Plant protein-coding gene families: emerging bioinformatics approaches

Manuel Martinez

Centro de Biotecnología y Genómica de Plantas (UPM-INIA), Campus Montegancedo, Universidad Politécnica de Madrid. Autovía M40 (Km 38), 28223-Pozuelo de Alarcón, Madrid, Spain

Protein-coding gene families are sets of similar genes with a shared evolutionary origin and, generally, with similar biological functions. In plants, the size and role of gene families has been only partially addressed. However, suitable bioinformatics tools are being developed to cluster the enormous number of sequences currently available in databases. Specifically, comparative genomic databases promise to become powerful tools for gene family annotation in plant clades. In this review, I evaluate the data retrieved from various gene family databases, the ease with which they can be extracted and how useful the extracted information is.

Classification of plant protein-coding genes into families

Classification of protein-coding genes into families is based on the structure, function and evolution of the proteins they encode and is widely accepted as a crucial tool for functional genomics. Gene families can be defined either as sets of evolutionarily related genes shared by several different species and with often similar biological functions, or as a set of homologous genes within one species (species-specific gene family). Although some gene families appear to be more dynamic during evolution and have species-specific gene members, others are more conserved and orthologous genes (i.e. genes sharing common ancestry that have diverged by speciation) can be found in evolutionarily distant but related species or spread across different kingdoms of life (see examples in [1]). Orthologous genes are particularly useful for the characterisation of unannotated proteins by identifying annotated counterparts that share high sequence identity. Sequence similarity searches by BLAST [2] or FASTA [3] programs have been used for the automatic classification of sequences, but a major limitation of this approach is that both tools treat each position in the query sequence with equal importance, constraining their ability to detect divergent homologues. Traditional signature databases, such as Pfam [4] or PROSITE [5], use alignments of multiple sequences to detect specific residues or motifs conserved among a set of homologous proteins. These motifs (or signatures) have been shown to be important for protein functionality and are able to define a family of proteins. Signature databases provide many clusters of prokaryotic and eukaryotic genes, which are useful for annotating proteins based on amino acid sequence similarities.

Recently, the increasing number of sequenced genomes has led to the development of novel bioinformatics tools for the analysis of gene families based on comparative genomics. These tools are based on methods that involve new clustering methodologies that will be particularly useful for identifying protein-coding gene families in plants. So far, most analyses on plant gene families have been performed by laborious searches using genomic and transcriptomic databases (for some recent examples, see [6–9]). Now, in addition to traditional signature-based repositories, emerging comparative genomic databases specific for plants promise to be powerful tools for discovering orthologous genes and provide the basis for more accurate gene family annotation in plant clades.

Bioinformatics tools and WWW-based databases

Different bioinformatics databases are available online and can be used to perform sequence-based analyses of gene families. Over the past few years, some of these databases have rapidly become obsolete or have not been properly updated, whereas new databases are continuously being created. The main databases currently available for gene family classification are detailed in Table 1. Traditional gene family databases based on signatures have been extensively used to classify proteins. Sequence signatures are typically derived from multiple sequence alignments that have been manually curated. These databases use different methodologies to produce protein signatures (Box 1), such as sequence clustering, regular expressions, profiles, or hidden Markov models (HMMs), and a variable degree of biological information on well-characterised proteins. Furthermore, they differ in the information they use to characterise the clusters (i.e. functional sites, functional conserved motifs, functional domains and structural domains) and the primary sequence storage database that provides the sequences. Most databases use the UniProtKB sequence database [10], which is a curated database with two sections. (i) UniProtKB/Swiss-Prot contains manually annotated records with information extracted from literature and curator-evaluated computational analysis; and (ii) UniProtKB/TrEMBL contains high-quality computationally analysed records enriched with automatic annotation and classification, and includes translations of all the coding sequences present in the ENA/GenBank/DDBJ Nucleotide Sequence Databases [11–13]. Thus, it contains redundant

Table 1. Bioinformatics tools for plant gene family analyses

Bioinformatic tool/URL	Database source	Clustering method	Cluster information based on	Protein families or signatures
Signature databases				
ProtClustDB Dec 2 2010/ http://www.ncbi.nlm.nih.gov/proteinclusters	NCBI RefSeq	Clique based	Functional domains	627757, 10885 (curated)
Pfam 25.0/ http://pfam.sanger.ac.uk/	UniProtKB	HMMs	Functional domains	12273 (Pfam-A)
PROSITE 20.68/ http://expasy.org/prosite/	UniProtKB	Patterns, profiles	Functional sites	1598
PRINTS 41.1/ http://www.bioinf.manchester.ac.uk/dbbrowser/PRINTS/index.php	UniProtK	Fingerprints	Functional conserved motifs	2050
ProDom 2006.1/CG267/ http://prodom.prabi.fr/prodom/current/html/home.php	UniProtKB/267 completed genomes (one from plants)	MKDOM2	Functional domains	574656/301126
SMART 6.1/ http://smart.embl-heidelberg.de/	UniProtKB/760 completed genomes (one from plants)	HMMs	Functional domains	895
TIGRFAMs 10.0/ http://www.jcvi.org/cms/research/projects/tigrfams/overview/	UniProtKB	HMMs	Functional domains	4025
PIRSF 2.74/ http://pir.georgetown.edu/pirwww/dbinfo/pirsf.shtml	UniProtKB	HMMs	Functional domains	3248 (curated)
SUPERFAMILY 1.75/ http://supfam.cs.bris.ac.uk/SUPERFAMILY/	1452 completed genomes (27 from plants)/UniProtKB/PDB	HMMs	SCOP domains	2019
GENE3D 10.0.0/ http://gene3d.biochem.ucl.ac.uk/Gene3D/	1867 completed genomes	HMMs	CATH domains	2549
PANTHER 7.0/ http://www.pantherdb.org/	48 completed genomes (three from plants)	HMMs	Functional domains	6594
Integrative signature databases				
InterPro 31.0/ http://www.ebi.ac.uk/interpro/	UniProtKB	Signature integration	Gene3D, HAMAP, PANTHER, Pfam, PIRSF, PRINTS, ProDom, PROSITE, SMART, SUPERFAMILY, TIGRFAMs signatures	21185
CDD 2.26/ http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml	NCBI Database	PSSMs	NCBI-curated domains, Pfam, SMART, COGs, ProtClustDB signatures	41593
Comparative genomic databases				
PLAZA 2.0/ http://bioinformatics.psb.ugent.be/plaza/	23 completed genomes	TribeMCL, OrthoMCL,	InterPro signatures, GO annotations	32332
Phytozome 7.0/ http://www.phytozome.net/	25 completed genomes	DASW-based	Pfam, KOG, KEGG, PANTHER signatures	249439
GreenPhylDB 2.0/ http://greenphyl.cirad.fr/v2/cgi-bin/index.cgi	16 completed genomes	TribeMCL	InterPro signatures, UniProtKB entries, KEGG pathway entries	8227 (level 1)
EnsemblPlants 8.0/ http://plants.ensembl.org/index.html	15 completed genomes (five non-plant)	Hcluster_sg based	InterPro signatures, GO annotations, UniProtKB entries, UniGene entries	35183

