Andrea Carpi
Giovanni Di Maira
Marco Vedovato
Valeria Rossi
Tiziana Naccari
Massimo Floriduz
Mario Terzi
Francesco Filippini

Department of Biology
and C.R.I.B.I.,
University of Padua,
Padua, Italy

Comparative proteome bioinformatics: Identification of a whole complement of putative protein tyrosine kinases in the model flowering plant *Arabidopsis thaliana*

Phosphorylation by protein tyrosine kinases is crucial to the control of growth and development of multicellular eukaryotes, including humans, and it also seems to play an important role in multicellular prokaryotes. A plant tyrosine-specific kinase has not been identified yet; hence, plants have been suggested to share with unicellular eukaryote yeast a tyrosine phosphorylation system where a limited number of stress proteins are tyrosyl-phosphorylated only by a few dual-specificity (serine/threonine and tyrosine) kinases. However, preliminary evidence obtained so far suggests that tyrosine phosphorylation in plants depends on the developmental conditions. Since sequencing of the genome of the model flowering plant *Arabidopsis thaliana* has been recently completed, we have performed a bioinformatic screening of the whole *Arabidopsis* proteome to identify a model complement of *bona fide* protein tyrosine kinases. *In silico* analyses suggest that $< 4\%$ of *Arabidopsis* kinases are tyrosine-specific kinases, whose gene expression has been assessed by a preliminary polymerase chain reaction screening of an *Arabidopsis* cDNA library. Finally, immunological evidence confirms that the number of *Arabidopsis* proteins specifically phosphorylated on tyrosine residues is much higher than in yeast.

Keywords: *Arabidopsis thaliana* / Bioinformatics / Motif / PROSITE / Tyrosine kinase PRO 0257

1 Introduction

Approximately 30% of cellular proteins contain covalently bound phosphate because reversible phosphorylation is a central mechanism in the modulation of protein function, able to regulate almost all aspects of cell life in both prokaryotes and eukaryotes [1]. In fact, the phosphorylation of a protein can alter its behaviour in almost every conceivable way: modulation of its intrinsic biological activity, half-life, subcellular location and docking with other proteins [1]. In eukaryotes, protein phosphorylation mainly occurs on serine (Ser), threonine (Thr) and tyrosine (Tyr) residues, hence eukaryotic protein kinases are grouped together according to their substrate specificities: serine/threonine kinases (PSTKs) specifically phosphorylate Ser and Thr residues [2], tyrosine kinases (PTKs) specifically phosphorylate Tyr residues [3], while "dual-specificity" kinases (DSKs) phosphorylate Ser/Thr as well as Tyr residues [4]. In bacteria, histidine (His), aspartic acid (Asp) and glutamic acid (Glu) are the preferred target of kinases [5]. Phosphorylation of Ser and Thr residues occurs also in prokaryotes and it is mediated by kinases substantially different from eukaryotic PSTKs [6]; instead, tyrosine phosphorylation is an extremely rare event in prokaryotes [7].

Phosphotyrosine represents a minor fraction among phosphorylated residues in eukaryotic cells [3]; this notwithstanding, PTKs and protein tyrosine phosphatases (PTPs) play a central role in mechanisms underlying control of human and animal growth and development, as many receptor-like transducers (RTKs and RPTPs) and cytoplasmic PTK/PTPs are deeply involved in the regulation of cell shaping, adhesion and migration and in responses to extracellular signals [8–10]. Tyrosine phosphorylation has been suggested to be a primary indicator of signal transduction in multicellular organisms [11], because it is virtually absent in unicellular eukaryotes, such as *Saccharomyces cerevisiae*. In fact, tyrosine phosphorylation is mediated in yeast by a few DSKs and PTPs [4, 12].

Higher plants share many Ser/Thr-specific kinases and phosphatases and a few histidine kinases, related to bacterial two-component systems [13–15]. In addition, plants also have DSKs [16, 17], of which the most known phos-

Correspondence: Dr. Francesco Filippini, Department of Biology, University of Padua, viale G. Colombo 3, 35131 Padova, Italy
E-mail: francesco.filippini@unipd.it
Fax: +39-049-8276260

Abbreviations: DSK, dual specificity kinase; **FN**, false negative; **FP**, false positive; **MAPK**, Mitogen-activated protein kinase; **PSTK**, protein serine-threonine kinase; **PTK**, protein tyrosine kinase; **PTP**, protein tyrosine phosphatase; **PYP**, phosphotyrosyl proteins; **sir**, single illegal residue; **TAIR**, The *Arabidopsis* information resource; **tir**, twin illegal residue; **USPK**, unknown-specificity protein kinase

 0173-0835/02/1111–1494 $17.50+.50/0

phorylate conserved Thr-x-Tyr motifs of mitogen-activated protein kinases (MAPK) [18, 19]. Plants also have PTPs, including the product of plant oncogene *rolB* [20–22]. Since a plant tyrosine-specific kinase has not been identified yet, it has been suggested that in plants, similarly to yeast, only DSKs mediate tyrosine phosphorylation of a few stress proteins [23]. However, yeast is a unicellular eukaryote, while higher plants are multicellular eukaryotes (as well as animals). Relevance of protein tyrosine phosphorylation for multicellular organisms is confirmed by evidence that even in the multicellular prokaryote *Myxococcus xanthus,* the patterns of tyrosine phosphorylation change during development, indicating a possible role for this regulatory modification during aggregation and sporulation [7]. Preliminary evidence obtained so far suggests a role for tyrosine phosphorylation in plant development: in fact, in *Daucus carota* tyrosine phosphorylation depends on the developmental conditions [24], while protein tyrosine kinase inhibitors are able to alter embryo pattern in fucoid algae by inhibiting the establishment of zygotic polarity [25].

Sequencing of the whole genome of model flowering plant *Arabidopsis thaliana* has recently been completed, hence its whole proteome sequence has been inferred [26]; the number of detected putative proteins is continuously updated and to date it is roughly 30 000. Although PTKs are crucial to animal cell life, they represent a minor fraction of the proteomic kinase complement. In fact, the human genome has been estimated to encode about 2000 protein kinases [27], of which roughly 100 (*i.e.*, $\sim$5%) are PTKs [28].

