

Research

# The repertoire of protein kinases encoded in the draft version of the human genome: atypical variations and uncommon domain combinations

A Krupa and N Srinivasan

Address: Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560 012, India.

Correspondence: N Srinivasan. E-mail: ns@mbu.iisc.ernet.in

Published: 13 November 2002

*Genome Biology* 2002, **3**(12):research0066.1–0066.14

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2002/3/12/research/0066>

© 2002 Krupa and Srinivasan, licensee BioMed Central Ltd  
(Print ISSN 1465-6906; Online ISSN 1465-6914)

Received: 1 July 2002

Revised: 25 September 2002

Accepted: 11 October 2002

## Abstract

**Background:** Phosphorylation by protein kinases is central to cellular signal transduction. Abnormal functioning of kinases has been implicated in developmental disorders and malignancies. Their activity is regulated by second messengers and by the binding of associated domains, which are also influential in translocating the catalytic component to their substrate sites, in mediating interaction with other proteins and carrying out their biological roles.

**Results:** Using sensitive profile-search methods and manual analysis, the human genome has been surveyed for protein kinases. A set of 448 sequences, which show significant similarity to protein kinases and contain the critical residues essential for kinase function, have been selected for an analysis of domain combinations after classifying the kinase domains into subfamilies. The unusual domain combinations in particular kinases suggest their involvement in ubiquitination pathways and alternative modes of regulation for mitogen-activated protein kinase kinases (MAPKKs) and cyclin-dependent kinase (CDK)-like kinases. Previously unexplored kinases have been implicated in osteoblast differentiation and embryonic development on the basis of homology with kinases of known functions from other organisms. Kinases potentially unique to vertebrates are involved in highly evolved processes such as apoptosis, protein translation and tyrosine kinase signaling. In addition to coevolution with the kinase domain, duplication and recruitment of non-catalytic domains is apparent in signaling domains such as the PH, DAG-PE, SH2 and SH3 domains.

**Conclusions:** Expansion of the functional repertoire and possible existence of alternative modes of regulation of certain kinases is suggested by their uncommon domain combinations. Experimental verification of the predicted implications of these kinases could enhance our understanding of their biological roles.

## Background

Many cellular stimuli are channeled through a number of signaling molecules that result in alterations in the transcriptional and translational status of gene(s) to bring about the desired response. Reversible protein phosphorylation is one of the key mechanisms commonly used in signal transduction

to alter the functional states of the signaling proteins. Covalent attachment of a phosphate group to target proteins by protein kinases modulates the functional status of a myriad of proteins involved in cellular signaling networks. Protein kinases are essential to the regulation of diverse cellular processes including metabolism, stress responses,

transcription, translation, DNA replication and cell-cycle-control [1,2]. In the multicellular eukaryotes they are also involved in more evolved functions such as organ and limb development, neuronal signaling, apoptosis and cell-cell communication [3,4]. Abnormal functioning of protein kinases has been implicated in a large number of human diseases and many of them have been identified as proto-oncogenes [5,6].

Serine/threonine kinases and tyrosine kinases make up one of the largest known protein families. Commonly referred to as the eukaryotic protein kinases (ePKs), these share a common three-dimensional fold [7]. The catalytic domain comprises an amino-terminal lobe containing mostly  $\beta$  strands and a carboxy-terminal lobe containing mainly  $\alpha$  helices. The ATP-binding and substrate-binding sites are located in and around the cleft between the two lobes. The ePKs have been further classified into various subfamilies on the basis of the sequence similarity of the catalytic kinase domains [1,8,9]. However, the members of a subfamily often have similar substrate specificities and modes of regulation at a general level. ePKs are regulated by diverse means that include binding of an additional subunit, autophosphorylation and interaction with other domains within the polypeptide chain [10].

Using the first drafts of the human genome data [11], we report here on the repertoire of protein kinases in the human genome and their domain structures. We have classified the human kinases following Hanks and Hunter's [1,8,9] classification scheme and present them online at our kinases in genomes (KinG) website [12] and as Additional data files available with the online version of this paper. This list of kinases has been carefully arrived at after eliminating sequences that lack critical functional residues, in particular the ATP-binding glycine-rich motif and the catalytic aspartate [7]. These residues are indicated in the multiple sequence alignments of catalytic domains given in the additional data files. Fragments of protein kinases of less than 200 residues and lacking the functional residues have not been considered in the analysis. These include 24 gene products in the list with less than 200 residues that are probably pseudogenes. Hence, all the kinases used in the analysis are likely to be active kinase gene products and are unlikely to be pseudogenes. However, this paper discusses particularly those members of the kinase family with unusual domain combinations and those variants not studied so far, to the best of our knowledge, by experiment. We have compared the human kinase gene products with those encoded in the genomes of fruit fly [13], worm [14] and yeast [15]. A search in the human genome has also been carried out to identify the homologs of various subfamilies of bacterial protein kinases [16].

## Results and discussion

From our survey of the human genome using various sensitive family-profile search methods, the data available in

public databases, the presence of functional sequence motifs and manual analysis (see Materials and methods), we chose a dataset of 448 sequences for analysis of their domain structures and analysis of sequences not studied experimentally so far. We originally found 556 sequences of sufficient length and showing statistically significant E-values in the alignment with *bona fide* protein kinases. Although these sequences are expected to adopt the fold of a protein kinase, 108 of them lacked the glycine-rich motif or the catalytic base aspartate, or both, and hence do not have the essential requirements of a functional protein kinase (see Materials and methods for details). We have not considered these 108 sequences in the current analysis and confined it to 448 clear-cut cases. In addition, 11 lipid kinases have also been identified; these belong to the class of phosphoinositide-3 and phosphoinositide-4 kinases having structural similarity to the catalytic domain of protein kinases. The list of all the putative protein kinases used in this analysis, their classification into groups, multiple sequence alignment with indications of critical functional residues and the domain structures are given at [12] and in the Additional data files. The completed genomes of *Drosophila melanogaster*, *Caenorhabditis elegans* and *Saccharomyces cerevisiae* have also been surveyed and we used datasets containing 257, 474 and 122 gene products, respectively, containing regions with significant sequence similarity to the catalytic kinase domain. The number of protein kinases in the genomes analyzed here varies slightly from those reported by other groups [11,13-15,17] and in the public databases. This disparity is likely to be due to the sensitivity of the search methods and frequent updating of the draft version of the human genome. In the current analysis, the protein kinase homologs have been detected using a variety of sensitive methods and also by manual inspection to exclude most of the kinase-like sequences lacking functionally critical residues and those gene products that are very short. Our present study is focused only on the clear cases that are unlikely to be altered as a result of further changes in the dataset. In particular, we have used a stringent criterion in our search procedures (see Materials and methods) to avoid false positives. We have also consulted the datasets of kinases extracted from websites such as Superfamily [18,19], InterPro [20] and kinase.com [21], as well as ensuring the presence of functional sequence motifs, in order to decide on the list of kinases that is here.

Most human protein kinases have been classified into major subfamilies proposed by Hanks and Hunter [1,8,9], as described in Materials and methods. The most highly populated subfamilies are CAMK, receptor/non-receptor tyrosine kinases, CMGC and AGC (see Materials and methods for abbreviations) containing 81, 94, 72 and 69 proteins respectively. The casein kinase subfamily has only 12 members in the human genome. The protein kinases not included in the above subfamilies belong to other subfamilies, which include the polo family, MEK, MEKK, PAK,

NimA, mixed-lineage kinase (MLK), activin-TGF $\beta$  receptors, Wee1-Mik1, Raf and kinases involved in translational control. A detailed analysis of the members of the classical kinase subfamilies will be published elsewhere. Recent studies have identified four novel subfamilies of eukaryotic protein kinase-like sequences in bacteria [16] and homologous members of these subfamilies have been identified in the human genome. They include three homologs from the RIO1 subfamily, and eight and four members from the PID261 and ABC1 subfamilies, respectively. No homologs have been identified for the AQ578 subfamily of the bacterial protein kinases.