data. Some databases use information from completed genome projects. Among these, ProtClustDB [14] uses the RefSeq protein collection source from NCBI, which contains non-redundant sets of curated protein sequences from eukaryotic and prokaryotic genomes [15]. ProDom [16] and SMART [17] combine information from completed genomes and the UniProtKB database. PANTHER [18] signatures are based on sequences from 48 completed genomes. SUPERFAMILY [19] and Gene3D [20] also use completed genomes as a database source, but they classify the clusters based on the three-dimensional domains from the SCOP [21] and CATH [22] databases.

In addition to single databases, integrative databases provide a powerful resource to use to classify proteins on multiple levels: from protein families to structural superfamilies and functionally close subfamilies. The most widely used integrative databases are InterPro [23] and CDD [24]. InterPro integrates signatures from 11 major signature databases (Gene3D, HAMAP [25], PANTHER, Pfam, PIRSF [26], PRINTS [27], ProDom, PROSITE, SMART,

Box 1. Clustering methods

Protein classification into families implies the creation of methods for clustering homologous sequences. These methods can be designed to create protein signatures from multiple sequence alignments and grouped proteins according to sequence similarities, or to cluster homologous sequences based on pairwise comparison of full-length protein sequences from BLAST searches. The databases compiled in this review use the following clustering methods:

Methods based on multiple sequence alignments

Regular expressions: computer-readable formula for a pattern, which is a short conserved motif of amino acids found within a protein sequence.

Profiles: matrix of position-specific amino acid weights in which each position provides a score of the likelihood of finding a particular amino acid at a specific position in the sequence. Fingerprints are sets of conserved motifs that are modelled using profiles.

Hidden Markov Models: statistical models based on Bayesian methods, which use probabilities rather than scores found in profiles.

Methods based on pairwise comparisons from BLAST searches

Clique-based: based on BLAST scores modified by protein length × alignment length. Clusters (or cliques) consist of protein sets in which, for any given protein in the cluster, all the other members of the cluster have a greater modified score to this protein than does any protein outside of the cluster.

MKDOM2: based on recursive PSI-BLAST searches. It relies on the assumption that the shortest amino acid sequence corresponds with a single domain, and can be used as a query to screen the database with the PSI-BLAST program, to cluster homologous domains.

TribeMCL: based on the Markov Cluster (MCL) algorithm for graph clustering by flow simulation. The method does not operate directly on sequences but on a graph that contains similarity information obtained from BLAST searches. Global patterns of sequence similarity are detected and used to partition the similarity graph into protein families.

OrthoMCL: based on the same method as TribeMCL; it generates clusters of proteins where each cluster consists of orthologs or 'recent' paralogs.

DASW-based: based on an all-versus-all dual affine Smith-Watermann alignment (DASW). Consists of the accretion of paralogs to mutual best-hit ortholog seed using outgroup scores as thresholds and a clustering metric based on sequence similarity using DASW alignment.

Hcluster_sg-based: based on a sparse graph of protein relations constructed from the scores obtained from a WUBlastp1Smith-Waterman pairwise comparison of each gene against every other gene. Hcluster_sg performs hierarchical clustering under mean distance. It reads an input file that describes the similarity between two sequences, and groups the two nearest nodes at each step.

SUPERFAMILY and TIGRFAMs [28]); whereas CDD imports signatures from Pfam, SMART, COGs [29] and ProtClustDB, and uses position-specific scoring matrices (PSSMs) to derive database search models. All aforementioned signature databases use sequences obtained from prokaryotic and eukaryotic species, but without a focus on plants.