It cannot be excluded that PTKs might represent $<$ 5% of the plant kinase complement; hence, *Arabidopsis* PTKs might have escaped characterisation performed so far. In fact, most of the $>$ 900 putative protein kinases detected *in silico* in the *Arabidopsis* proteome have been only tentatively attributed to the Ser/Thr class [26], but functional evidence on phosphorylation specificity has been obtained only for a limited number of such kinases.

In order to identify a complement of *bona fide* plant PTKs, we have performed an *in silico* screening of the whole *Arabidopsis* proteome. Similarity searches using a PTK domain as a sequence probe (from a single kinase or derived as a consensus from any data set) result in nonspecific extraction from databases of both PSTKs and PTKs, and *vice versa*. In fact, the protein kinase domain is too conserved among Ser/Thr and Tyr kinases to allow specific detection of putative PSTKs or PTKs. This problem may be overcome by a pattern-based screening, performed using as sequence probes two signatures from the well-known "functional" database PROSITE [29], which contains "similarity-independent" amino acid signatures conserved in protein domains sharing a common biochemical moiety (*e.g.* ligand-binding, catalytic activity, post-translational modification).

Although single motifs are not able *per se* to confer specificity for Tyr phosphorylation [30], two PROSITE signatures for subdomain VIb of the catalytic site of eukaryotic protein kinases represent very specific sequence markers for PSTKs (accession PS00108) and PTKs (accession PS00109). Although they have been defined and improved almost a decade ago, their current precision (see table 1) is still very high (in the range 95–100%). In order to further improve the extraction of putative PSTKs and PTKs, we had to consider the problem of false negatives (true hits, endowed with the proband activity but showing a degenerate signature). Thus, we screened the *Arabidopsis* proteome using also slightly degenerate PROSITE signatures as sequence probes. After having extracted from the *Arabidopsis* proteome a set of signature-positive hits, each hit has been analyzed to discard proteins showing a kinase signature but representing nonkinase proteins or nonfunctional kinases (*e.g.* lacking one or more subdomains, or mutated at residues crucial to the enzyme activity). The sequences of putative PTKs have been further analyzed to group them based on their conserved or specific moieties.

Then, shifting to prediction-driven "wet" biology, an *Arabidopsis* cDNA library has been screened with specific oligonucleotide primers, in order to get preliminary evidence on the expression of genes encoding all members of the putative proteomic complement of *Arabidopsis* PTKs. Considering that the main difference between PS00108 and PS00109 signatures depends on the residue at position VII (hence, possibly even on a single nucleotide pair), amplified cDNA regions containing the PS00109 motif have been sequenced, to verify they match to the corresponding genomic sequences. Finally, immunological evidence has been obtained that the number of Tyr-phosphorylated proteins is much higher in *Arabidopsis* than in yeast.

2 Materials and methods

2.1 Bioinformatic analyses

2.1.1 PROSITE signatures PS00108 and PS00109

Complete information on signatures PS00108 and PS00109 are available at the PROSITE pages of the ExPASY server (http://www.expasy.ch/prosite); however, relevant information is reported in brief in Table 1.

Table 1. Protein kinases active-site signatures

PROSITE accession:	PS00108
Current[a] Precision[b]:	99.89%
Current[a] Recall[c]:	92.04%
PROSITE syntax[d]:	[LIVMFYC]-x-[HY]-x-D-[LIVMFY]-K-x(2)-N-[LIVMFYCT](3)
PatMatch syntax[e]:	[LIVMFYC]X[HY]XD[LIVMFY][K]XXN[LIVMFYCT][LIVMFYCT][LIVMFYCT]
PROSITE accession:	PS00109
Current[a] Precision[b]:	94.79%
Current[a] Recall[c]:	98.41%
PROSITE syntax[d]:	[LIVMFYC]-x-[HY]-x-D-[LIVMFY]-[RSTAC]-x(2)-N-[LIVMFYC](3)
PatMatch syntax[e]:	[LIVMFYC]X[HY]XD[LIVMFY][RSTAC]XXN[LIVMFYC][LIVMFYC][LIVMFYC]

a) Current Precision and Recall are calculated on data from Release 40.7 of SWISS-PROT (Dec. 2001)
b) Precision is calculated as true hits / (true hits + false positives)
c) Recall is calculated as true hits / (true hits + false negatives)
d) For details on the PROSITE syntax, see: http://www.expasy.ch/tools/scnpsit3.html
e) For details on the PatMatch syntax, see: http://www.*Arabidopsis*.org/cgi-bin/patmatch/nph-patmatch.pl#syntax

2.1.2 Pattern scanning of the *Arabidopsis* proteome

A first analysis of the complete *Arabidopsis* proteome has been performed using as scanning probes the non-degenerate PROSITE signatures PS00108 and PS00109 (see Table 1) and as scanning tool the "PatMatch" utility, available at the web site of TAIR (The *Arabidopsis* Information Resource [31]). Two further scanning steps have been performed using degenerate PS00108 and PS00109 PatMatch probes (see Table 2). In particular, eighteen "single illegal residue" (sir) probes have been derived modifying, in each pattern, only one position among I, III, V, VI, VII, X, XI, XII and XIII, *i.e.* positions accepting only one or a group of amino acid residues. Positions II, IV, VIII and IX had not to be changed as they are already completely degenerate in both original signatures (see Table 1, where they are indicated by "x"). Also "twin illegal residue" (tir) probes were created, accepting at two positions (among III, V, VII, X and XIII) only illegal residues (as reported in Table 3) found to be compatible with a conserved kinase domain. Also illegal Asn or Gln residues at position VII were considered, because their contribution to Ser/Thr or Tyr phosphorylation specificity is unknown.

2.1.3 Elimination of possible false positives from extracted sequences

The amino acid sequence of each extracted putative kinase has been used as a BLAST [32] probe to screen the Conserved Domain Database [33] (CDD, http://www.ncbi.nlm.nih.gov/Structure/cdd/wrpsb.cgi) and it was scanned for the presence of protein kinase profiles (PROSITE PS50011 and/or Pfam PF00069) analysing in the EBI BLAST output all links to the InterPro [34] web site (http://www.ebi.ac.uk/interpro/) as well as using the "ProfileScan" tool at the ExPASY server (http://www.expasy.ch). Sequences showing a PS00108 or PS00109 signature but lacking a conserved kinase domain or showing a kinase domain truncated at crucial regions were considered as possible false positives (FP).