### Protein kinases with uncommon combination of domains

The high degree of functional diversification among the protein kinases is made possible by their ability to interact with large numbers of cellular proteins. These interactions are mediated through additional subunits or domains of the kinase that are regulatory or act as protein-interaction modules. The functions of these non-catalytic domains thus suggest the biological roles of the kinase and the specificity of the proteins or other ligands that bind to their specific domains. We identified 89 single-domain kinases encoded in the human genome, which are probably regulated by separate subunits. Clearly, a majority (359 sequences - about 80%) of the human protein kinases contain at least one domain other than the catalytic kinase domain.

The current survey has revealed the domain combinations of some protein kinases in which the role of the associated domains in relation to the overall biological function of the kinase is unclear. Figure 1 shows the sequence of domains of protein kinases with uncommon domain combinations. MARK-4 (ENSP246363; this id can be related to the corresponding Ensembl id by inserting five zeros after ENSP; a table showing the gi codes that correspond to these ids is available as an additional data file), whose catalytic domain has highest similarity to the CAMK subfamily, is associated with UBA (ubiquitin-associated domain) and KA1 (kinase-associated domain) (Figure 1a). MARK-4 is involved in mitogenic signaling and is crucial to microtubule disruption [22]. The presence of the UBA suggests a role for this kinase in integrating mitosis with the ubiquitination pathway, which is needed to break down the microtubules, by means of its association with ubiquitin through its UBA domain. The NF $\kappa$ B inhibitor kinase (IKK) (ENSP220806) [23,24] has a key role in phosphorylating the inhibitor of the transcription factor NF $\kappa$ B, I $\kappa$ B, leading to its ubiquitination and degradation, the subsequent activation of NF $\kappa$ B and enhancement of transcription. The presence of a ubiquitin domain (Ubiq-b) at the carboxyl terminus of IKK (Figure 1b) suggests that degradation of I $\kappa$ B is mediated through the ubiquitin domain directly, or by its association with other ubiquitins. The human genome encodes two additional homologs of IKK. A non-receptor tyrosine kinase (ENSP081325) [25]

associated with Cdc42 (a GTPase) contains an SH3 domain, which mediates protein-protein interactions through the polyproline conformation, and a UBA domain (Figure 1f). MARK-4, Cdc42-associated protein kinase and IKK therefore suggest an active involvement of protein kinases in the ubiquitin pathways.

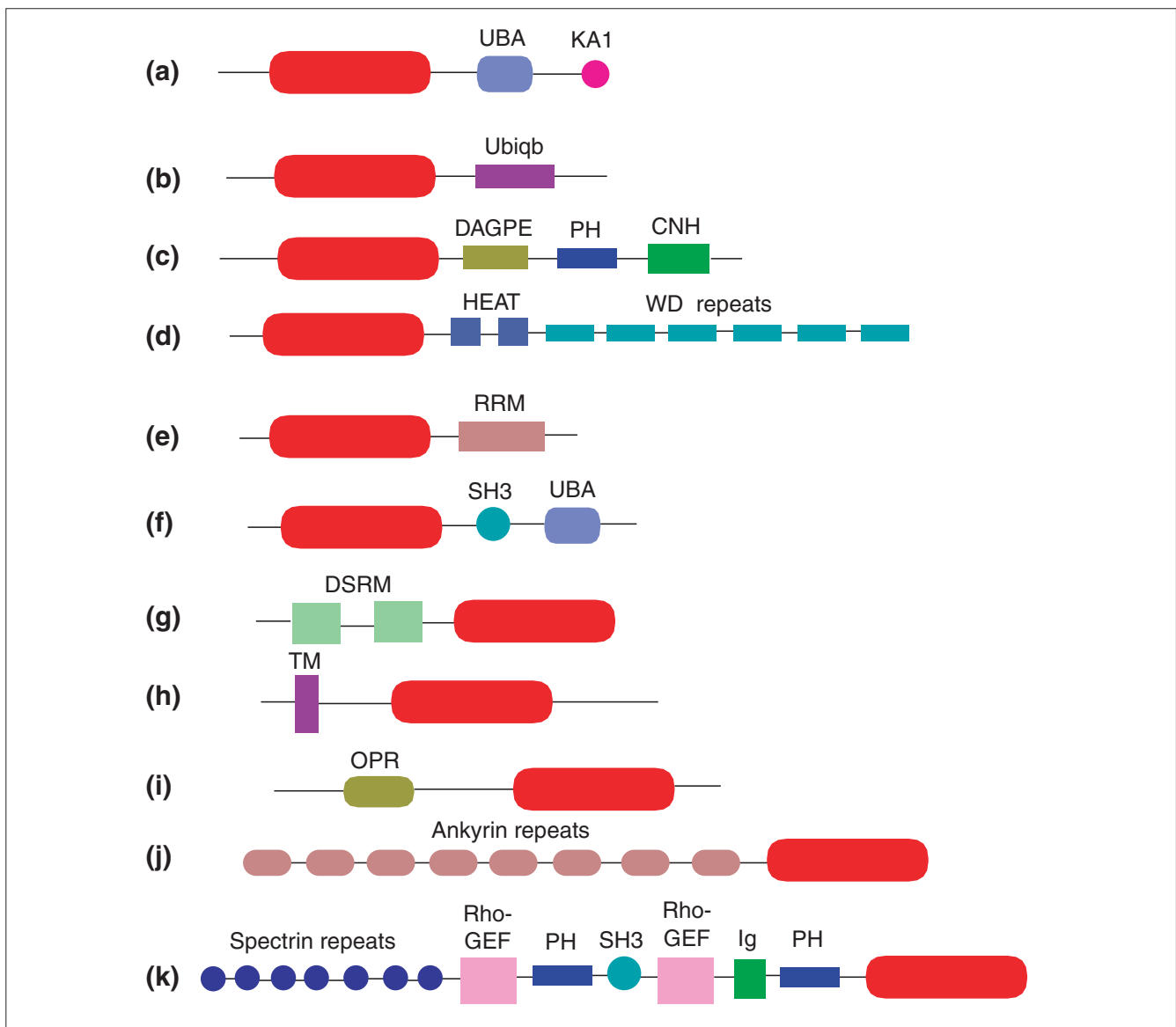
A protein kinase (ENSP232797) with two HEAT and six WD repeats following the catalytic domain has been identified (Figure 1d). The HEAT motifs are known to form  $\alpha$ -helical repeats, resulting in suprahelical structures that serve as protein-recognition interfaces [26] as understood for the PR65 subunit of phosphatase 2A [27]. The domain structure and the sequence pattern of the catalytic domain of this protein kinase are distinct from those of known kinase subgroups, suggesting a distinct subgroup.

The protein kinases involved in the regulation of protein translation are activated in various ways. In addition to the well-known members of this class, namely eIF2 $\alpha$  kinase (ENSP242081) [28] and protein kinase-R (PKR) (ENSP233057; Figure 1g) [29], the current study revealed the presence of a protein kinase with a single-strand RNA-binding motif (RRM) (Figure 1e). This kinase (ENSP235784) is known to be associated with the stathmin gene implicated in leukemia [30]. The presence of RRM suggests that RNA binding could be influential for its activity and suggests a probable role in the phosphorylation of proteins controlling translation or RNA-associated proteins.

A DMPK (myotonic dystrophy kinase)-like protein kinase (ENSP216542; Figure 1c) [31] has a phorbol-ester binding (DAG\_PE) domain followed by a pleckstrin homology (PH) domain and a CNH domain. The catalytic domain has high similarity to the kinases of the AGC group. The presence of the PH domain may suggest an association with the membrane. The PH domain could also mediate interaction with small G proteins, as DMPK is known to interact with Cdc42, a GTPase proximal to the membrane [31]. Hence, the domain combination of this kinase is consistent with a regulatory mechanism mediated by the Cdc42 GTPase and its action downstream in the small G-protein-mediated pathways.

Another kinase encoded in the human genome (gi13652062) shows similarity to the mitogen-activated protein kinase kinase (MAPKK) catalytic domain and contains an octicosapeptide repeat (OPR) domain which is known to occur in an isoform of protein kinase C (PKC) [32]. This divalent cation-binding domain suggests an influence of divalent cations in mitogenic signaling.