Recent advances in genomics studies have increased current knowledge of plant genomes. According to the Genomes OnLine Database (GOLD, http://www.genomesonline.org/) [30], more than 20 plant genomes have already been completed and there are more than 200 ongoing plant genomic projects, most of which are benefitting from the use of next-generation sequencing technologies. This continuously increasing number of available plant genomes and new sequencing genome projects has led to the creation of specific comparative genomic databases for plants. These databases can be used to perform comparative analyses, and to study the genome organisation, evolutionary origin and composition of gene families. Based on orthologous genes, comparative genomics provides a powerful approach for applying biologically functional information gained from model species to crops. Furthermore, phylogenetic analysis can provide useful information about the processes that have contributed to the evolutionary divergence of genes [31,32]. The most comprehensive comparative genomic databases that focus on plant gene families are PLAZA [33], Phytozome, GreenPhylDB [34] and EnsemblPlants [35] (Table 1), although they differ in the method used to perform the cluster analysis (Box 1). Clustering methods usually involve pairwise comparisons of full-length protein sequences. PLAZA uses graph-based clustering methods implemented in TribeMCL [36] and OrthoMCL [37]; GreenPhylDB also uses TribeMCL. Phytozome clusters are based in the accretion of paralogs using outgroup scores, whereas EnsemblPlants uses hclusters_sg to generate clusters from a sparse graph of protein relations based on BLAST scores. In addition, PLAZA, GreenPhylDB and EnsemblPlants use phylogenetic inferences to identify biologically relevant duplication and speciation events. The three programs construct the phylogenetic trees by the maximum likelihood method PhyML [38], and use different methods to reconcile the trees (NOTUNG in PLAZA,SDI/RIO in GreenPhylDB, and TreeBeST in EnsemblPlants). In addition, these databases differ in the genomes they include, the external databases they use for functional annotation of the clusters, and the number of protein families they contain. PLAZA also provides multispecies colinearity views. This is an emerging trend for studying gene families, because more sequences from closely related genomes are becoming available.

Protein-coding family analyses: peptidase families and their inhibitors

Over the past few years, many studies have focused on characterising the extent of a gene family in a plant species. Recent examples include protein families with different molecular functions, such as transcription factors [39,40], catalytic enzymes [7,9], transporters [8] or molecular transducers [41]. Given the many papers on different plant gene families, a comprehensive coverage of all the different reports is not feasible. Here, gene families of plant peptidases and their inhibitors have been selected based on: (i) the number of recent reports with information on the size of peptidase/inhibitor gene families; and (ii) the existence of a specific database for both peptidases and inhibitors (i.e. the MEROPS database, http://merops.sanger.ac.uk/) [42]. Peptidases are enzymes that hydrolyse peptide bonds and are encoded by approximately 2% of genes in all living organisms. Peptidase activity is tightly controlled by protein–protein interactions with their inhibitors, the expression and activities of which are also under strict regulation. Peptidases were formerly classified, based on the chemical mechanism of catalysis, as serine,

Table 2. Entire peptidase and peptidase inhibitor gene families characterised in plants

Clan/Family	Family description		ethodology	Refs
Peptidases				
AA/A1	Pepsin-like	1	HMMER's Genome + non-redundant	[46]
	Pepsin-like	1	BLASTp Genome	[65]
	Pepsin-like	1	BLASTp Genome	[66]
	Pepsin-like	1	BLASTp Genome + tBLASTn cDNAs + Motif Scan	[44]
CA/C1A	Papain-like	1	HMMER's Genome + nr	[46]
	Papain-like	2	BLASTp Genome + Motif Scan	[67]
	Papain-like	10	BLASTp Genome + tBLASTn cDNAs/ESTs	[51]
CD/C11	Clostripain-like	16	BLASTp Genome + tBLASTn cDNAs/ESTs	[1]
CD/C13	Legumain-like	10	BLASTp Genome + tBLASTn cDNAs/ESTs	[51]
	Legumain-like	16	BLASTp Genome + tBLASTn cDNAs/ESTs	[1]
	GPI:protein transamidase-like	16	BLASTp Genome + tBLASTn cDNAs/ESTs	[1]
CD/C14	Metacaspase-like	1	BLASTp Genome	[68]
	Metacaspase-like	16	BLASTp Genome + tBLASTn cDNAs/ESTs	[1]
CD/C50	Separase-like	16	BLASTp Genome + tBLASTn cDNAs/ESTs	[1]
PA/S1	Chymotrypsin-like	2	BLASTp Genome + Motif Scan	[67]
SB/S8	Subtilisin-like	1	HMMER's Genome + non-redundant	[46]
	Subtilisin-like	1	BLASTp Genome	[69]
SC/S10	Serine carboxypeptidase-like	1	PSI-BLAST Genome + tBLASTn cDNAs	[48]
	Serine carboxypeptidase-like	1	HMMER's Genome + cDNAs	[45]
	Serine carboxypeptidase-like	2	BLASTp Genome (Phytozome)	[43]
SK/S14	Clp-like	2	BLASTp Genome + Motif Scan	[67]
SJ/S16	Lon-like	2	BLASTp Genome + Motif Scan	[67]
ST/S54	Rhomboid-like	2	BLASTp Genome + Motif Scan	[67]
MA/M41	FtsH-like	1	BLASTp Genome	[70]
	FtsH-like	2	BLASTp Genome + Motif Scan	[67]
Inhibitors				
IC/I3	Kunitz trypsin inhibitor	1	BLASTp Genome + tBLASTn ESTs	[49]
ID/I4	Serpins	3	BLASTp/PSI-BLAST Genome + tBLASTn ESTs	[71]
IG/I13	Chymotrypsin-like inhibitor	1	BLASTp Genome + tBLASTn cDNAs	[72]
IH/I25	Cystatins	3	BLASTp Genome + tBLASTn ESTs	[50]
	Cystatins	10	BLASTp Genome + tBLASTn cDNAs/ESTs	[51]

cysteine, threonine, aspartic, glutamic, asparagine or metallo peptidases. Currently, the MEROPS database is based on a hierarchical classification of homologous sets of peptidases and peptidase inhibitors grouped into families by sequence comparison. The families are sorted into clans by three-dimensional structural similarity.