Similarity searches were performed screening nonredundant protein and/or nucleotide databases (Swissall or GenPept, EMBL or GenBank) at the European (http://www.ebi.ac.uk/blast2) or American (http://www.ncbi.nlm.nih.gov/BLAST) bioinformatic servers by BLASTP and TBLASTN [32], using BLOSUM62 matrix and default settings. Homologous sequences were aligned using the CLUSTALW [35] tool available at the EBI server (http://www2.ebi.ac.uk/clustalw) and the Megalign software (DNASTAR). Receptor-like or cytoplasmic locations were predicted using PSORT [36] and further transmembrane prediction tools available at the ExPASY server.

2.2 cDNA library, PCR amplification and analysis of the amplified fragments

The Arabidopsis 3' Matchmaker cDNA library (Clontech, Palo Alto, CA, USA) was derived by reverse transcription of mRNAs from pooled green vegetative tissues of 3-week old *Arabidopsis* plants, wild-type ecotype Columbia. cDNAs of the library are cloned in *Eco RI* site in pGAD10 plasmids (GenBank Accession: U13188) in

Escherichia coli cells. The library was amplified on Petri dishes with agarised culture medium and then divided into ten aliquots, each corresponding to 6×10^5 independent cDNA clones. Plasmid DNA from replicates of each aliquot was extracted using Plasmid Mega kits (Qiagen, Hilden, Germany) and stored at −80°C.

PCR primers for the amplification of the motif-containing region of each extracted putative PTK were projected using in combination Oligo 4.05 (National Biosciences, Plymouth, MN, USA), Primer Express (Applied Biosystems, Foster City, CA, USA) and Primer Select (DNASTAR, Madison, WI, USA). Primers were projected in order to amplify, for each putative PTK, a coding region (300 to roughly 700 bp; average length = 502 bp) including the degenerate or nondegenerate PROSITE motif used as a probe to extract the kinases. Analytical PCR amplifications were performed in a 20 μl reaction volume in a model 9700 thermal cycler (Applied Biosystems), using the Expand™ High-Fidelity *Taq* polymerase and dNTP mix (Roche, Basel, Switzerland), 100 ng of library DNA and 60 pmoles each of reverse and forward primer. PCR amplification started with a 5 min denaturation step at 95°C, followed by 35 to 45 cycles of (denaturation, 30 s at 95°C; annealing, 45 s at varying temperature, depending on the primers' couple; elongation, 50 s at 72°C) and by a final elongation step (5 min at 72°C) before chilling down to 4°C. Amplified DNA fragments were electrophoresed together with 100 bp DNA ladder (Amersham Biosciences, Little Chalfont, Bucks, UK) onto 1.5% agarose gel containing 0.5 μg/mL ethidium bromide, and the molecular weight of the amplified fragments was determined using a Chemi-Doc image documentation system and Quantity One software (Bio-Rad, Hercules, CA, USA). Restriction analyses were performed using separately each appropriate restriction enzyme on 0.5 μg of any proband amplified fragment.

2.3 Plant and yeast culture, tissue explants

Arabidopsis thaliana seeds, ecotype Columbia (Lehle Seeds, Round Rock, TX, USA), were sterilised by standard ipochloride/ethanol/water washings, incubated 48 h at 4°C (vernalisation) and then put onto wet ground or incubated in germination buffer, *i.e.* Murashige & Skoog medium (Duchefa, Haarlem, The Netherlands) at 50% salt concentration. Germinating seeds and plantlets were incubated at 24°C, 80–90% humidity. *Arabidopsis* plants were washed with tap water, then with distilled water; after washing, plants were immediately homogenized or frozen with liquid nitrogen. Wild-type *S. cerevisiae* cells were grown for 8 h in standard YPAD medium (Invitrogen, Carlsbad, CA, USA) at 30°C under continuous agitation at 270 rpm, collected and then immediately homogenized or frozen with liquid nitrogen.

2.4 Sample preparation and immunoblotting

Plant homogenates, SDS-PAGE [37], immunoblotting, displacement with phosphoamino acids and on-blot dephosphorylation were performed as reported [24]. Proteins were extracted from yeast cells using the Y-PER reagent (Perbio, Helsingborg, Sweden), following the manufacturer's instructions. Chemiluminescent signal was followed before saturation using a Chemi-Doc image documentation system and Quantity One software (Bio-Rad).

3 Results

3.1 Preliminary *in silico* screening of the *Arabidopsis* proteome

As a first step towards an *in silico* identification of a model complement of putative plant PTKs, screening of the whole *Arabidopsis* proteome has been performed using as a scanning tool the "PatMatch" program available at the TAIR [31] web site (http://www.*Arabidopsis*.org/home.html), and as pattern probes two PROSITE signatures (accessions: PS00108, PS00109), which are specific to the catalytic domain of respectively PSTKs and PTKs (see Section 2.1.2 and Table 1). Setting the database option as: "GenPept, PIR and SWISS-PROT", the PS00108 and PS0019 probes extracted, respectively, 1016 and 39 hits. However, after having eliminated 68 redudant sequences, the correct hit numbers were, respectively, 955 and 32.

A further group of putative kinases has been identified by a second screening step, performed using eighteen slightly degenerate "single illegal residue" (sir) patterns, derived from PS00108 or PS00109 (Section 2.1.2 and Table 2). All sequences, extracted by degenerate or nondegenerate PS00108/PS00109 signatures, were scanned for the presence of a region homologous to the conserved catalytic domain of protein kinases. This allowed us to discard, as possible false positives (FPs), sequences sharing a PS00108 or PS00109 signature but lacking a kinase domain (Table 3).