The Trio kinase (triple-function domain) (ENSP252004) [33] belonging to the CAMK group is unique in having two guanine-nucleotide-exchange factor (GEF) domains, one specific for Rho and the other for Rac (Figure 1k). As the amino-terminal GEF is involved in the activation of Jun

**Figure 1**

Schematic representation of uncommon domain combinations in protein kinases encoded in the human genome. **(a)** MARK-4 (ENSP246363); **(b)** IKK (ENSP220806); **(c)** DMPK (myotonic dystrophy kinase)-like protein kinase (ENSP216542); **(d)** function unknown (ENSP232797); **(e)** stathmin-associated kinase (ENSP235784); **(f)** non-receptor tyrosine kinase associated with Cdc42 (ENSP081325); **(g)** protein kinase-R (ENSP233057); **(h)** function unknown (ENSP251707); **(i)** MAP kinase kinase (gi3652062); **(j)** protein kinase, function unknown (ENSP234972); **(k)** Trio kinase (ENSP252004). Protein kinase domains are shown in red and the other domains are labeled as follows: CNH, Citron Nik I-like kinase homology domain; DAGPE, diacylglycerol/phorbol ester-binding domain; DSRM, double-stranded RNA-binding motif; GEF, guanine-nucleotide exchange factor; HEAT, a conserved domain in huntingtin, elongation factor 3, a-regulatory subunit of PP2A and Tor1 protein; Ig, immunoglobulin domain; KA1, kinase-associated domain; OPR, octicosapeptide repeat; PH, pleckstrin homology domain; RRM, RNA recognition motif; SH3, Src homology 3 domain; TM, transmembrane region; UBA, ubiquitin-associated domain; Ubiqb, ubiquitin domain. In (k) the amino- and carboxy-terminal RhoGEF domains are known to bind to Rho and Rac GTPases, respectively.

kinases and in the production of membrane ruffles [33], it appears that the kinases upstream to Jun kinases could be the probable targets of the Trio kinase. Another member of the CAMK group (gi13651132; not shown in Figure 1) contains SH3, PDZ and guanylate kinase domains. The PDZ domain may mediate the membrane-association property of this kinase. Whereas the SH3 domain might mediate inter-

protein or interdomain interactions, it is intriguing to note the existence of a potential catalytic guanylate kinase domain alongside a protein kinase domain.

A protein kinase of the CMGC group (ENSP234972) with eight ankyrin repeats at the amino terminus has the characteristic sequence of the C-helix of cyclin-dependent kinases (CDKs)

conserved in the catalytic domain (Figure 1j). Interaction of the CDK inhibitors belonging to the p19INK family occurs through the ankyrin repeats of the inhibitor [34], and the inhibitors exist as separate genes (not fused to the catalytic kinase domain as in ENSP234972). Hence, the role of ankyrin domains in this member of the CDK-like kinase subfamily of the CMGC group could be intramolecular regulation.

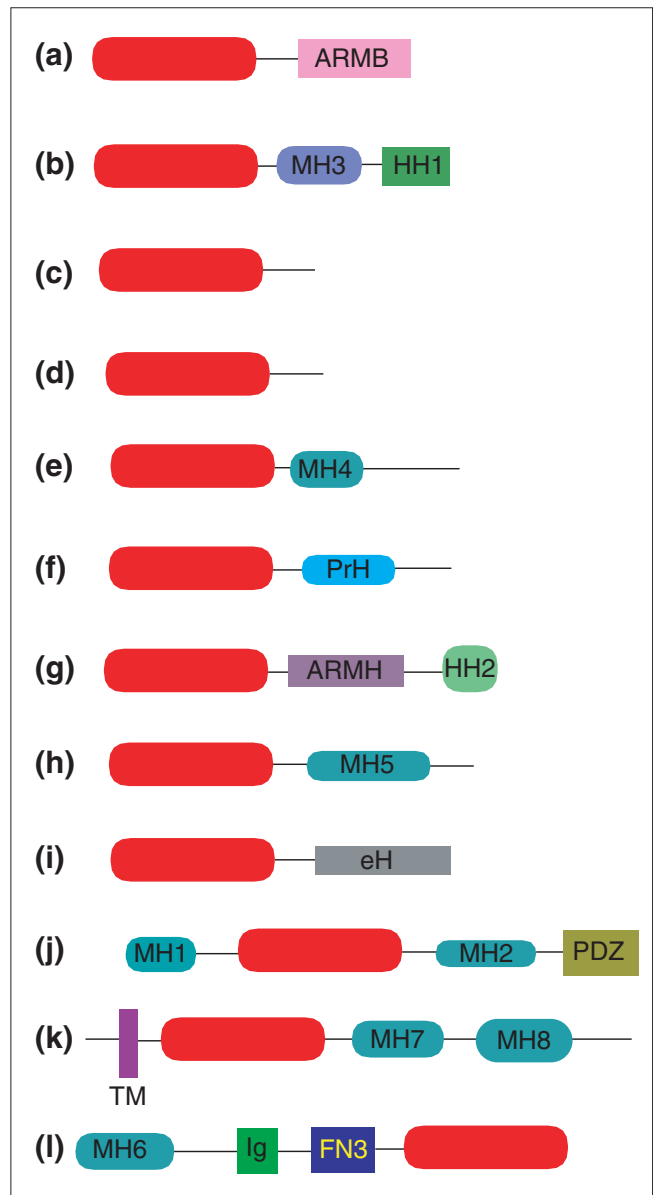
Also identified were CDK-7-related kinase (ENSP234626) with an insertion of 173 residues in the activation loop that continues as the subdomain X, following which is another insertion of 113 residues that extends as the carboxy-terminal end of the catalytic domain. Serine/threonine kinase-23 of the CMGC group has an insertion of 134 residues between the catalytic loop and activation loop and a second insertion occurs following the subdomain X, which is 86 residues long. An unannotated kinase gene product corresponding to the TREMBL entry tr|CAC39299| has an insertion of 144 residues between the catalytic loop and the activation loop. The dual-phosphorylation-regulated kinase homolog (ENSP247840) also has an insertion between the X subdomain and the carboxy-terminal end of the catalytic domain. These insertions could constitute a separate domain and probably confer different substrate specificity to the kinases, as they are located in the carboxy-terminal lobe.

### Hypothetical protein kinases not studied experimentally so far

Human protein kinases have been extensively studied for a long time. However, the functions and mode of regulation of several of them are still unclear. Our analysis has added to the list of protein kinases whose functions are not yet well understood. In almost all these cases, closely similar homologs of these domains are found in other eukaryotes. We have identified at least 12 human gene products with a clear similarity to protein kinases, but to the best of our knowledge these genes are unexplored by experimental analysis. A list of these kinases with their domain assignments is shown in Figure 2. Two (ENSP250481 and GenBank gi 13652558) contain only a single functional (catalytic) domain, (Figure 2c,d). The rest have long amino-terminal and carboxy-terminal extensions. The catalytic domains in most of these putative protein products could be associated to a known group (subfamily) of protein kinases because of a high similarity of the catalytic regions.

Two protein kinases (ENSP207880 and ENSP188359) (Figure 2a,g) show repeat motifs of the ARM repeats family. However, the catalytic domains of one of them (Figure 2g) are similar to the AGC subfamily, whereas the other kinase does not seem to belong to any of the known subfamilies. The ARM repeat is evolutionarily related to the HEAT repeat, and they usually act as scaffold modules enabling interaction with other proteins [35].