A summary of recent analyses of plant peptidase/inhibitor gene families is detailed in Table 2. This shows plant peptidase and peptidase inhibitor protein families classified based on their putative peptidase/inhibitor catalytic role and the species in which they have been characterised. From 2004 to date, complete families of aspartic (A), cysteine (C), serine (S) and metallopeptidases (M) have been annotated in various plant species, but mainly in the model plant *Arabidopsis thaliana* (*Arabidopsis*). Early studies relied on a single plant species. With the increase of completely sequenced genomes and to unravel evolutionary changes in a specific gene family, analysis have incorporated species from multiple plant clades, ranging from algae to basal and land plants.

Extensive searches in recent reports indicate that the methodology used to interpret gene composition of gene families of peptidases and their inhibitors was similar to those used to investigate gene families with other molecular functions. It is remarkable that none of these reports made use of searches with WWW-based gene family databases. For peptidases, only one report used the comparative genomic database Phytozome [43]. However, the report did not extract sequences from a keyword search for gene families, but instead used BLAST searches in the genomic databases included in Phytozome. The method most frequently used to find peptidases or peptidase inhibitors from a determined family is the automatic search function for sequences by BLAST on genomic sequences, proteomic sequences derived from them, or in collections of cDNA or transcript assemblies. These BLAST searches have usually been performed using one or several known amino acid sequences for proteins that belong to the selected gene family. Alternatively, HMMs have been constructed and used for the search [44,45]. Most analyses include additional manual curation for identified proteins to refine selection of genes belonging to the different families. This curation normally uses a multiple alignment to identify and eliminate repeated sequences as a first step. Additionally, family members can be confirmed by searching for gene family motifs in signature databases [44,46]. Often, protein models derived from the automatic prediction of proteins from protein-coding genes have to be manually curated. The status of a complete genomic project ranges from the standard draft to the finished annota-

Table 3. Examples of WWW-based analyses of plant gene families in *Arabidopsis*[a]

Bioinformatic tool	Legumains	Metacaspases	Cystatins	Zf-DOF	GRAS	ARF
Genomes	5 genes, 8 gene models	9 genes, 11 gene models	7 genes, 8 gene models	36 genes, 45 gene models	33 genes, 36 gene models	23 genes, 35 gene models
Dedicated databases						
MEROPS/ PlnTFDB	Peptidase C13 legumain family (5)	Peptidase C14 caspase family (9)	Inhibitor I25 cystatin family (7)	C2C2-Dof family (36, 45 gene models)	GRAS family (33, 36 gene models)	ARF family (23, 35 gene models)
Signature databases						
ProtClustDB	Clusters: 2 (5, 411)	Clusters: 8 (10)	Clusters: 7 (7)	Clusters: 43 (48)	Clusters: 32 (34)	Clusters: 26 (36)
Pfam	PF01650 Peptidase C_13 (12)	PF00656 Peptidase_ C14 (11)	PF00031 Cystatin (11)	PF02701 Zf-Dof (62)	PF03514 GRAS (44)	PF02362 B3 (101) PF02309 Aux/IAA (156) PF06507 ARF (39)
PROSITE	No hit	No hit	PS00287 Proteinase inhibitor I25, cystatin, (7)	PS50884 Zf-Dof-2 (62) PS01361 Zf-Dof-1 (62)	PS50985 GRAS (43)	PS50863 B3 (119) PS50962 Aux/IAA-ARF (387)
PRINTS	PR00776 Hemoglobinase	No hit	PR00295 Stefina	No hit	No hit	No hit
ProDom	Clusters: 3 (6, 5, 5)	Clusters: 4 (9, 7, 4, 3)	Clusters: 2 (7, 1)	Clusters: 3 (31, 2, 1)	Clusters: 8 (36, 35, 34, 17, 3, 3, 3, 2)	Clusters: 11 (47, 35, 26, 26, 25, 22, 20, 2, 2, 2, 1)
SMART	No hit	No hit	SM00043 (7)	No hit	No hit	No hit
TIGRFAMs	No hit	No hit	No hit	No hit	No hit	No hit
PIRSF	No hit	No hit	No hit	No hit	No hit	No hit
SUPERFAMILY	No hit	SCOP52129 Caspase-like (6)	SCOP54403 Cystatin/ monellin (23)	No hit	No hit	SCOP101936 DNA-binding pseudobarrel (224) SCOP54277 CAD and PB1(118)
GENE3D	No hit	CATH 3.40.50.1460 (10)	CATH 3.10.450.10 (18)	No hit	No hit	CATH 2.40.330.10 (123)
PANTHER	PTHR12000 (5)	No hit	PTHR11413 (7)	No hit	No hit	No hit
Integrated signature databases						
InterPro	IPR001096 (12)	IPR011600 (11)	IPR000010 (11) IPR018073 (5) IPR020381 (16)	IPR003851 (62)	IPR005202 (44)	IPR003311 (380) IPR003340 (114) IPR010525 (78) IPR011525 (387) IPR015300 (132)
CDD	cl02159	cl00042	cl09238	cl03664	cl03514	cl05824, cl03557, cl03528
Comparative genomic databases						
PLAZA	HOM000842 (4)	HOM001286 (6) HOM000940 (3)	HOM000450 (7)	HOM000095 (36)	HOM000041 (36)	HOM000105 (24)
Phytozome	29032379 (4) 29026375 (1)	29013081 (3) 29001759 (6)	Clusters: 7 (2, 2, 1, 1, 1, 1, 1)	Clusters: 6 (14, 9, 5, 2, 1, 1)	Clusters: 10 (11, 6, 4, 4, 2, 2, 1, 1, 1, 1)	Clusters: 11 (19, 12, 7, 6, 5, 4, 4, 3, 2, 1, 1)
GreenPhylDB	21756 Peptidase C13 family (7)	21380 Caspase family (11)	21238 Cysteine Protease Inhibitor Family / Cystatin (8)	20969 C2C2-DOF family (45)	20939 GRAS family (36)	25036 ARF family (35)
EnsemblPlants	Plant (4) Pan taxonomic (5)	Plant (6) Pan-taxonomic (9)	Plant (4) Pan-taxonomic (6)	Plant (21) Pan-taxonomic (36)	Plant (7) Pan-taxonomic (30)	Plant (3) Pan-taxonomic (24)