Roughly 14% of sequences extracted by signature PS00108 are possible FPs; hence, PS00108 precision calculated on the *Arabidopsis* proteome is lower than the ≤100% value reported in the PROSITE documentation page. However, we have to take into account that the data set used to calculate the precision of PROSITE signatures is SWISS-PROT, while proteomes inferred from whole genome sequences, including the *Arabidopsis* proteome, contain several false coding sequences (wrong gene predictions and pseudogenes).

Table 2. Single illegal residue (sir) patterns in PatMatch syntax[a)]

Sir-PS00108 patterns:	
I:	[^LIVMFYC]X[HY]XD[LIVMFY]KXXN[LIVMFYCT][LIVMFYCT][LIVMFYCT]
III:	[LIVMFYC]X[^HY]XD[LIVMFY]KXXN[LIVMFYCT][LIVMFYCT][LIVMFYCT]
V:	[LIVMFYC]X[HY]X[^D][LIVMFY]KXXN[LIVMFYCT][LIVMFYCT][LIVMFYCT]
VI:	[LIVMFYC]X[HY]XD[^LIVMFY]KXXN[LIVMFYCT][LIVMFYCT][LIVMFYCT]
VII:	[LIVMFYC]X[HY]XD[LIVMFY][^K]XXN[LIVMFYCT][LIVMFYCT][LIVMFYCT]
X:	[LIVMFYC]X[HY]XD[LIVMFY]KXXN[^N][LIVMFYCT][LIVMFYCT][LIVMFYCT]
XI:	[LIVMFYC]X[HY]XD[LIVMFY]KXXN[^LIVMFYCT][LIVMFYCT][LIVMFYCT]
XII:	[LIVMFYC]X[HY]XD[LIVMFY]KXXN[LIVMFYCT][^LIVMFYCT][LIVMFYCT]
XIII:	[LIVMFYC]X[HY]XD[LIVMFY]KXXN[LIVMFYCT][LIVMFYCT][^LIVMFYCT]
Sir-PS00109 patterns:	
I:	[^LIVMFYC]X[HY]XD[LIVMFY][RSTAC]XXN[LIVMFYC][LIVMFYC][LIVMFYC]
III:	[LIVMFYC]X[^HY]XD[LIVMFY][RSTAC]XXN[LIVMFYC][LIVMFYC][LIVMFYC]
V:	[LIVMFYC]X[HY]X[^D][LIVMFY][RSTAC]XXN[LIVMFYC][LIVMFYC][LIVMFYC]
VI:	[LIVMFYC]X[HY]XD[^LIVMFY][RSTAC]XXN[LIVMFYC][LIVMFYC][LIVMFYC]
VII:	[LIVMFYC]X[HY]XD[LIVMFY][^RSTAC]XXN[LIVMFYC][LIVMFYC][LIVMFYC]
X:	[LIVMFYC]X[HY]XD[LIVMFY][RSTAC]XX[^N][LIVMFYC][LIVMFYC][LIVMFYC]
XI:	[LIVMFYC]X[HY]XD[LIVMFY][RSTAC]XXN[^LIVMFYC][LIVMFYC][LIVMFYC]
XII:	[LIVMFYC]X[HY]XD[LIVMFY][RSTAC]XXN[LIVMFYC][^LIVMFYC][LIVMFYC]
XIII:	[LIVMFYC]X[HY]XD[LIVMFY][RSTAC]XXN[LIVMFYC][LIVMFYC][^LIVMFYC]

a) For details on PatMatch syntax, see: http://www.*Arabidopsis*.org/cgi-bin/patmatch/nph-patmatch.pl#syntax

It is noteworthy that, among the 28 putative PTKs extracted by PS00109, two showed a conserved pattern for RIO1 family proteins (PROSITE accession: PS01245, *E* value: 8e-77). We found that a signature degeneration is more acceptable in putative PSTK than in putative PTKs. In fact, while roughly 66% of sequences extracted by sir-PS00108 patterns share a conserved kinase domain, less than 11% of sequences extracted by sir-PS00109 are putative PTKs (Table 3). In particular, only a few illegal residues are compatible with a conserved kinase domain in putative PTKs (Table 3).

After the FP elimination step, only eight putative kinases sharing an illegal residue at position VII (Asn or Gln) were found, of which one showed a PS01245 pattern for RIO1 proteins. The other seven were named unknown-specificity protein kinases (USPKs); in fact, their capacity to phosphorylate specifically either Ser/Thr or Tyr residues cannot be inferred *a priori*, as just position VII is crucial to discriminate among PSTKs (sharing only residue K at VII) and PTKs (sharing R, S, T, A or C).

A third screening was performed using ten "twin illegal residues" patterns (tir-PS00109, see Section 2.1.2) to identify *in silico* further putative PTKs. It is noteworthy that, among extracted sequences (Table 4), six out of the eight putative PTKs were extracted by patterns degenerate at position V. After these first steps of pattern search and elimination of most FPs, fifty *Arabidopsis* putative PTK sequences were retained for further analyses, of which 24 were extracted by degenerate PS00109 patterns. In addition, three putative RIO1 kinases and nine putative USPKs were retained for further analysis.

Table 3. *Arabidopsis* sequences extracted by original and degenerate PS00108/PS00109 patterns

Pattern: PS00108				PS00109			
	ES[a)]	FP[a)]	pPSTK[b)]	ES[a)]	FP[a)]	pPTK[b)]	
	955	132	823	32	4	28	
Pattern: sir-PS00108				sir-PS00109			
	ES[a)]	FP[a)]	pPSTK[b)]	ES[a)]	FP[a)]	pPTK[b)]	cir[d)]
sir[c)]:							
I	24	1	23	0	0	0	–
III	22	15	7	49	48	1	N
V	68	20	48	35	23	12	[ACIMNR]
VI	1	1	0	1	1	0	–
X	19	10	9	45	43	2	[IK]
XI	3	3	0	4	4	0	–
XII	2	1	1	13	13	0	–
XIII	12	0	12	2	1	1	G
Total	151	51	100	149	133	16	

a) ES, Extracted Sequences; FP, False Positive sequences;
b) pPSTK, putative PSTKs; pPTK, putative PTKs;
c) degenerate position representing the single illegal residue (sir)
d) cir, illegal residues compatible with a putative PTK domain.