A protein kinase (ENSP251197; Figure 2j) with a PDZ domain and regions homologous to mouse kinases (MH1



**Figure 2**

Schematic representation of the domain assignments to hypothetical protein kinases encoded in the human genome. Protein kinase domains are shown in red. The other domains are labeled as follows: ARMB, ARM repeats of  $\beta$ -catenin-like proteins; ARMH, HEAT-related ARM-repeats; eH, homologous regions found in proteins of eukaryotes (*C. elegans*, *D. melanogaster*, *Arabidopsis thaliana*); HH, homologous region found in human proteins; MH, homologous regions found in proteins of mammals (mouse and human); PDZ, a conserved domain in PSD95, DLG and ZO-1; PrH, homologous regions found in proteins of primates (*Macaca* species and human); TM, transmembrane region. **(a)** function unknown (ENSP207880); **(b)** function unknown (ENSP252105); **(c)** function unknown (gi13652558; homologs have been found only in eukaryotes); **(d)** function unknown (ENSP250481; homologs have been found only in mouse); **(e)** function unknown (ENSP234334); **(f)** function unknown (ENSP232788); **(g)** function unknown (ENSP188359); **(h)** predicted BMP2-inducible kinase (ENSP251485); **(i)** protein kinase involved in embryonic development (ENSP250887); **(j)** protein kinase associated with the dystrophy complex (ENSP251197); **(k)** protein kinase similar to mouse apoptosis-associated kinase (ENSP251707); **(l)** protein kinase of unknown function (ENSP244847).

and MH2) is currently known only from the human and mouse genomes. The kinase domain of this gene product has a sequence identity of 71% to the microtubule-associated testis-specific serine/threonine kinase (MAST-205) [36]. MAST-205 is known to be associated with microtubules that are part of contractile structures. The regions at the amino and carboxyl termini of the kinase are homologous with the regions flanking the kinase domain of the syntrophin-associated protein kinase implicated in muscular dystrophy [37]. Translocation of the proteins constituting the dystrophy complex is aided by the PDZ domains of syntrophins that interact with the syntrophin-associated kinase [38]. From the domain combination observed in ENSP251197, a role for this protein kinase in the phosphoregulation of protein components of the dystrophy complex cannot be precluded. Assignment of biological function by virtue of similar domain combinations is consistent with previous studies [39] where conservation of domain order in proteins of similar domain combinations is suggested as an important means of functional conservation among various multidomain proteins. However, definite implication of ENSP251197 in the regulation of proteins of the dystrophy complex will require experimental studies.

Also included in this list is the protein kinase ENSP250887 (Figure 2i), which is closely related to HrPOPK1 of *Halocynthia roretzi* and Lok1p protein kinase of *Drosophila*. Both these are implicated in the establishment of the embryonic axis [40,41], suggesting a probable involvement of this human kinase in embryonic development.

A tyrosine kinase (ENSP251707, Figure 2k) with a predicted transmembrane region and similarity to the mouse apoptosis-associated tyrosine kinase [42] is identified in the human genome, suggesting a probable role of this human tyrosine kinase in apoptosis. If a search of the sequence databases is made, a large number of different tyrosine kinases appear as hits, following the mouse apoptosis-associated tyrosine kinase as the top hit.

A homolog of the mouse BMP2-inducible protein kinase (ENSP251485, Figure 2h) has also been identified in the human genome. The BMP2-inducible kinase has a regulatory role in osteoblast differentiation [43], suggesting a similar role for the human kinase.

### Non-kinase domains associated with human protein kinases

There are 61 distinct non-kinase domains associated with the catalytic domains of the human protein kinases. A few

domains are shared by more than one protein kinase whose catalytic domains belong either to the same subfamily (such as the AGC group) or to different subfamilies. In the current study we have also investigated the extent of similarity in the non-kinase domains shared by protein kinases with catalytic domains from different subfamilies. The domains analyzed include DAG\_PE, SH2, SH3 and PH.

The phylogenetic analysis of protein kinases and their associated domains was carried out as described in Materials and methods. Phylogenetic trees were obtained using CLUSTAL W [44]. The SH3 and SH2 domains are usually involved in the recruitment of proteins to their specific targets. In the case of protein tyrosine kinases, however, they also have regulatory roles. The SH2 domains of the non-receptor tyrosine kinases are known to translocate the corresponding kinases to specific receptors or other proteins associated with receptors, which activate them. So the specificity of these tyrosine kinases is also dependent on the adaptor domains that target them to specific proteins.

We have aligned the sequences of SH3, SH2 and kinase domains of human non-receptor tyrosine kinases separately. Dendrograms based on the sequence similarity of SH3, SH2 and catalytic kinase domains (Figure 3a-c) of the potential cytoplasmic tyrosine kinases encoded in the human genome show a similar clustering of the various gene products independently of the domains considered for the clustering. BMX kinase, a member of the Tec family, is a notable exception. The SH2- and the kinase-domain-based dendrograms suggest the inclusion of BMX kinase in the cluster of other Tec family members. However, BMX kinase is excluded from the Tec family cluster in the SH3-based tree, where it forms an outgroup (Figure 3c). This dissimilarity is also apparent from the sequence identity of the SH3 domains of various members of the Tec family with the BMX SH3 domain (< 18%), unlike the kinase domains (> 60%) and the SH2 domains (> 45%). It has also been observed that the Syk and ZAP kinases, each with two SH2 domains in addition to the kinase domain, show distinct clustering of the amino-terminal and the carboxy-terminal SH2 domains (Figure 3b). The amino-terminal SH2 forms a distinct cluster in the former, whereas the carboxy-terminal SH2 is grouped with the carboxy-terminal Src kinase and megakaryocyte-associated tyrosine kinase. The sequence identity between ZAP kinase and Syk amino-terminal SH2 domains (> 50%) and the carboxy-terminal SH2 domains (> 50%) is greater than the sequence identity of the amino- and carboxy-terminal SH2 domains (< 31%) in either of the kinases.

**Figure 3** (see figure on the next page)

Dendrograms based on the sequence similarity of various domains found in cytoplasmic tyrosine kinases. Members of different groups of tyrosine kinases are indicated in different colors. The consensus domain organization of kinases of the specific subfamily is indicated at the nodes of corresponding subfamilies. (a) Catalytic kinase domain; (b) SH2 domain; (c) SH3 domain. The bootstrap values obtained for 1,000 replicates are shown.