[a]The number of *Arabidopsis* sequences is given in parenthesis.

tion [47]. Standard draft sequences are likely to harbour many poor-quality regions and can be relatively incomplete. Thus, some genes can be missed or misarranged. In these cases, the inclusion in the analysis of sequences from transcript databases is recommended to ensure that the complete gene family is recorded [48]. When various databases have been scanned, a second sequence alignment should be performed to avoid potential protein redundancy and select unique protein sequences [44,45,49]. Alternatively, when plant genomic sequences are not yet available, advantage might be taken of large amounts of EST-derived sequences, as is the case for barley (*Hordeum vulgare*). Searches in those transcript assembly databases have been crucial in determining the approximate number of members that form several peptidase/inhibitor families in barley [50,51].

WWW resources: *Arabidopsis* protein-coding gene families

Difficulties in establishing the entire plant protein-coding gene families reported in the abovementioned studies highlight the importance of having confidence in the data obtained from protein-coding gene family databases. Each database has different aims, underlying methods and models, and they will not return exactly the same set of proteins. As an example, the retrieved results against different databases have been evaluated for several gene families encoding peptidases/inhibitors and transcription factors (Table 3). These families have been selected based on the following three criteria: (i) all are well-known families whose occurrence in the fully sequenced *Arabidopsis* genome has been previously reported; (ii) dedicated databases for these families exist, allowing a more complete comparison between these databases and signature and comparative genomic databases; and (iii) their widely conserved amino acid regions make them good candidates to be correctly clustered in gene family databases.

As previously described, the MEROPS database compiles all information on peptidases and their inhibitors. Two families of peptidases, the C13 legumain family and the C14 metacaspase family, and a family of peptidase inhibitors, the I25 cystatin family, were selected. The C13 family is formed by legumains or vacuolar processing enzymes (VPE) and glycosylphosphatidylinositol (GPI):-protein transamidases [1]. More than 300 C13 peptidase sequences are available in the MEROPS database that share an active site formed by a His and a Cys and the conserved amino acids that surround them. C14 is a broad family with more than 500 sequences in the MEROPS database and includes caspases, paracaspases and metacaspases I and II. In plants, only metacaspases have been described [1]. Both types of metacaspase share the His and Cys residues and most of the amino acids surrounding them but differ in their global domain architecture. The I25 family is formed by cystatins, which are proteins able to inhibit cysteine peptidases from the papain subfamily C1A and the C13 legumain family [52]. Cystatins are also a well-known family in eukaryotes and viruses, accounting for more than 300 sequences in the MEROPS database. The three-dimensional structures of these proteins have been resolved, showing their high level of conservation (for examples, see [53–55]). In recent genomic- and evolutionary-wide analyses, the composition of these three gene families in *Arabidopsis* has been described. The C13 family includes four legumains (five gene models) and one GPI:-protein transamidase (three gene models); the C14 family is formed by three metacaspases type I and six type II (11 gene models); and seven different members have been found for the I25 family (eight gene models) [1,50,51].

For transcription factors, there are many specific databases, such as PlnTFDB (http://plntfdb.bio.uni-potsdam.de/v3.0/) [56]. Three different families of transcription factors specific for plants, the zinc finger, DNA-binding with one finger (Zf-DOF), GRAS (GAI, RGA, SCR) and auxin response factor (ARF) families have been selected here. The Zf-Dof family is a particular class of zinc finger protein characterised by a conserved region of 50 amino acids with a C2–C2 finger structure, associated with a basic region [57]. The GRAS gene family is an important plant-specific gene family that features a variable amino-terminus and a highly conserved carboxyl-terminus that contains five recognisable motifs, including two leucine heptad repeats [58]. Finally, the ARF transcription factors form a more complex family. ARFs consist of modular domains that can function independently of one another [59]. Most ARFs contain an amino-terminal DNA binding domain that is classified as a plant-specific B3-type, which is also found in a variety of plant transcription factors, such as the LEC2-ABI3–VAL (LAV), Related to ABI3 and VP1 (RAV) and Reproductive meristem (REM) families [60]. Furthermore, ARFs contain a middle region that functions as an activation or repression domain, and a carboxy-terminal dimerisation domain that is related in terms of its amino acid sequence to domains III and IV in Aux/IAA proteins [59]. In contrast to peptidases and their inhibitor gene families, extensive duplications during evolution have led to the large number of transcription factor members in gene families. 36 Zf-DOF genes (45 gene models), 33 GRAS genes (36 gene models) and 23 ARF genes (35 gene models) have been annotated in the *Arabidopsis* genome [58,61,62].

Examples of *Arabidopsis* gene families obtained from database searches are compiled in Table 3. BLAST scans were performed using as a query a representative *Arabidopsis* sequence from each analysed gene family (legumains, *At*2g25940; metacaspases, *At*1g79330; cystatins, *At*2g40880; Zf-DOFs, *At*3g45610; GRAS, *At*1g07520; and ARFs, *At*1g19220). The dedicated databases MEROPS and PlnTFDB were used as controls. Although MEROPS does not include different gene models for every protein, it correctly groups the *Arabidopsis* proteins into their families. PlnTFDB gives the correct number of family members, including the putative splicing variants for each transcription factor.

