± 0.157	0.417 ± 0.068
L. longiflorum 17.6 *vs.* 16.5	0.287 ± 0.090	0.284 ± 0.051
I. nil 17.2 *vs.* 18.8	0.650 ± 0.212	0.149 ± 0.034***
Z. mays 17.8 *vs.* 17.5	0.182 ± 0.080	0.040 ± 0.017***
	Con	Rad
B. Pairwise estimates of conservative (Con) and radical (Rad) substitutions among the Class II sequences		
L. longiflorum 18.2 *vs.* 17.6	0.100 ± 0.021	0.060 ± 0.022
L. longiflorum 18.2 *vs.* 16.5	0.383 ± 0.038	0.261 ± 0.046***
L. longiflorum 17.6 *vs.* 16.5	0.275 ± 0.038	0.226 ± 0.043
I. nil 17.2 *vs.* 18.8	0.203 ± 0.027	0.167 ± 0.035
Z. mays 17.8 *vs.* 17.5	0.063 ± 0.017	0.024 ± 0.014***

Values are means ± SE. * indicates that Ks is significantly greater than Ka at the 0.05 probability level, ** 0.01 level and *** 0.001 level.

versification of gene products (Ohta 1991). Gene duplication has long been recognized as an important process in genome evolution. Once a gene duplicates, the new copy can accumulate substitutions and eventually diverge enough that a new function becomes possible. Gene duplication and divergence has been examined theoretically (Nagalaki 1984; Walsh 1987, 1995; Ohta 1988a–c, 1991). This study has shown that gene duplication, sequence divergence and gene conversion have all played a role in the evolution of the small heat-shock protein genes in plants. The small heat-shock protein genes have evolved from the single gene found in most animals and fungi into a large super gene family in angiosperms. The diversification of small heat-shock proteins in plants may reflect molecular adaptations to stressful conditions unique to plants as well as evolution of functions not related specifically to high temperature stress. Analysis of patterns of substitutions reveals that the selective constraints on the small heat-shock protein gene families are not identical. Differences in selective constraint frequently reflect functional differences. This suggests that functional divergence has occurred among the small heat-shock proteins in plants.

Evolutionary relationships among small heat-shock protein gene families: The order of the gene duplications that gave rise to the five small heat-shock protein gene families is not known and cannot be deduced from the phylogenetic analysis of the available sequences. More data on small heat-shock proteins in

TABLE 4

Comparisons of nonsynonymous substitutions among gene families

Species comparison	CP	Class II	Class I	ER
A. Nonsynonymous substitutions, Ka, per site				
T. aestivum vs. Z. mays	0.063 ± 0.015	0.086 ± 0.025	0.080 ± 0.017	
G. max vs. P. sativum	0.083 ± 0.017	0.110 ± 0.019	0.063 ± 0.015	0.104 ± 0.017
A. thaliana vs. P. sativum	0.143 ± 0.022	0.219 ± 0.029	0.136 ± 0.022	0.209 ± 0.025
A. thaliana vs. G. max	0.121 ± 0.012	0.255 ± 0.031	0.106 ± 0.027	0.231 ± 0.028

Species comparison	CP *vs.* II	CP *vs.* I	CP *vs.* ER	II *vs.* I	II *vs.* ER	I *vs.*ER
T. aestivum vs. Z. mays	NS	NS		NS		
G. max vs. P. sativum	NS	NS	NS	**	NS	***
A. thaliana vs. P. sativum	***	NS	***	***	NS	***
A. thaliana vs. G. max	***	NS	***	***	NS	***

Values are means ± SE. * indicaes that the Ka for the two gene classes are different at the 0.05 probability level, ** 0.01 level and *** 0.001 level; NS indicates that the Ka for the two genes are not statistically different.