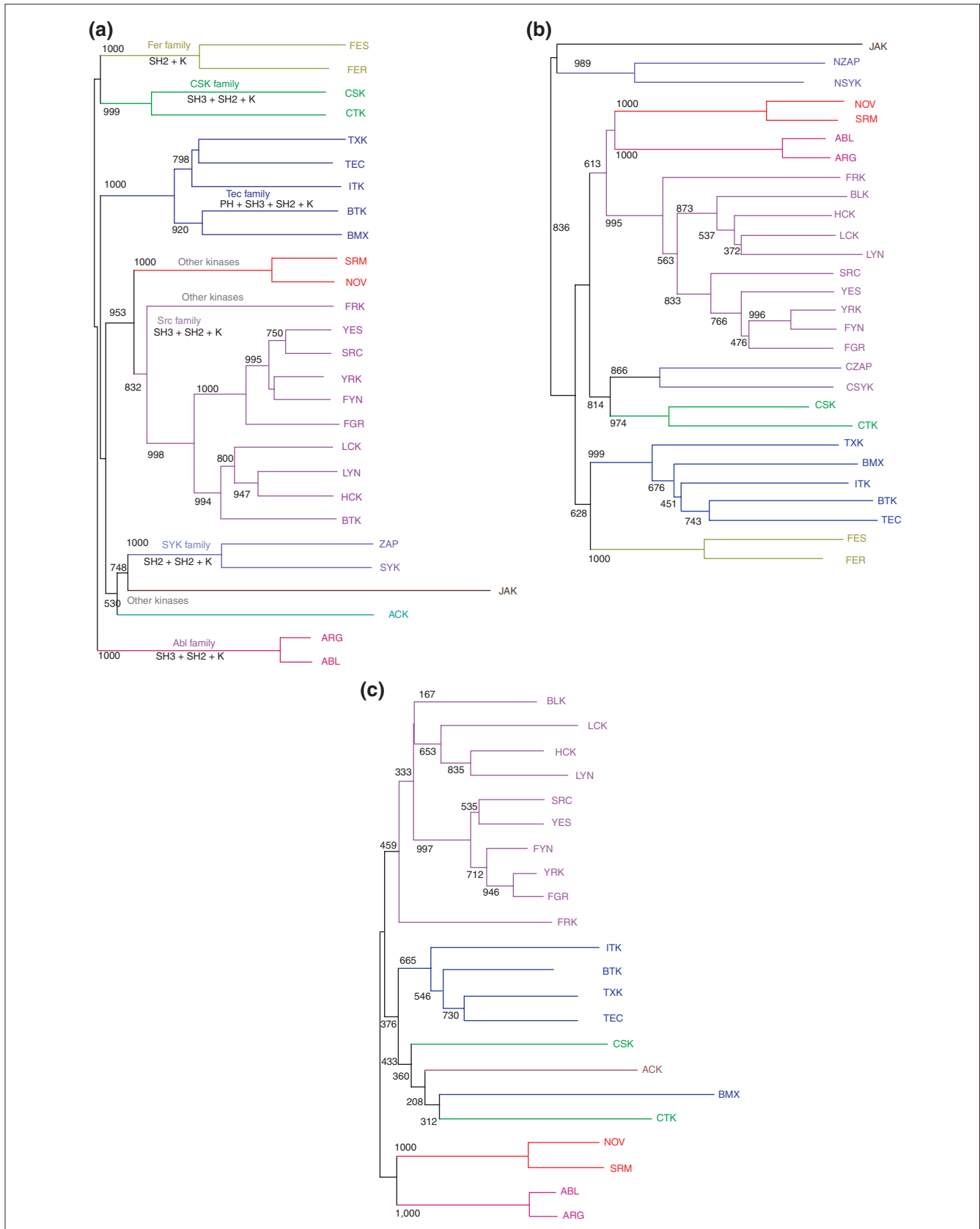


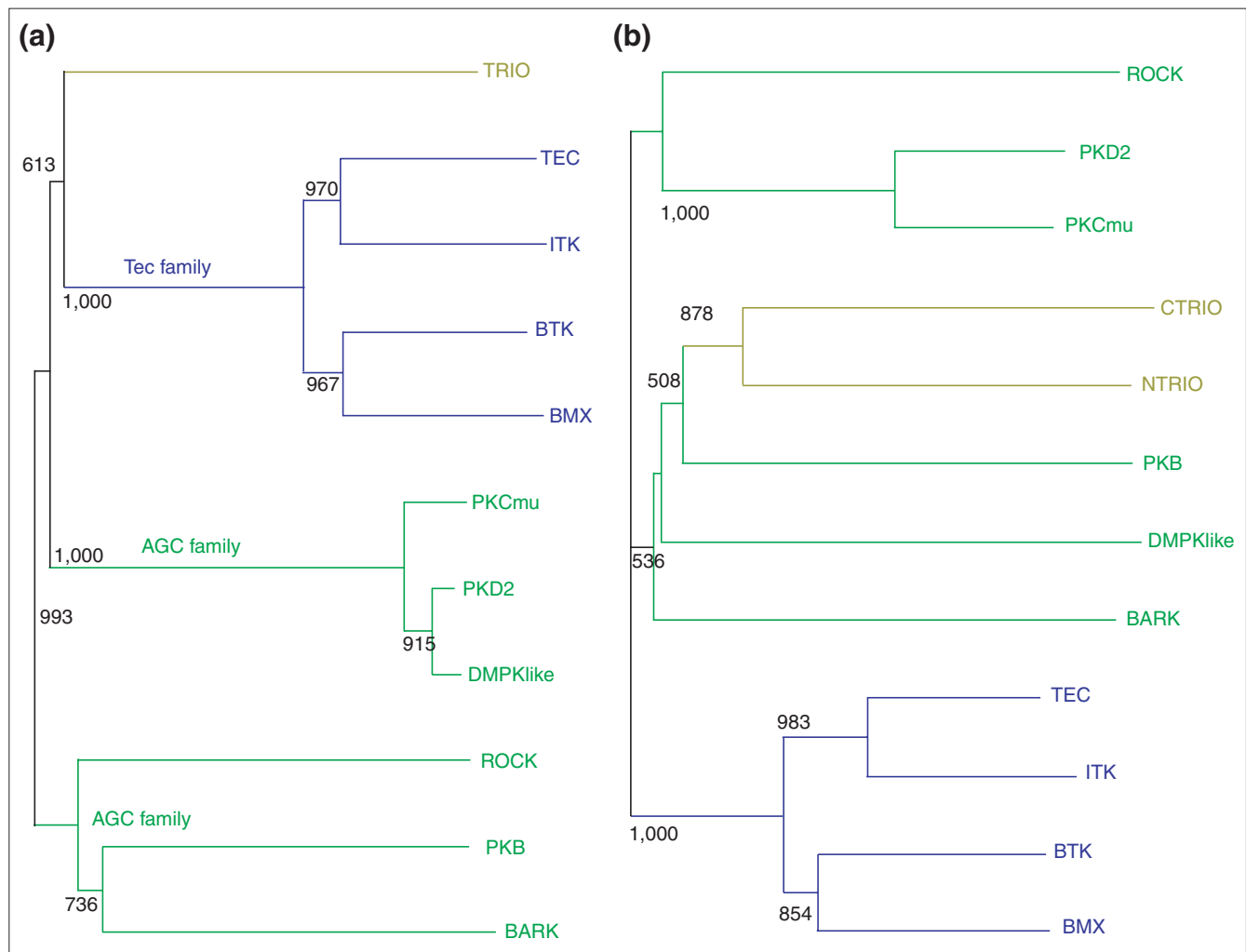
Figure 3 (see legend on the previous page)

In the protein kinases, diacylglycerol-binding (DAG\_PE) domains are restricted to the AGC and Raf subfamilies. The PKC isoforms  $\beta$ ,  $\gamma$ ,  $\theta$ ,  $\eta$ ,  $\epsilon$  and  $\mu$  have two DAG\_PE domains in tandem and amino-terminal to the kinase catalytic domain. The amino-terminal DAG\_PE domains of  $\beta$ ,  $\gamma$ ,  $\theta$ ,  $\eta$  and  $\epsilon$  isoforms form a distinct group with sequence identity greater than 68%, whereas the carboxy-terminal DAG\_PE domains of the  $\beta$ ,  $\epsilon$ ,  $\theta$ ,  $\eta$  and  $\gamma$  cluster together (65%) (dendrogram not shown). The PKC $\mu$  isoform is an exception where both the amino- and carboxy-terminal DAG\_PE domains cluster together (sequence identity 54%) with the carboxy-terminal DAG\_PE clusters.

The PH domains show variation in their biochemical properties. Some of the PH domains are known to bind to different proteins and certain PH domains can bind small-molecule ligands such as phospholipids. The protein kinases with PH

domains include the members of the Tec (cytoplasmic tyrosine kinases) and AGC (serine/threonine kinase) families. In the case of PDK1 and PKB, two enzymes involved in the phosphoinositide kinase pathways, the PH domains are involved in the translocation of the enzymes to the membrane, resulting in enzyme activation [45]. As already well documented, the corresponding catalytic domains of the two families form distinct clusters. The dendrogram based on the PH domains varies significantly from that based on the catalytic domains (Figure 4a,b). The PH domains of the PKC $\mu$  and PKD2 also form a distinct cluster, in contrast to the clustering of their kinase domains with the other AGC members.

These comparisons of the kinases and their associated domains suggest the probable mode of evolution of the multidomain kinases. In general, the grouping of kinases by the extent of the similarity in their catalytic domains is similar to



**Figure 4**

Dendrograms based on sequence similarity of domains in kinases belonging to Raf, AGC (serine/threonine kinase group) and Tec families (cytoplasmic tyrosine kinases) of protein kinases. **(a)** Catalytic kinase domain; **(b)** PH domain. Bootstrap numbers are shown on the nodes.