Signature databases based on clustering techniques gave hits for all the sequences analysed. ProtClust gave several sequences for each family that fit approximately with the number of gene models in the *Arabidopsis* genome. However, these sequences were grouped in several non-curated clusters for each family, with the exception of C13 legumains, which are included in two different clusters, one of them formed by the four actual C_13 legumains, and a second one encompassing a unique

GPI:protein transamidase. ProDom clusters are based on domains. Therefore, each protein could be assigned to several clusters, and the most numerous can be chosen as the most reliable result for a cluster. The results for metacaspases, cystatins and GRAS families corresponded with the correct number of members found in *Arabidopsis*. However, the results for the legumains, Zf-DOF and ARF families did not correspond with the known numbers of members of the most numerous clusters. With the exception of the ARF family, the Pfam database scored one single entry for each sequence analysed. Pfam classifies domains, rather than global proteins, which explains the inclusion of ARFs in three different families: the proteins with a B3 domain, the proper ARFs and the proteins with an Aux/IAA dimerisation domain. When the number of sequences from *Arabidopsis* was requested from the Pfam database, higher numbers of sequences than expected were retrieved, confirming the existence of duplicated sequences in the source databases. The PROSITE database assigned the cystatin, Zf-DOF, GRAS and ARF sequences to protein families found in *Arabidopsis* with similar problems to those observed in the Pfam database, namely duplication of sequences and assignment of ARF sequence to different families. Databases based on structural information, such as GENE3D and SUPERFAMILY, found homology only for metacaspase, cystatin and ARF sequences. With the exception of metacaspases, these databases predicted very high numbers of cystatins and ARFs for *Arabidopsis*, which suggests that both proteins share a common fold with several related proteins. The remaining signature databases gave poor results for gene family analysis. The TIGRFAMs and PIRSF databases did not find hits in any sequence; SMART only found a hit in the cystatin sequence; and PRINTS found hits on the cystatin and legumain sequences, but without any information on the occurrence of members of this family in individual species. Likewise, PANTHER, which uses genomic sequences, found hits only on cystatin and legumain sequences, but gave the correct number of family members in *Arabidopsis*.

InterPro and CDD integrative signature databases performed much better. They found a single entry for most protein families based on the results from individual database searches. However, ARFs were found in five different InterPro and three different CDD families, and cystatins also appeared in three different InterPro entries. Because the annotation of InterPro families is based on results from signature databases such as Pfam, higher numbers of sequences than expected were present in the InterPro families.

By contrast, searches for *Arabidopsis* sequences in the comparative genomic databases produced the most reliable results. However, the different methods involved in the creation of clusters in these databases implied variations in the retrieved results. The best results were obtained from the GreenPhylDB, which is the only database that considers different splicing forms for *Arabidopsis* gene models; with the exception of legumains, for which one gene model was not retrieved, all members of each gene family tested were correctly clustered and named in this database. All clusters were at level one, the top level in the hierarchy, except for the ARF family, which was clustered at level two, but was also included at level one in a more extended family that contained other proteins with B3 or Aux/IAA domains. The PLAZA database also gave good results. It was able to sort metacaspases I and II, and legumains and GPI:protein transamidases into different clusters. For transcription factors, this database found the correct number of Zf-DOF members in *Arabidopsis*, but included one mitochondrial and three ribosomal proteins in the GRAS and ARF clusters, respectively. Phytozome clusters gave less useful information on *Arabidopsis* family members in the gene family database. The most accurate results were obtained for C13 and C14 peptidases, which were clustered in two groups resembling the subfamilies known for these two gene families. The remaining families were split into several clusters, showing proteins from other gene families in clusters with homology to the ARF sequence. Finally, gene family information in EnsemblPlants was more difficult to find and gave inconsistent results. When the Plant Compara database was used, clusters included a very low number of proteins for each family analysed. By contrast, the Pan-taxonomic database, which has a wider taxonomic scope, including genomes from prokaryotes, gave accurate numbers of proteins for *Arabidopsis* gene families.

In summary, although valuable information can be extracted from most signature and integrative signature databases, the best performance can be obtained from searches in comparative genomic databases, particularly GreenPhylDB and PLAZA.

Obtaining useful information from WWW databases

Based on the bioinformatics tools described in this review, an overview of how to obtain information or identify protein-coding gene families from databases is given in Figure 1. This flowchart includes different input and output points covering results from laboratory-obtained sequences or literature information to final sequences extracted from established and new databases. Information about sequences generated by laboratory sequencing or obtained from bibliographic databases can be obtained rapidly by searching the ENA/GenBank/DDBJ primary databanks from the European Bioinformatics Institute (EBI), the Nacional Center for Biotechnology Information and the Center for Information Biology-DNA Data Bank of Japan (CIB-DDBJ), respectively. The three databanks are in a collaborative network where sequences are shared. This implies that the same hits are achieved with a search in any of these three databases [63].

After primary information has been collected from databanks, a next step should be to perform searches in secondary or specialised databases. Most of these databases offer searching options by using keywords or BLAST analysis of sequences. Commonly, a BLAST search will retrieve more accurate information than will a keyword search, because keyword searches access the entire entry of a sequence, signature or family, which can lead to the identification of incorrect gene family sequences or to vital entries being missed. Depending on their availability and the expected result of the search, BLAST searches with a sequence of interest can be performed in five different types of secondary database: (i) genomic databases; (ii)