TABLE 5

Comparisons of conservative and radical substitutions among gene families

Species comparisons	CP		Class II		Class I		ER	
	Con	Rad	Con	Rad	Con	Rad	Con	Rad
T. aestivum vs.	0.105	0.048	0.149	0.033	0.115	0.089		
Z. mays	± 0.012	± 0.019**	± 0.0245	± 0.016***	± 0.022	± 0.027		
G. max vs.	0.123	0.018	0.148	0.052	0.088	0.072	0.108	0.084
P. sativum	± 0.022	± 0.012***	± 0.024	± 0.021***	± 0.019	± 0.025	± 0.019	± 0.024
A. thaliana vs.	0.145	0.070	0.254	0.094	0.195	0.170	0.210	0.132
P. sativum	± 0.024	± 0.025***	± 0.029	± 0.028***	± 0.026	± 0.035	± 0.025	± 0.030***
A. thaliana vs.	0.131	0.073	0.273	0.131	0.211	0.172	0.230	0.122
G. max	± 0.022	± 0.025**	± 0.030	± 0.031***	± 0.027	± 0.036	± 0.026	± 0.028***

Values are means ± SE. Con, conservative; Rad, radical. * indicates that Con is greater than Rad at the 0.05 probability level, ** 0.01 level, and *** 0.001 level.

early plants will be needed to determine the order of gene duplications. This work is in progress.

Plesofsky-Vig *et al.* (1992) hypothesized that the CP-localized protein may have been transferred to the plant nucleus from a photosynthetic endosymbiont and therefore the CP protein family is only distantly related to the other plant small heat-shock protein families. The sequence conservation among the small heat-shock proteins argues against the hypothesis of an endosymbiotic origin of the CP protein. All of the plant small heat-shock proteins share a plant consensus region in the C-terminal domain, in addition to the heat-shock region that is shared with other eukaryotic small heat-shock proteins. The plant consensus region is not conserved in other eukaryotic small heat-shock proteins (Vierling 1991; Plesofsky-Vig *et al.* 1992; Jong *et al.* 1993). If the CP proteins were bacterial in origin, they would not share this region with the other plant small heat-shock proteins. It is then more likely that early in the plant lineage a single small heat-shock protein gene existed that had the plant consensus region. Multiple duplications of this gene gave rise to the many small heat-shock protein gene families early in the evolution of plants (*i.e.*, at least before the rise of the angiosperms).

Evolutionary relationships within small heat-shock protein gene families: The relationships of the genes for CP- and ER-localized proteins are congruent with organismal relationships and therefore these two gene families are most likely composed of orthologous genes. The phylogenetic relationships among the class I sequences is however more complex.

The phylogenetic relationships of the class I sequences suggests that gene conversion is occurring among some but not all of the class I genes. When gene conversion is frequent, all paralogous genes involved in the gene conversion event will be each others closest relatives in a phylogenetic analysis (Sanderson and Doyle 1992). When gene conversion does not occur at all or very infrequently, each group of paralogous genes will reflect organismal relationships (Sanderson and Doyle 1992). The class I sequences from *D. carota, M. sativum, G. max* and *A. thaliana* are all most closely related to other con-specific class I genes, *i.e., A. thaliana* HSP 17.6, 17.4 and 18.2. This pattern suggests that either there are new duplications in each species, or, more likely, that gene conversion is maintaining sequence similarity among class I genes. The relationships of the class I sequences could also be explained by numerous independent duplications within each lineage. However if gene duplications were this frequent, one would expect to see many more small heat-shock proteins than have been observed.

The sequence divergence among the class I genes within species suggests that while gene conversion occurs it is not frequent. In a study of globin genes, Fitch *et al.* (1991) were able to detect which portion of the gene was undergoing gene conversion by constructing trees using different regions of the globin genes. Trees constructed separately from small heat-shock protein gene sequences for the N and C terminal domain had the same topology as the trees based on the entire gene sequence. This indicates that gene conversion is not limited to either the N or C terminal domains. A similar pattern to that seen with the small heat-shock protein genes was reported with the genes for the small subunit of ribulose bisphosphate carboxylase (Meager *et al.* 1989).

Comparisons of some of the class II genes suggest that functional divergence is occurring within the class II family. It has been previously established that some of the class II genes are developmentally expressed. However it is not known if the differences in expression reflect differences in function. The class II genes in both *Z. mays* (Atkinson *et al.* 1993) and *L. longiflorum* (Bouchard 1990) are expressed during heat-shock and flower development. The *I. nil* HSP 17.2 gene is induced during heat shock and is also induced by changes in photoperiod (Krishna *et al.* 1992). The *I. nil* HSP 18.8 gene is heat-inducible but is not induced by changes in photoperiod (Krishna *et al.* 1992).