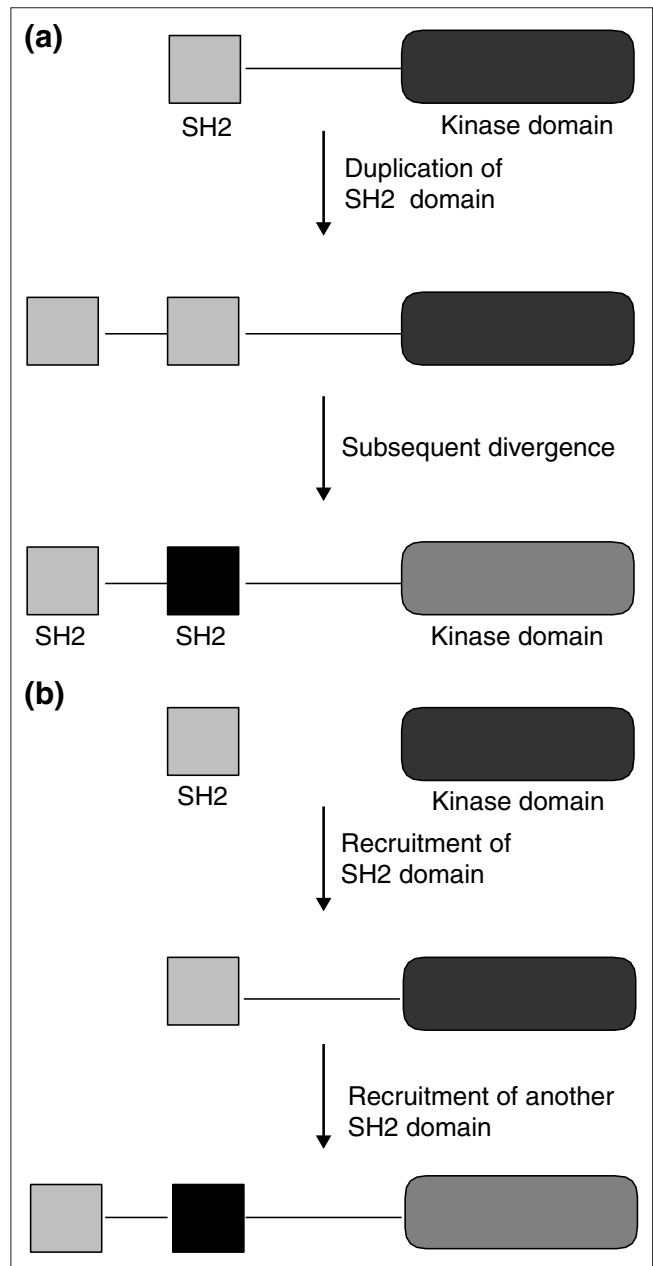
the clustering according to their associated domains, implying a coevolution of the catalytic domains and associated domains. This is also consistent with the interactions of the associated domains with a specific group of ligands that include small molecules and other signaling proteins. However, there are some exceptions, as described above. The distinct clustering of the amino- and carboxy-terminal SH2 and DAG\_PE domains suggests that the protein kinases containing them could have evolved in two stages, in any of the two possible ways as shown in Figure 5. In the first case (Figure 5a) a kinase domain would have evolved with one of the two repeat domains. Subsequent duplication of the SH2 or the DAG\_PE domain followed by divergence of the kinase-containing protein would have resulted in a kinase with two dissimilar SH2 or DAG\_PE domains. Alternatively, the domain combination would have occurred by independent recruitment of two distinct SH2 or DAG-PE domains (Figure 5b), resulting in the final gene product. Such diverse but homologous domains in a single gene product could enable distinct specificity and affinity for cognate ligands.

The occurrence of significantly diverse homologous domains in the same protein kinase, as evident from their distinct clustering, could also suggest the involvement of such kinases in the cross-talk between different signaling pathways. Such domain diversity could imply an involvement of such kinases in more than one pathway, serving as upstream and downstream effectors for more than one signaling cascade.

**Human protein kinases with no detectable closely related proteins encoded in the fruit fly and worm genome databases**

In view of the gene duplications across various genomes there is a possibility of expansion of certain subfamilies of the protein kinases, in one genome compared to another genome. This section discusses only the extreme cases where no homologs with significant similarity in sequence as well as domain composition to the protein kinase encoded in the human genome could be identified in the genomes of two lower multicellular eukaryotes of complete genome data. Such an analysis of the human protein kinases has enabled us to identify protein kinases that are potentially specific to vertebrates. We have compared the human protein kinases with those of two other multicellular organisms, *D. melanogaster* and *C. elegans*.

As protein tyrosine kinases are involved in sophisticated functions such as cellular differentiation and proliferation, one would expect the complexity of the signaling in higher organisms could be attributed partly to them. The apoptosis-associated tyrosine kinase (AAYK) [42], Fl-cytokine receptor kinase, c-Fms [46] and mast/stem-cell growth factor receptor kinase [47] identified in the human genome could not be detected in the worm and fruit fly genomes. The closest homologs of AAYK are neurospecific receptor kinase (NRK)



**Figure 5** Schematic representation of the possible mechanisms of evolution of multidomain protein kinases with more than one homologous but distinct non-kinase domain, as exemplified in the case of the Syk family of cytoplasmic tyrosine kinases. (a) Duplication of the SH2 domain followed by distinct functional specialization of duplicated domains; (b) Recruitment of two homologous but distinct SH2 domains by domain fusion.

with Fz and kringle domains in fly (30% identity with the kinase domain) and an open reading frame (ORF) CE00747 with laminin, epidermal growth factor (EGF) and FN3 domains in worm (51% identity with the kinase domain). However, no distinct functional domains other than the catalytic kinase domain could be identified in AAYK.

FL-cytokine receptor also seems to lack a functional counterpart in the fly and worm genomes, although the closest homologs are platelet-dependent growth factor (PDGF) receptor and vascular endothelium growth factor (VEGF) receptor-related kinase (PVR) and fibroblast growth factor (FGF) receptor family member ORF CE28238 in the fly and worm genomes, respectively. However the domain organization of PVR and CE28238 are distinct from that found in the FL-cytokine receptors. A functional homolog of the c-Fms and the mast/stem-cell growth factor receptor kinase is apparently absent in the fly and the worm genomes. The closest sequence neighbors of these c-Fms and mast/stem-cell growth factor receptor kinases are the FGF receptor kinase and PVR, respectively, in the fly, whose domain structures are different from that of the c-Fms and the mast/stem-cell growth factor receptor kinase.

All the above-mentioned kinases are involved in highly specialized cytokine- and growth-factor-mediated signaling pathways, and a few are specific for hematopoietic cells (mast/stem-cell receptor kinase) and immune responses (macrophage-stimulated receptor kinase), which explains their absence from lower multicellular organisms.

Certain members of a group of kinases involved in the protein translation, such as heme-regulated initiation (HRI) factor kinase (gi11125768) [28] and interferon-induced dsRNA-dependent kinase (PKR) (ENSP233057) [48], are at present found only in the human genome among the completed and publicly available genomes. This is consistent with the interferon-induced signaling pathways and the heme-regulated pathways being specific for vertebrates, and hence the lack of their effector domains in lower organisms. The domain structures of the homologs of the HRI kinase and PKR in worm and fly, eukaryotic initiation factor kinase eIF2 $\alpha$ -kinases, are different from those found in higher organisms. The worm and fly have two repeats of bacterial PQQ domains preceding the catalytic kinase domain in the sequence. In the case of human HRI kinase, no functional domain can be assigned to this region, although heme binding to this region to influence its activity is suggested. The human dsRNA-dependent kinase has two consecutive dsRNA-binding motifs preceding the catalytic kinase domain in the sequence. This is also consistent with its activity being influenced by interferon, an antiviral agent. Binding of the viral RNA to the dsRNA-binding motifs of the kinase could be regulated by interferon, and subsequently trigger a signaling cascade designed to evade viral infection [49].

The possible absence of such kinases in lower organisms such as fly and worm implies the existence of an alternative mode of defense against viral infection. Hence, the occurrence of heme-binding as well as RNA-binding kinases in higher organisms suggests that the translation machinery of these organisms has been associated in more than one

way to the protein kinases, linking them to the various signaling cascades.

Another important class of kinases, the receptor-interacting protein kinases (RIPKs) (ENSP220751), associated with receptors triggering apoptosis [50] were identified in the human genome but could not be found in invertebrates. The closest homologs of the RIPKs in the worm genome are the Raf kinases, which lack the caspase-recruitment domain (CARD) found in RIPKs. MAPKKK (TAK1) in the fly genome is the closest homolog of the RIPKs, and also lacks a CARD domain. This suggests that the components involved in controlling the signaling pathways mediating cell death in humans could be different from those in lower organisms. RIPKs with a CARD domain in higher organisms must have evolved to provide additional regulation to the apoptotic signaling pathways, and could also act as a link between its substrates and other components of the apoptotic machinery.