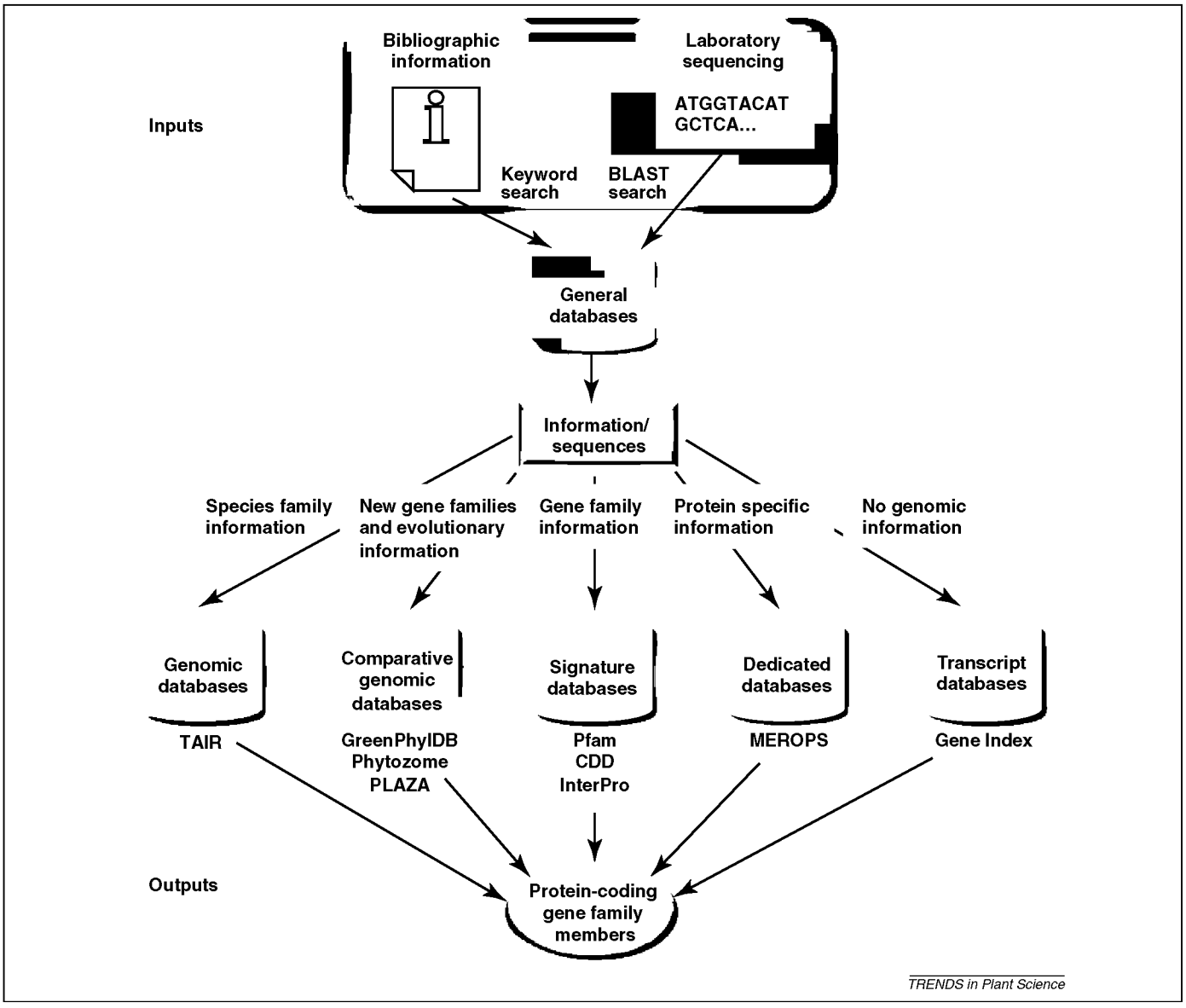

Figure 1. Flowchart of gene family analyses. Inputs are bibliographic information on a selected gene family or nucleotidic/amino acid sequences obtained in the laboratory. Keyword or BLAST searches on general databases give more information and sequences on the gene family. Depending on their availability and the expected result of the search, BLAST searches can be performed in five different types of secondary database: genomic, comparative genomic, signature, dedicated, or transcript database. Examples of databases for each type are shown. Results on the number of members for a selected protein-coding gene family can be different for each database.

comparative genomic databases; (iii) signature databases; (iv) dedicated databases; and (v) transcript databases.

Genomic databases offer the best accuracy on the genomic sequence of a species, because they are regularly updated. Some examples are The Arabidopsis Initiative Resource (TAIR; http://www.arabidopsis.org/) or The Rice Genome Annotation Project (http://rice.plantbiology.msu.edu/). Most genomic databases support their functional annotation with an active link to signature databases. If the sequence in question belongs to a species whose genome has been sequenced and annotated, the best information about the members of the family it belongs to will be retrieved from searches in its own genomic database.

Comparative genomic databases such as GreenPhylDB, Phytozome or PLAZA, are based on methods that use pairwise comparisons of full-length protein sequences and typically involve clustering techniques. Clustering methods can be applied to classify many sequences rapidly, in an automated manner, and with reasonable accuracy, particularly for the Tribe-MCL algorithm implemented in PLAZA and GreenPhylDB [64]. This has been demonstrated with the above examples and enables genes to be grouped into families not covered by signature methods. The plant comparative genomic databases are the best choice for the identification of members of a protein family in related species, which is particularly interesting for phylogenetic analyses and the prediction of gene function in evolutionary biology.

Signature databases, such as Pfam and Prosite, are an alternative way of obtaining information about the members of a gene family. Searches in signature databases allow genes to be grouped according to similarities with known sequence signatures. Signature-based methods are routinely used for gene function annotation and most give broad information about proteins in every family, which is exceptionally good in the Pfam database. However, these methods

have different limitations, such as the incorrect resolution of gene family substructures, missing gene families with yet uncharacterised motifs or domains, or insufficient updating. As an example of these limitations, most signature databases failed to provide correct results for the protein families above selected. Integrative signature databases, such as InterPro and CDD, aim to overcome limitations of individual signature databases. Improvements have been made mostly in the hierarchical organisation of gene families. Links to their integrated databases makes them a useful first option for signature database searches.