Table 4. *Arabidopsis* sequences extracted by tir-PS00109 patterns

irp[a)] 1	irp[a)] 2	ES[b)]	FP[b)]	pPTK[c)]	pPUSK[c)]
III	V	11	10	1	–
III	VII	0	0	–	0
III	X	1	1	0	–
III	XIII	1	0	1	–
V	VII	6	4	–	2
V	X	38	33	5	–
V	XIII	1	1	0	–
VII	X	0	0	–	0
VII	XIII	0	0	–	0
X	XIII	1	0	1	–
Total		59	49	8	2

a) irp 1 and 2, illegal residue positions 1 and 2
b) ES, Extracted Sequences; FP, False Positive sequences
c) pPTK, putative PTK; pPUSK, putative Protein Unknown-Specificity Kinase

3.2 Further screenings to eliminate false positives

In order to eliminate eventual FPs having escaped the first selection steps, the group of 59 putative PTKs + USPKs was scanned for the presence of profiles PROSITE PS50011 and/or Pfam PF00069, both specific to the conserved catalytic domain of eukaryotic protein kinases [38, 39]. In addition, the presence of either a single or multiple kinase signature was assessed for each sequence. Such an approach could not be followed to identify eventual FPs among putative RIO1 kinases, because the catalytic properties of RIO1 proteins are still unknown and even their kinase activity is only hypothetical [40].

Six putative PTKs and one putative USPK were discarded as possible FPs or putative PSTKs, of which (i) two showed an overall length roughly half a complete catalytic kinase domain; (ii) two lacked a kinase profile; and (iii) three showed a PS00108 signature at the right domain position.

Further analysis of the 44 putative PTKs and eight putative USPKs having overcome the profile/second signature scanning step was performed by aligning the putative catalytic domain of each sequence to the conserved catalytic domain of eukaryotic protein kinases [38, 39], starting from the PS00109 signature at subdomain VIb. This step was aimed at identifying sequences lacking one or more subdomain/region/motif known to be crucial to the catalytic activity [39, 41–49], such as *e.g.* the activation loop, which generally begins with a conserved Asp-Phe-Gly sequence (DFG motif) and ends at a conserved Ala-Pro-Glu (APE motif) [49]. This step allowed us to eliminate eight possible FPs, of which four lacked two or more *N*- or *C*-terminal subdomains, and four showed mutations at crucial motifs such as *e.g.* DFG and APE. Hence, 47 putative kinases (36 PTKs, eight USPKs and three RIO1 kinases) were retained for further analyses.

3.3 Analysis of sequences from false negatives

PS00108 and PS00109 signatures show high precision values; however, false negatives (FNs, *i.e.* proteins with demonstrated PSTK or PTK activity but not extracted by signatures cited above) from the SWISS-PROT+TrEMBL database can be analysed to get suggestions on the possible specificity of putative *Arabidopsis* kinases extracted by the degenerate patterns. In fact, we considered the possibility that different sets of "illegal residues" might be specifically represented in the degenerate signatures of known FN proteins endowed with either PSTK or PTK activity. This might suggest a possible specificity (PSTK or PTK) for putative kinases extracted by the degenerate patterns, *i.e.* representing putative FN kinases. Such a prediction analysis was performed in three steps: (1) analysis of illegal residues in the PROSITE signatures from known FN kinases; (2) pattern match screening of the SWISS-PROT/TrEMBL database with all sir- and tir-PS00109 patterns found to be compatible with a conserved kinase domain in the *Arabidopsis* proteome; and (3) analysis of links among specific illegal residues and kinase activity.

Most FN-PSTKs and FN-PTKs from the PROSITE information pages for PS00108 and PS00109 show single illegal residues, which are ≥ 2 in six FN-PSTKs. Mismatches in the PS00109 signature from FN-PTKs can be found at positions I, V and XI, while in the PS00108 signature from FN-PSTK they concern almost all positions, including VII. A screening of the SWISS-PROT+TrEMBL database with those sir- and tir-PS00109 patterns having extracted putative *Arabidopsis* PSTKs or PTKs resulted in the extraction of several animal and nonanimal sequences, of which 17 were FN kinases with experimentally demonstrated Ser/Thr or Tyr specificity, not already reported in the PROSITE documentation list.

Such extracted FNs showed illegal residues only at positions V, VII and X; in particular, Asn at position V was shared by five FN-PTKs, while Asn or Gln illegal residues at position VII, characterising USPKs, were shared by 12 FN PSTKs (Table 5). This latter finding strongly suggests that USPKs are possibly PSTKs and not PTKs. Instead, a possible specificity for Ser/Thr or Tyr residues cannot be inferred in the case of putative kinases sharing an illegal residue Asn at position V and/or Lys at position X. Also for the other illegal residues from extracted

Table 5. Illegal residues degenerating PS00108/PS00109 signatures in reported or extracted[a] false negative PSTKs (S) or false negative PTKs (Y)

IR[b]	Position								
	I	III	V	VI	VII	X	XI	XII	XIII
Ala	–	–	–	–	–	–	–	S (12)	–
Arg	S (1)	–	–	–	S (1)	–	–	–	S (3)
Asn	–	–	Y (1+5); S (3)	–	S (1)	–	–	–	–
Asp	–	–	–	–	–	S(1)	–	–	–
Cys	–	–	–	S (1)	–	–	–	–	–
Gln	–	–	–	–	–	S(1)	–	–	–
Glu	–	–	S (1)	–	S (2)	–	–	–	–
Gly	–	–	–	–	–	–	–	–	–
His	–	–	–	–	S (1)	–	–	–	–
Ile	–	–	–	–	–	–	–	–	–
Leu	–	–	–	–	–	–	–	–	–
Lys	S (1)	–	S (2)	–	–	Y(+1); S (1)	–	–	–
Met	–	–	–	–	–	–	–	–	–
Phe	–	S (1)	–	–	–	–	–	–	–
Pro	–	–	–	–	–	–	–	–	–
Ser	Y (2)	–	–	–	S (1)	–	–	–	S (1)
Thr	S (2)	–	–	–	S (1)	–	Y(1)	–	–
Trp	S (1)	–	–	–	–	–	–	–	–
Tyr	–	–	–	–	–	–	–	–	–
Val	–	–	–	–	–	–	–	–	–

a) Number of *extracted* FNs is reported between parentheses as "+n", soon after number of *reported* FNs
b) IR, illegal residue

Arabidopsis putative PTKs, no further indications can be inferred analysing sequences from known and extracted FNs. After the FN analysis, only three putative RIO1 kinases and 36 putative PTKs were retained for the last steps of *in silico* classification, PCR screening and sequencing of the motif-containing regions.