In comparisons of both *Z. mays* and *I. nil* class II sequences synonymous substitutions are significantly greater than nonsynonymous substitutions. However, the patterns of amino-acid replacement substitution (conservative *vs.* radical) between the *I. nil* small heat-shock protein genes indicates that there may be functional divergence among the *I. nil* small heat-shock proteins.

Rapid divergence after gene duplication has been reported for other genes (Li and Gojobori 1983; Li 1985; Goodman *et al.* 1987). In these cases there was enough phylogenetic information to place the timing of gene duplications on a phylogenetic tree and to asses rates of nonsynonymous substitutions before and after the duplication events. These studies show that while the rate of nonsynonymous substitution may be high immediately after duplication, this rate does eventually slow down. The difficulty with interpreting the *L. longiflorum* data is that we do not have sequences from other closely related organisms and so it is not possible to date the duplications. There may be no selective constraint on the *L. longiflorum* HSP 16.5 kDa protein at all; it may be drifting with neutral substitutions. Another possibility is that after the duplication event this gene had a burst of nonsynonymous substitutions but is now under selection to maintain a new function. The ratio of Ks to Ka is both a function of selective constraints and the time since duplication. The equality in rates of synonymous and nonsynonymous substitutions may reflect the fact that the synonymous substitutions, which accumulate as function of time, are now reaching the level of the nonsynonymous substitutions. If these genes were sampled sometime in the future, Ks would be higher than Ka.

It is unlikely that the *L. longiflorum* HSP genes are pseudogenes. They are expressed and they do not have any misplaced start or stop codons. They have the conserved class II consensus region, in addition to the conserved plant heat-shock domain and the eukaryotic heat-shock domain. If they were pseudogenes, they would accumulate amino acid replacements at the same rate across the entire sequence and these conserved regions would not be maintained. Most likely the *L. longiflorum* genes are recently duplicated genes that are in the process of diverging in both sequence and function from an ancestral gene. More complete sampling within Lilium and related taxa will be needed before this can be determined with greater confidence.

Selective constraints among the small heat-shock protein gene families: The differences in evolutionary rate among the small heat-shock protein gene families found in this study suggest that these gene families have diverged in function. Equality of rates of nonsynonymous substitutions indicate that proteins are under similar selective constraints. The CP proteins have significantly fewer nonsynonymous substitutions than the class II sequences. The ER and class II sequences have significantly more nonsynonymous substitutions than the class I sequences. There are also differences in the ratio of conservative to radical amino acid replacement substitutions among the gene families. If the ratio of conservative to radical replacements reflects functional constraints, then the class I sequences are functionally distinct from the other classes. Recent *in vitro* studies indicate that some small heat-shock proteins can act as molecular chaperones (Jakob *et al.* 1993; Merck *et al.* 1993; Jakob and Buchner 1994; Lee *et al.* 1995). If the small heat-shock proteins are molecular chaperones, the differences in selective constraint revealed by this study suggest that the individual small heat-shock protein families may have very different substrate specificities. It is also possible that some small heat-shock protein families may have evolved entirely new functions.

The evolution of the small heat-shock proteins in plants from a single gene to a very large multigene family composed of at least five gene families is an important example of gene family diversification. The application of molecular evolutionary analysis to DNA and amino acid sequences of unknown function can help to establish paralogous groupings and, most importantly, can identify possible instances of functional divergence. The assumption underlying this analysis is that sequence divergence reflects functional divergence. Where functional differences have already been established for other proteins (Karlin *et al.* 1992), this has proved to be true. Our ability to obtain DNA and amino acid sequences has far outstripped our ability to conduct detailed *in vitro* and *in vivo* studies of protein function. The use of sequence analysis can help in the formulation of hypotheses concerning function that can then be tested in the laboratory.

This work was supported by a postdoctoral fellowship from the Research Training Group in the Analysis of Biological Diversification at the University of Arizona. I thank G. Bharathan, N. Moran, R. Sheehy and E. Vierling for many helpful discussions. A. Colwell, G. Bharathan, L. McDade, N. Moran and E. Vierling provided useful comments on previous versions of this manuscript.

LITERATURE CITED

Arrigo, A.-P., and J. Landry, 1994 Expression and function of the low-molecular-weight heat-shock proteins, pp. 335–373 in *The Biology of Heat Shock Proteins and Molecular Chaperones,* edited by R. I. Morimoto. Cold Spring Harbor Laboratory Press, Plainview, NY.