Dedicated databases are focused on a specific group of related protein families. These databases commonly give specialised information on the included proteins that cannot be retrieved from signature databases. The MEROPS database for peptidases and their inhibitors and the PlnTFDB for transcription factors are two examples of specialised databases that are regularly updated and offer valuable information for each particular case.

Finally, transcript databases are based on ESTs or cDNA sequencing projects. Some examples of databases focused on these sequences are the Transcript Assemblies Database (http://plantta.jcvi.org/index.shtml) and the Gene Index Project (http://compbio.dfci.harvard.edu/tgi/). For several plant species without genome sequencing projects, databases with extensive collections of EST or cDNA sequences are available. In these cases, searches of these databases combined with manual curation of retrieved sequences can provide important information about members of a gene family in a not yet sequenced plant species. In addition, many researches do not use gene-by-gene searches, but are shifting towards data analysis when information for a set of genes is required. Several databases offer tools for bulk analysis, such as the integrated signature database CDD and the comparative genomic database PLAZA.

Concluding remarks

The rapid increase in available genome sequences has produced an enormous volume of raw information that needs to be processed to extract information about gene family architecture and evolution. As a result, new, more accurate and faster tools for genome-wide gene family classification are emerging. As more genomes are sequenced, more users will demand the appropriate tools for genome-wide classification and annotation of different gene families. Available signature databases are based on programs that perform automatic non-curated classification of gene families based on sequence analyses. This automatic classification has been shown to predict gene families with a high level of accuracy. Many gene families have been comprehensively annotated and described in detail. These annotations provide a solid base for comparative gene family classification. However, emerging plant-specific comparative genomic databases based on new clustering techniques have become more accurate tools for the prediction of new protein families. The development of these new databases and the implementation of phylogenetic tree construction methods to infer gene families and orthologous genes will be crucial once more plant genomes are sequenced. Furthermore, data integration is one of the major challenges in this field. Although many new gene family studies are being published, only a small part of this knowledge is incorporated into the gene family databases. Strategies that allow integration of published data and re-evaluation of protein-coding gene families into databases should therefore be developed.

Acknowledgements

I thank Isabel Diaz, Miguel Angel Moreno-Risueño and Ignacio Rubio-Somoza for critical reading of the manuscript. I thank database authors for useful feedback information. Financial support from the Ministerio de Ciencia e Innovación (project BFU2008-01166) is gratefully acknowledged.

References

1 Cambra, I. *et al.* (2010) Clan CD of cysteine peptidases as an example of evolutionary divergences in related protein families across plant clades. *Gene* 449, 59–69

2 Altschul, S.F. *et al.* (1990) Basic local alignment search tool. *J. Mol. Biol.* 215, 403–410

3 Pearson, W.R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol.* 183, 63–98

4 Finn, R.D. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.* 38, D211–D222

5 Sigrist, C.J. *et al.* (2010) PROSITE, a protein domain database for functional characterization and annotation. *Nucleic Acids Res.* 38, D161–D166

6 Del Bem, L.E. and Vincentz, M.G. (2010) Evolution of xyloglucan-related genes in green plants. *BMC Evol. Biol.* 10, 341

7 Lu, M. *et al.* (2010) Identification and analysis of the germin-like gene family in soybean. *BMC Genomics* 11, 620

8 Plett, D. *et al.* (2010) Dichotomy in the NRT gene families of dicots and grass species. *PLoS ONE* 5, e15289

9 Tyler, L. *et al.* (2010) Annotation and comparative analysis of the glycoside hydrolase genes in *Brachypodium distachyon*. *BMC Genomics* 11, 600

10 UniProt_Consortium (2009) The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res.* 38, D142–D148

11 Kaminuma, E. *et al.* (2011) DDBJ progress report. *Nucleic Acids Res.* 39, D22–D27

12 Leinonen, R. *et al.* (2011) The European Nucleotide Archive. *Nucleic Acids Res.* 39, D28–D31

13 Benson, D.A. *et al.* (2011) GenBank. *Nucleic Acids Res.* 39, D32–D37

14 Klimke, W. *et al.* (2009) The National Center for Biotechnology Information's Protein Clusters Database. *Nucleic Acids Res.* 37, D216–D223

15 Pruitt, K.D. *et al.* (2007) NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res.* 35, D61–D65

16 Bru, C. *et al.* (2005) The ProDom database of protein domain families: more emphasis on 3D. *Nucleic Acids Res.* 33, D212–D215

17 Letunic, I. *et al.* (2009) SMART 6: recent updates and new developments. *Nucleic Acids Res.* 37, D229–D232

18 Mi, H. *et al.* (2010) PANTHER version 7: improved phylogenetic trees, orthologs and collaboration with the Gene Ontology Consortium. *Nucleic Acids Res.* 38, D204–D210

19 de Lima Morais, D.A. *et al.* (2011) SUPERFAMILY 1.75 including a domain-centric gene ontology method. *Nucleic Acids Res.* 39, D427–D434

20 Lees, J. *et al.* (2010) Gene3D: merging structure and function for a thousand genomes. *Nucleic Acids Res.* 38, D296–D300

21 Andreeva, A. *et al.* (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.* 36, D419–D425

22 Cuff, A.L. *et al.* (2011) Extending CATH: increasing coverage of the protein structure universe and linking structure with function. *Nucleic Acids Res.* 39, D420–D426

23 Hunter, S. *et al.* (2009) InterPro: the integrative protein signature database. *Nucleic Acids Res.* 37, D211–D215

24 Marchler-Bauer, A. *et al.* (2011) CDD: a Conserved Domain Database for the functional annotation of proteins. *Nucleic Acids Res.* 39, D225–D229