3.4 *In silico* analysis of the putative PTKs

In order to get suggestions on possible functions and involvement in cell pathways of each identified putative PTKs, similarity analysis and domain/profile/pattern scanning of each sequence were performed. BLAST [32] searches, CLUSTALW [35] alignments and PSORT [36] predictions allowed us to group the putative PTKs in three major groups.

Group 1 kinases seem to fall in the RLK-Pelle group reported by Shiu and Bleecher [51], although they are putatively endowed with Tyr-specific activity. Group 1 PTKs can be further divided in three subgroups: (i) putative RTKs, sharing a receptor-like topology and, in most members, leucine-rich repeats (LRR) [52] in the predicted extracellular domain; (ii) putative cytoplasmic PTKs related to Pti1, a downstream interactor of the Pto kinase (a plant homologue of Drosophila Pelle [53]), and (iii) putative PTKs related to wall-associated kinases (WAK) [54].

Group 2 consists of proteins sharing a RIO1 family domain and putatively endowed with tyrosine kinase activity, while kinases in Group 3 are similar to MAPK kinases or eukaryotic initiation factor 2 alpha (eIF2alpha) protein kinases [55], which are both endowed with DSK activity. A periodically updated table, reporting such groups of putative kinases with links to their EBI and TAIR entry pages, is located at our web page: http://www.bio.unipd.it/molbinfo/Docs.html.

3.5 Expression of the putative kinase genes and validation of their signatures

A region from each putative kinase cDNA, containing sequence encoding the PROSITE signature, was amplified by PCR using appropriate primer pairs and all ten aliquots of an *Arabidopsis* cDNA library derived from wild-type green tissues. This allowed us to get a preliminary idea of the expression of each putative gene and to confirm signature-coding sequences.

Panels in Fig. 1 are representative results from PCR experiments to amplify the signature-containing region of each putative PTK gene. The expression level is conven-

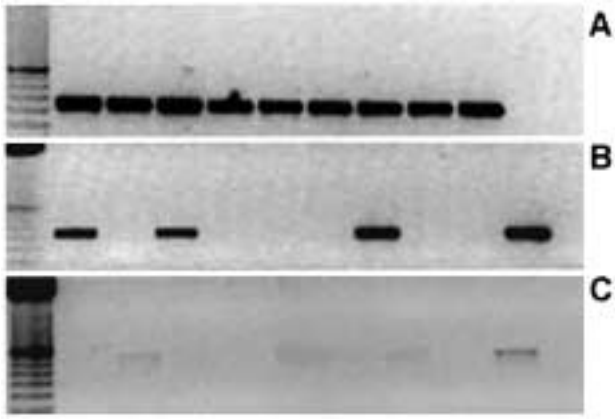

Figure 1. PCR screening of the ten aliquots of a 3'-Matchmaker cDNA library from *Arabidopsis thaliana* green tissues. Roughly 150 ng of amplified DNA were electrophoresed in 1% Tris/acetate/EDTA on 1.5% agarose gel, in the presence of 0.5 μg/mL ethidium bromide. Amplification of gene encoding P9Q4AS (panel A) is representative for "high" expression level. Results shown for the amplification of genes Q9LSC2 (panel B) and Q9FHT0 (panel C) are representative for "intermediate" and "low" expression levels, respectively.

tionally reported here as "high" or "intermediate" when a right-length fragment was specifically amplified in respectively ≥ 8 (*e.g.* panel A), or ≥ 4 aliquots (*e.g.* panel B). Finally, panel C is an example of "low" expression, while black lanes after up to 45 cycles (not shown) were arbitrarily considered as "no" expression (at least in terms of absence in the cDNA library). "High" or "intermediate" expression levels are shown by >75% of screened genes, while expression of roughly 15% genes was undetectable. Expression data are also reported and regularly updated at our web site, in the previously cited table. Finally, sequencing confirmed genomic data for all amplified fragments.

3.6 Comparative immunoanalysis of phosphotyrosine patterns in *Arabidopsis* and yeast

Proteins from freshly prepared "total extracts" [24] from suspension-growing yeast or *Arabidopsis* cells were electrophoresed by SDS-PAGE, blotted onto nitrocellulose filters and immunostained using recombinant antiphosphotyrosine antibodies [24]. Immunoblotting experiments were performed in replicates and tyrosine phosphorylation specificity was confirmed as reported [24].

Figure 2 shows that *Arabidopsis* total extract contains a much higher number of phosphotyrosyl-proteins (PYPs) than yeast, from low to high molecular weight. Obviously, there is no demonstrated link between the PYPs in the blot and the putative PTKs identified *in silico*, because Tyr phosphorylation may depend on both auto- or trans-phosphorylation as well as on the phosphorylation of PTK substrates. Instead, our data demonstrate that PYP patterns from *Arabidopsis* are more similar to animal than yeast patterns; hence, possibly in plants several kinases are capable of phosphorylating Tyr residues.

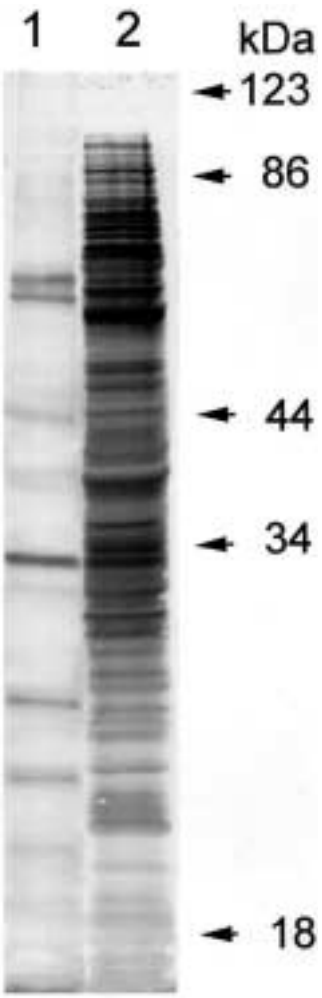

Figure 2. Immunoblotting with RC20 recombinant antibodies [24] of crude extracts (30 μg/lane) from four-day, wild-type yeast (left lane) and *Arabidopsis* (right lane) cells. Molecular weight markers are indicated.