Atkinson, B. G., M. Raizada, R. A. Bouchard, J. R. Frappier and D. Walden, 1993 The independent stage-specific expression of the 18-kDa heat shock protein genes during microsporogenesis in *Zea mays* L. Dev. Genet. **14:** 15–26.

Becker, J., and E. A. Craig, 1994 Heat-shock proteins as molecular chaperones. Eur. J. Biochem **219:** 11–23.

Beckmann, R. P., L. A. Mizaen and W. J. Welch, 1990 Interaction of Hsp 70 with newly synthesized proteins: implications for protein folding and assembly. Science **248:** 850–854.

Boorstein, W. R., T. Ziegelhoffer and E. A. Craig, 1994 Molecular evolution of the HSP70 multigene family. J. Mol. Evol. **38:** 1–17.

Bouchard, R. A., 1990 Characterization of expressed meiotic prophase repeat transcript clones of *Lilium:* meiosis-specific expression, relatedness, and affinities to small heat shock protein genes. Genome **33:** 68–79.

Craig, E. A., J. S. Weissman and A. L. Horwich, 1994 Heat shock proteins and molecular chaperones: mediator of protein conformation and turnover in the cell. Cell **78:** 365–372.

Eernisse, D. J., 1992 DNA translator and aligner—hypercard utilities to aid phylogenetic analysis of molecules. CABIOS **8:** 177–184.

Felsenstein, J., 1993 PHYLIP (Phylogeny Inference Package), Seattle.

Fitch, D. H. A., W. J. Bailey, D. A. Tagle, M. Goodman, L. Sieu *et al.,* 1991 Duplication of the Yn-globin gene is mediated by repetitive L1 LINE sequences in an early ancestor of simian primates. Proc. Natl. Acad. Sci. USA **88:** 7396–7400.

Genetics Computer Group, 1991 GCG Package. Genetics Computer Group, Madison, WI.

Gething, M.-J., and J. Sambrook, 1992 Protein folding in the cell. Nature **355:** 33–45.

Goodman, M., J. Czelusniak, B. F. Koop, D. A. Tagle and J. L. Slightom, 1987 Globins: a case study in molecular phylogeny. Cold Spring Harbor Symp. Quant. Biol. **52:** 875–890.

Hughes, A. L., 1993a Contrasting evolutionary rates in the duplicate chaperonin genes of *Mycobacterium tuberculosis* and *M. leprae.* Mol. Biol. Evol. **10:** 1343–1359.

Hughes, A. L., 1993b Nonlinear relationships among evolutionary rates identify regions of functional divergence in heat-shock protein 70 genes. Mol. Biol. Evol. **10:** 243–255.

Hughes, A., T. Ota and M. Nei, 1990 Positive darwinian selection promotes charge profile diversity in the antigen-binding cleft of class I major-histocompatibility-complex molecules. Mol. Biol. Evol. **7:** 515–525.

Jakob, U., and J. Buchner, 1994 Assisting spontaneity: the role of HSP90 and small HSPs as molecular chaperones. Trends Biochem. Sci. **19:** 205–211.

Jakob, U., M. Gaestel, K. Engel and J. Buchner, 1993 Small heat shock proteins as molecular chaperones. J. Biol. Chem. **268:** 1517–1520.

Jong, W. W. d., J. A. M. Leusnissen and C. E. M. Voorter, 1993 Evolution of the alpha-crystallin/small heat-shock protein family. Mol. Biol. Evol. **10:** 103–126.

Karlin, S., V. Brendel and P. Bucher, 1992 Significant similarity and dissimilarity in homologous proteins. Mol. Biol. Evol. **9:** 152–167.

Knack, G., Z. Liu and K. Kloppstech, 1992 Low molecular mass heat-shock proteins of a light-resistant photoautotrophic cell culture. Eur. J. Cell Biol. **59:** 166–175.

Kobayashi, T., E. Kobayashi, S. Sato, Y. Hotta, N. Miyajima *et al.,* 1994 Characterization of cDNAs induced in meiotic prophase in Lily micorosporocytes. DNA Research **1:** 15–26.