These potentially unique kinases therefore may be responsible for fine-tuning of the most highly specialized signaling pathways in humans. They may hence lack counterparts in the lower organisms, or their functions are carried out by some subgroups of enzymes. It is of interest to search for these kinases in other higher eukaryotic organisms such as mouse as and when the data is made freely available.

## Conclusions

The functional diversification of the kinases in the human genome, to meet the requirements of complex network of signaling pathways, is quite apparent from the analysis. The functional repertoire seems to be further expanded by protein kinases of uncommon domain composition. Shuffling of signaling domains or modules among gene products containing the various catalytic kinase domains seems to be one of the obvious means of generating such diverse biological roles. Identification of protein kinases with uncommon domain combinations suggests the possible existence of alternative modes of regulation of a few protein kinases, which appear to be distinct from the previously known and well characterized regulatory strategies.

The functional promiscuity of the various modules involved in signal transduction enables diverse biological interactions to be achieved using a limited number of modules. The occurrence of combinations of these modules would narrow the range of functional diversity, however. Given the functional diversity associated with these modules, their co-occurrence with domains of distinct known function, investigated using a computational approach in the current study, enables us to predict the biological implications of such uncommon domain combinations in relation to the overall functions of protein kinases. Experimental verification of these predictions could enhance our understanding of the specific biological roles of these protein kinases.

## Materials and methods

The complete set of predicted protein sequences from the open reading frames (ORFs) of the human genome (build 24; 3 July 2001) was obtained from the National Center for Biotechnology Information (NCBI) [51] and from Ensembl (version 1.1.0; April 2001 [52]) databases. Using sensitive sequence-profile matching algorithms, we have searched for kinases in both these versions of the human genome as there are differences between them. An overwhelming majority of the recognized kinases are common to the NCBI and Ensembl versions, but we have identified distinct kinases by comparing the two lists. Genome data from *S. cerevisiae* and *D. melanogaster* have been obtained from [51] and *C. elegans* from the Sanger Centre, UK [53].

We have used multiple sensitive sequence search and analysis methods - PSI-BLAST [54], IMPALA [55] and HMMer, which matches hidden Markov models (HMMs) [56]. These programs have been previously benchmarked ([55,57,58] and N. Mhatre and N.S., unpublished work) and we have used a stringent cutoff for E-values (0.0005 in PSI-BLAST,  $10^{-8}$  in IMPALA and 0.1 for HMMer) for identifying sequences with significant similarity to protein kinases. A six-node multi-processor Linux cluster machine (built by CDC Linux, Inc.) and several Linux-driven PCs have been used in these searches, using stand-alone versions of the programs. The list of predicted human kinases used for further analysis has been arrived at after careful cross-referencing between the results of these methods as well as manual scrutiny for a variety of factors such as length of the kinase domains and presence of critical functional residues. Protein kinases from the human genome and other organisms extracted from databases such as SUPERFAMILY [18,19], InterPro [20] and kinase.com [21] have been compared and cross-referenced with our initial list of kinases manually, taking into consideration the presence of functional motifs (see below) and the lengths of putative kinase domains.

Domain assignment to the non-catalytic regions of the kinase-containing genes has been made using the HMM search methods by querying each of the kinase-containing sequences against the 3,071 protein family HMMs available in the Pfam database [59]. Transmembrane segments have been detected using TMHMM [60].

### Criteria for identification of sequences with protein kinase functional motifs

Among the various sequence motifs characteristics of protein kinases [1], the glycine-rich motif with the pattern GXGXXG and the catalytic base aspartate [10] are suggested to be critical for the accommodation of ATP and phosphorylation respectively [7]. These are essential properties of a protein kinase. After eliminating several sequences on the basis of the results of sensitive profile matching, cross-referencing with other databases and manual analysis, there are 556 sequences, but not all of them have the glycine-rich

motif and catalytic base. Although these 556 sequences showed statistically significant matches (E-values better than 0.0001 in PSI-BLAST) with the sequences of proteins shown experimentally to be protein kinases, and hence are likely to adopt the fold of a protein kinase, not all these sequences may function as protein kinases. We have used the glycine-rich motif and the presence of the catalytic base as essential criteria to identify likely functional kinases.

In protein kinases generally, apart from the glycine-rich motif a positively charged residue (most often a lysine residue equivalent to Lys 72 of cyclic-AMP-dependent protein kinase) is also relevant to the accommodation of ATP. However, the number of residues that separate the glycine-rich motif and the charged residue along the primary structure is highly variable among the protein kinases. More important, certain subfamilies of protein kinases, such as MEK-related WNK [61,62], do not have such a positively charged residue supporting ATP but, nevertheless, function as kinases. Many known kinases are not entirely consistent with the glycine-rich motif as well (see, for example, the PROSITE pattern [63] for protein kinases). For example, protein kinases such as those in the casein kinase 2 subset lack one of the three glycy residues in the consensus pattern. On the basis of these considerations, the following criteria have been arrived at and used for identifying subset of sequences that possess essential sequence motifs to function as protein kinases. First, at least two out of the three glycy residues in the glycine-rich motif should be present. Second, an aspartate residue should be located in the sequence at a position equivalent to the catalytic base of known protein kinases.

Of 556 sequences with statistically significant similarity to protein kinases, 53 are found to lack the aspartate catalytic base and hence were not considered for further analysis. Most of these sequences also lack the typical glycine-rich motif. Among the remaining 503 sequences, 55 lacked two or all of the three glycy residues in the characteristic protein kinase glycine-rich region. However, in these 55 sequences an aspartate residue aligns with the catalytic aspartate of known protein kinases. These sequences were also not included in the detailed analysis, leaving 448 sequences for further analysis. The number of protein kinase-like sequences identified in the current analysis is of the order of the numbers reported by others [11,64].

While the two above criteria are used here to identify putative protein kinases, it should be noted that families of eukaryotic lipid kinases [65], and bacterial antibiotic-phosphorylating enzymes [66] and lipopolysaccharide-phosphorylating enzymes [67], which are evolutionarily related to eukaryotic protein kinases and share the fold, lack two or all of the three glycy residues in the glycine-rich motif. Nevertheless, they can phosphorylate substrates, although the substrates are commonly not proteins or peptides. It will

therefore be worthwhile to explore experimentally whether those sequences identified in the current analysis with aspartate at the catalytic base position, but which lack two or all the three glycy residues, could function as kinases.

### Classification of human protein kinases into subfamilies

A multiple sequence alignment of a collection of related proteins can be transformed into a position-specific scoring matrix (PSSM) or profile to use for further searches. The extent of sequence similarity and of conserved residues reflected in the PSSM will determine the effectiveness of matching of a query sequence. A PSSM generated from closely related family members forming a subfamily could be more effective than an overall family PSSM in specifically identifying further members of the subfamily. Thus a PSSM can be generated by considering the family of all the related sequences, or several PSSMs can be generated by considering subsets of the sequences to form subfamily profiles.

We classified 448 protein kinases according to the scheme of Hanks and Hunter [1]. This classification is based on the kinase catalytic domain. The most closely related catalytic domains have been classified by Hanks and Hunter into subfamilies - AGC (including the protein kinase A, protein kinase G and protein kinase C families), CMGC (the cyclin-dependent kinase, MAP kinase, glycogen synthase kinase, casein kinase 2 families), CAMK (calcium/calmodulin-dependent kinase), RTK (receptor tyrosine kinases), and so on. A large number of eukaryotic protein kinases have been classified into Hanks and Hunter subfamilies, and the sets of classified kinases are available in the Protein Kinase Resource (PKR) [68,69]. We have used the multiple sequence alignment of protein kinases belonging to each of the subfamilies available in PKR to generate a profile of each subfamily that encodes the pattern of conservation and variation in amino-acid sequence characteristic of that group. These PSSMs form a library of profiles representing the signature sequence pattern of the different kinase subfamilies.