4 Discussion

The *Arabidopsis* genome encodes proteins from 11 000 families, similar to the functional diversity of *Drosophila melanogaster* and *Caenorhabditis elegans* and not so far from *Homo sapiens* [31]. This notwithstanding, a very important regulatory system such as tyrosine phosphorylation is yet uncharacterised in the model flowering plant *Arabidopsis*, although preliminary evidence suggests that tyrosine phosphorylation depends in plants on the developmental conditions [24]. The discovery of plant PTKs would confirm the existence of a complete protein tyrosine phosphorylation system in plants, starting the identification of elements possibly crucial to very important developmental transduction pathways.

As an alternative to start a complex biochemical approach to purify PTK-like activities, we followed a bioinformatic approach to get preliminary identification in the *Arabidopsis* proteome of the subset of protein sequences representing those putative kinases possibly able to phosphorylate Tyr residues (putative PTKs, DSKs and, perhaps, RIO1 kinases).

Screening with highly specific PROSITE signatures PS00108 and PS00109 has shown that the putative PTK/PSTK ratio in plants is similar to the ratio in animals. In fact, eliminating FPs and analysing all putative kinase sequences allowed us to identify 39 putative Tyr-phosphorylating enzymes (32 PTKs, 4 DSKs and 3 RIO1-like), representing roughly 3.5% of the *Arabidopsis* protein kinases, similarly to 4–5% hypothesized for humans. Most putative PTKs overcoming the FP-screening have

been extracted by the nondegenerate PS00109 signature; however, a significant number of further putative PTKs has been identified also by slightly degenerate patterns. Further, the analysis of known and newly extracted FN kinases (not only from plants) suggests that, in addition to Lys, residues Asn or Gln at position VII of the PROSITE signature might indicate PSTK activity.

It is noteworthy that, as well as in animals, RTKs are highly represented among PTKs; in fact, most putative PTKs from *Arabidopsis* seem to fall in the RLK-Pelle superfamily, which contains kinases playing a major role in the control of organ and tissue development, such as *e.g.* wall-associated kinases (WAKs)s, expressed at organ junctions, in shoot and root apical meristems, in expanding leaves [54].

A possible involvement in developmental pathways of the putative PTKs found in the *Arabidopsis* proteome sequence is in agreement to our previous evidence on a high number of PYP bands in carrot and on their variation depending on the developmental conditions [24]; experimental evidence from this work shows now in the model plant species *Arabidopsis thaliana* that the extent of tyrosine phosphorylation is much higher than in yeast, contradicting previous hypotheses on a similar low-complexity Tyr-phosphorylation system in both yeast and plants [23].

Deeper analyses concerning protein tyrosine phosphorylation patterns from *Arabidopsis* young and adult plants and the effect of PTK inhibitors (Floriduz *et al.*, manuscript in preparation) further demonstrate that plant cell PYP patterns and their developmental variation are more related to animal than to yeast cells. The identification of putative PTKs related to stress or mitogenic signalling pathways suggests that, similarly to animal PTKs, they might be involved also in the control of growth and stress responses, while the presence of PTK-specific signatures in putative kinases related to DSKs or to RIO1 suggests that these elements might be endowed with (or also with) tyrosine phosphorylation activity.

We have found that most genes encoding putative PTKs identified *in silico* are expressed in green vegetative tissues; however, lack of expression is demonstrated by the absence of PCR amplification, which might depend on (i) root-specific expression, (ii) expression at a different developmental stage or (iii) absence of an inducing signal.

Obviously, we cannot exclude that some identified genes might be catalytically inactive or endowed with a PSTK activity, but the high precision of PS00109 signature and preliminary experimental evidence suggest that the group of identified kinases may represent a good "starting point" to identify and characterize the first plant PTK and to unravel roles and functions of the tyrosine phosphorylation system in the plant kingdom.

We thank Raffaella Picco for useful discussions and encouragement and Sabrina Canova for her "lucky influence" in our screenings. This research has been supported by MURST ex 60% 2001 to FF and by CNR PF "Biotecnologie" to MT.

Received September 18, 2001

5 References

[1] Cohen, P., *Trends Biochem. Sci.* 2000, *25*, 596–601.

[2] Edelman, A. M., Blumenthal, D. K., and Drebs, E. G., *Annu. Rev. Biochem.* 1987, *56*, 567–613.

[3] Hunter, T., Cooper, J. A., *Annu. Rev. Biochem.* 1985, *54*, 897–930.

[4] Lindberg, R. A., Quinn, A. M., Hunter, T., *Trends Biochem. Sci.* 1992, *17*, 114–119.

[5] Yan, J. X., Packer, N. H., Gooley, A. A., Williams, K. L., *J. Chromatogr. A* 1998, *808*, 23–41.

[6] Cozzone, A. J., *J. Cell. Biochem.* 1993, *51*, 7–13.

[7] Frasch, S. C., Dworkin, M., *J. Bacteriol.* 1996, *178*, 4084–4088.

[8] Lemmon, M. A., Schlessinger, J., *Trends Biochem. Sci.* 1994, *19*, 459–463.

[9] Sato, T. N., Tozawa, Y., Deutsch, U., Wolburg-Buchholz, K. *et al.*, *Nature* 1995, *376*, 70–74.

[10] Brady-Kalnay, S. M., Tonks, N., *Curr. Opin. Cell Biol.* 1995, *7*, 650–657.

[11] Ullrich, A., Schlessinger, J., *Cell* 1990, *61*, 203–212.