Krishna, P., R. F. Felsheim, J. C. Larkin and A. Das, 1992 Structure and light-induced expression of a small heat-shock protein gene of *Pharbitis nil.* Plant Physiol. **100:** 1772–1779.

Lee, G. J., N. Pokala and E. Vierling, 1995 Structure and molecular chaperone activity of cytosolic small heat shock proteins from pea. J. Biol. Chem. **270:** 10432–10438.

Lenne, C., and R. Douce, 1994 A low molecular mass heat-shock proteins is localized to higher plant mitochondria. Plant Physiol. **105:** 1255–1261.

Li, W.-H., 1985 Accelerated evolution following gene duplication and its implication for the neutralist-selectionist controversy, pp. 333–352 in *Population Genetics and Molecular Evolution,* edited by T. Ohta and K. Aoki. Japan Scientific Society, Tokyo.

Li, W.-H., 1993 Unbiased estimation of the rates of synonymous and nonsynonymous substitution. J. Mol. Evol. **36:** 96–99.

Li, W.-H., and T. Gojobori, 1983 Rapid evolution of goat and sheep globin genes following gene duplication. Mol. Biol. Evol. **1:** 94–108.

Lindquist, S., and E. A. Craig, 1988 The heat-shock proteins. Annu. Rev. Genet. **22:** 631–677.

Meagher, R. B., S. Berry-Lowe and K. Rice, 1989 Molecular evolution of the small subunit of ribulose bisphosphate carboxylase: nucleotide substitution and gene conversion. Genetics **123:** 845–863.

Merck, K. B., P. J. T. A. Groenen, C. E. M. Voorter, W. A. D. Haard-Hoekman, J. Horowitz *et al.,* 1993 Structural and functional similarities of bovine alpha-crystallin and mouse small heat-shock protein. A family of chaperones. J. Cell Biol. **114:** 255–261.

Nagalaki, T., 1984 Evolution of multigene families under intrachromosomal gene conversion. Proc. Natl. Acad. Sci. USA **81:** 3796–3800.

Nover, L., 1991 *Heat Shock Response.* CRC Press, Boca Raton, FL.

Ohta, T., 1988a Evolution by gene duplication and compensatory advantageous mutations. Genetics **120:** 841–847.

Ohta, T., 1988b Multigene and supergene families, pp. 41–65 in *Oxford Survey in Evolutionary Biology,* edited by P. H. Harvey and L. Partridge. Oxford University Press, New York.

Ohta, T., 1988c Time for acquiring a new gene by duplication. Proc. Natl. Acad. Sci. USA **85:** 3509–3512.

Ohta, T., 1991 Multigene families and the evolution of complexity. J. Mol. Evol. **33:** 34–41.

Plesofsky-Vig, N., J. Vig and R. Brambl, 1992 Phylogeny of the alpha-crystallin-related heat shock proteins. J. Mol. Evol. **35:** 537–545.

Rensing, S. A., and U.-G. Maier, 1994 Phylogenetic analysis of the stress-70 protein family. J. Mol. Evol. **39:** 80–86.

Sanderson, M. J., and J. J. Doyle, 1992 Reconstruction of organismal and gene phylogenies from data on multigene families: concerted evolution, homoplasy, and confidence. Syst. Biol. **41:** 4–17.

Schneider, H.-C., J. Berthold, M. F. Bauer, K. Dietmeier, B. Guiard *et al.,* 1994 Mitochondrial Hsp70/MIM44 complex facilitates protein import. Nature **371:** 768–774.

Swofford, D. L., 1993 *PAUP: Phylogenetic Analysis Using Parsimony, Version 3.1.* Computer program distributed by the Illinois History Survey, Champaign, IL.

Vierling, E., 1991 The heat shock response in plants. Annu. Rev. Plant Physiol. Plant Mol. Biol. **42:** 579–620.

Walsh, J. B., 1987 Sequence-dependent gene conversion: can duplicated genes diverge fast enough to escape conversion. Genetics **117:** 543–557.

Walsh, J. B., 1995 How often do duplicated genes evolve new functions. Genetics **139:** 421–428.

Wu, C.-I., and W.-H. Li, 1985 Evidence for higher rates of nucleotide substitution in rodents than in man. Proc. Natl. Acad. Sci. USA **82:** 1741–1745.

Communicating editor: W.-H. Li