[12] Stern, D., Zheng, P., Beidler, D. R., Zerillo, C., *Mol. Cell. Biol.* 1991, *11*, 987–1001.

[13] Stone, J., Walker, J. C., *Plant Physiol.* 1995, *108*, 451–457.

[14] Smith, R. D., Walker, J. C., *Annu. Rev. Plant Physiol. Plant Mol. Biol.* 1996, *47*, 101–125.

[15] Urao, T., Miyata, S., Yamaguchi-Shinozaki, K., Shinozaki, K., *FEBS Lett.* 2000, *478*, 227–232.

[16] Hirayama, T., Oka, A., *Plant Mol. Biol.* 1992, *20*, 653–662.

[17] Ali, N., Halfner, U., Chua, N. H., *J. Biol. Chem.* 1994, *269*, 31626–31629.

[18] Mizoguchi, T., Hayashida, N., Yamaguchi-Shinozaki, K., Kamada, H., Shinozaki, K., *FEBS Lett.* 1993, *336*, 440–444.

[19] Zhang, S., Klessig, D. F., *Plant Cell* 1997, *9*, 809–824.

[20] Haring, M. A., Siderius, M., Jonak, C., Hirt, H. *et al.*, *Plant J.* 1995, *7*, 981–988.

[21] Fordham-Skelton, A. P., Skipsey, M., Eveans, I. M., Edwards, R., Gatehouse, J. A., *Plant Mol. Biol.* 1999, *39*, 593–605.

[22] Filippini, F., Rossi, V., Marin, O., Trovato, M. *et al.*, *Nature* 1996, *379*, 499–500.

[23] Xu, Q., Fu, H.-H., Gupta, R., Luan, S., *Plant Cell* 1998, *10*, 849–857.

[24] Barizza, E., Lo Schiavo, F., Terzi, M., Filippini, F., *FEBS Lett.* 1999, *447*, 191–194.

[25] Corellou, F., Potin, P., Brownlee, C., Kloareg, B., Bouget, F. Y., *Dev. Biol.* 2000, *219*, 165–182.

[26] The *Arabidopsis* Genome Initiative, *Nature* 2000, *408*, 796–815.

[27] Stenberg, K. A., Riikonen, P. T., Vihinen, M., *Nucleic Acids Res.* 1999, *27*, 362–364.

[28] Robinson, D. R., Wu, Y. M., Lin, S. F., *Oncogene* 2000, *19*, 5548–5557.

[29] Falquet, L., Pagni, M., Bucher, P., Hulo, N. *et al.*, *Nucleic Acids Res.* 2002 *30*, 235–238.

[30] Carrera, A. C., Borlado, L. R., Roberts, T. M., Martinez, C., *Biochem. Biophys. Res. Commun.* 1994, *205*, 1114–1120.

[31] Huala, E., Dickerman, A. W., Garcia-Hernandez, M., Weems, D. *et al.*, *Nucleic Acids Res.* 2001 *29*, 102–105.

[32] Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. *et al.*, *Nucleic Acids Res.* 1997, *25*, 3389–3402.

[33] Marchler-Bauer, A., Panchenko, A. R., Shoemaker, B. A., Thiessen, P. A. *et al.*, *Nucleic Acids Res.* 2002 *30*, 281–283.

[34] Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A. *et al.*, *Nucleic Acids Res.* 2001, *29*, 37–40.

[35] Thompson, J. D., Higgins, D. G., Gibson, T. J., *Nucl. Acids Res.* 1994, *22*, 4673–4680.

[36] Nakai, K., Kanehisa, M., *Genomics* 1992, *14*, 897–911, http://psort.nibb.ac.jp.

[37] Laemmli, U. K., *Nature* 1970, *277*, 680–685.

[38] Hanks, S. K., Quinn, A. M., Hunter, T., *Science* 1988, *241*, 42–52.

[39] Hanks, S. K., Hunter, T., *FASEB J.* 1995, *9*, 576–596.

[40] Leonard, C. J., Aravind, L., Koonin, E. V., *Genome Res.* 1998, *8*, 1038–1047.

[41] Gibbs, C. S., Zoller, M. J., *J. Biol. Chem*. 1991, *266*, 8923–8931.

[42] Knighton, D. R., Zheng, J. H., Ten Eyck, L. F., Ashford, V. A. *et al.*, *Science* 1991, *253*, 407–413.

[43] Zheng, J., Knighton, D. R., ten Eyck, L. F., Karlsson, R. *et al.*, *Biochemistry* 1993, *32*, 2154–2161.

[44] Vetrie, D., Vorechovsky, I., Sideras, P., Holland, J. *et al.*, *Nature* 1993, *361*, 226–233.

[45] Bossemeyer, D., *Trends Biochem. Sci.* 1994, *19*, 201–205.

[46] Taylor, S. S., Radzio-Andzelm, E., *Structure* 1994, *2*, 345–355.

[47] Cox, S., Radzio-Andzelm, E., Taylor, S. S., *Curr. Opin. Struct. Biol.* 1994, *4*, 893–901.

[48] Katso, R. M., Russell, R. B., Ganesan, T. S., *Mol. Cell Biol.* 1999, *19*, 6427–6440.

[49] Skamnaki, V. T., Owen, D. J., Noble, M. E., Lowe, E. D. *et al.*, *Biochemistry* 1999, *38*, 14718–14730.

[50] Johnson, L. N., Noble, M. E. M., Owen, D. J., *Cell* 1996, *85*, 149–158.

[51] Shiu, S. H., Bleecher, A. B., *Sci. STKE.* 2001, *113*, RE22.

[52] Ceulemans, H., De Maeyer, M., Stalmans, W., Bollen, M., *FEBS. Lett.* 1999, *456*, 349–351.

[53] Bogdanove, A. J., Martin, G. B., *Proc. Natl. Acad. Sci. USA* 2000, *97*, 8836–8840.

[54] Wagner, T. A., Kohorn, B. D., *Plant Cell.* 2001, *13*, 303–318.

[55] Lu, J., O'Hara, E. B., Trieselmann, B. A., Romano, P. R., Dever, T. E., *J. Biol. Chem.* 1999, *274*, 32198–32203.