If a query sequence consisting of the catalytic domain of a putative protein kinase is searched against this database of profiles, it could be expected that the subfamily to which the query kinase belongs would appear as the best hit with the lowest E-value, with some other subfamilies appearing with less significance. If the E-value for the top hit is significant (E-value less than 0.0001) then the query sequence is said to cluster best with that subfamily and is considered to be a putative member of that subfamily. For example, if the catalytic domain of ENSP224103 is queried against the database of profiles, the AGC subfamily appears at the top of the list with an E-value of  $10^{-121}$ , followed by the CAMK subfamily (E-value  $3 \times 10^{-68}$ ) and the polo subfamily (E-value:  $10^{-58}$ ). This means that the sequence pattern of ENSP224103 matches most significantly with the AGC

group of kinases and it is thus considered to be a member of this subfamily. Every human kinase sequence has been scanned against the subfamily profiles to associate it with the appropriate subfamily.

The program IMPALA [55] has been used to match the kinase query with the subfamily PSSMs. Although IMPALA was designed originally to identify distant relatives of a family, as we have constrained every PSSM by the subfamily sequence patterns, the program helps us identify the putative subfamily to which the query kinase catalytic domain belongs.

After associating the catalytic kinase domains to a Hanks and Hunter subfamily, the non-kinase domains in these proteins were identified using HMM and profile-searching in the Pfam database, as described above. The SwissProt database was consulted extensively to derive functional information on the various kinase-associated domains. Protein kinases that have been implicated in various diseases have been analyzed using information from the SwissProt and literature databases.

### Phylogenetic analysis of kinase and associated domains

The pairwise sequence distance between two protein domains has been calculated based on the percent identities obtained from the multiple sequence alignment derived from MALIGN [70]. The following expression was used to calculate a sequence-based dissimilarity measure between two proteins:  $D = -100.0 \ln(\text{pids}/100)$ , where pids is the percent identity between a given pair of protein domains. CLUSTAL W [44] has been used to generate trees. The robustness of the nodes in the trees generated by neighbor-joining methods has been tested by the bootstrap algorithm available in CLUSTAL W. Confidence limits have been tested by bootstrapping the trees 1,000 times. Trees and bootstrap values for various branch orders of domains are described in the text and shown in Figures 3 and 4.

### Note added in proof

While this paper was in review, a related paper was published by Hunter and coworkers [71] that classifies and compares the human protein kinases with protein kinases encoded in other eukaryotic genomes.

### Additional data files

A list of kinases identified in the human genome with their classification into groups, the multiple sequence alignment of catalytic regions with critical functional residues indicated, the domain structures of kinase-containing gene products, and a table with Ensembl ids and the corresponding gi ids are available as additional data files with the online version of this article. This information is also freely available from our KinG website [12].

## Acknowledgements

We thank the anonymous referees for several important suggestions for the improvement of our work and paper. We thank S. Abhiman and K.R. Abhinandan for their help in setting-up the KinG website and additional data files. A.K. is supported by a fellowship from the Council of Scientific and Industrial Research, India. This research is supported by the award of an International Senior Fellowship in biomedical sciences to N.S. by the Wellcome Trust, UK and by the computational genomics project supported by the Department of Biotechnology, India.

## References

- Hanks SK, Quinn AM, Hunter T: **The protein kinase family: conserved features and deduced phylogeny of the catalytic domains.** *Science* 1988, **241**:42-52.
- Pawson T: **Introduction: protein kinases.** *FASEB J* 1994, **8**:1112-1113.
- Zhang Z, Yu X, Zhang Y, Geronimo B, Lovlie A, Fromm SH, Chen Y: **Targeted misexpression of constitutively active BMP receptor-IB causes bifurcation, duplication, and posterior transformation of digit in mouse limb.** *Dev Biol* 2000, **220**:154-167.
- Frost DO: **BDNF/trkB signaling in the developmental sculpting of visual connections.** *Prog Brain Res* 2001, **134**:35-49.
- Lee MH, Yang HY: **Negative regulators of cyclin-dependent kinases and their roles in cancers.** *Cell Mol Life Sci* 2001, **58**:1907-1922.
- Irby RB, Mao W, Coppola D, Kang J, Loubeau JM, Trudeau W, Karl R, Fujita DJ, Jove R, Yeatman TJ: **Activating SRC mutation in a subset of advanced human colon cancers.** *Nat Genet* 1999, **21**:187-190.
- Zheng J, Knighton DR, ten Eyck LF, Karlsson R, Xuong N, Taylor SS, Sowadski JM: **Crystal structure of the catalytic sub-unit of cAMP dependent protein kinase complexed with MgATP and peptide inhibitor.** *Biochemistry* 1993, **32**:2154-2161.
- Hunter T: **A thousand and one protein kinases.** *Cell* 1987, **50**:823-829.
- Hanks SK, Quinn AM: **Protein kinase catalytic domain sequence database: identification of conserved features of primary structure and classification of family members.** *Meth. Enzymol* 1991, **200**:38-42.
- Johnson LN, Noble ME, Owen DJ: **Active and inactive protein kinases: structural basis for regulation.** *Cell* 1996, **85**:149-158.
- International Human Genome Sequencing Consortium: **Initial sequencing and analysis of the human genome.** *Nature* 2001, **409**:860-921.
- Supplementary information**  
[<http://hodgkin.mbu.iisc.ernet.in/~king>]
- Morrison DK, Murakami MS, Cleghon V: **Protein kinases and phosphatases in the *Drosophila* genome.** *J Cell Biol* 2000, **150**:F57-F62.
- Plowman GD, Sudarsanam S, Bingham J, Whyte D, Hunter T: **The protein kinases of *Caenorhabditis elegans*: A model for signal transduction in multicellular organisms.** *Proc Natl Acad Sci USA* 1999, **96**:13603-13610.
- Hunter T, Plowman GD: **The protein kinases of budding yeast: six score and more.** *Trends Biochem Sci* 1997, **22**:18-22.
- Leonard CJ, Aravind L, Koonin, EV: **Novel families of putative protein kinases in bacteria and archaea: evolution of the "eukaryotic" protein kinase superfamily.** *Genome Res* 1998, **8**:1038-1047.
- Robinson DR, Wu YM, Lin SF: **The protein tyrosine kinase family of the human genome.** *Oncogene* 2000, **19**:5548-5557.
- Gough J, Karplus K, Hughey R, Chothia C: **Assignment of homology to genome sequences using a library of Hidden Markov Models that represent all proteins of known structure.** *J Mol Biol* 2001, **313**:903-919.
- Superfamily** [<http://supfam.mrc-lmb.cam.ac.uk/SUPERFAMILY>]
- Aweiler R, Attwood TK, Bairoch A, Bateman A, Birney E, Biswas M, Bucher P, Cerutti L, Corpet F, Croning MDR, et al.: **The InterPro database, an integrated documentation resource for protein families, domains and functional sites.** *Nucleic Acids Res* 2001, **29**:37-40.
- kinase.com** [<http://www.kinase.com>]
- Drewes G, Ebnet A, Preuss U, Mandelkow EM, Mandelkow E: **MARK, a novel family of protein kinases that phosphorylate microtubule-associated proteins and trigger microtubule disruption.** *Cell* 1997, **89**:297-308.
- Shimada T, Kawai T, Takeda K, Matsumoto M, Inoue J, Tatsumi Y, Kanamaru A, Akira S: **IKK-i, a novel lipopolysaccharide-inducible kinase that is related to I $\kappa$ B kinases.** *Int Immunol* 1999, **11**:1357-1362.
- Li Q, Estepa G, Memet S, Israel A, Verma IM: **Complete lack of NF- $\kappa$ B activity in IKK1 and IKK2 double deficient mice: additional defect in neurulation.** *Genes Dev* 2000, **14**:1729-1733.
- Yang W, Cerione RA: **Cloning and characterization of a novel Cdc42-associated tyrosine kinase, ACK-2, from bovine brain.** *J Biol Chem* 1997, **272**:24819-24824.
- Groves MR, Barford D: **Topological characteristics of helical repeat proteins.** *Curr Opin Struct Biol* 1999, **9**:383-389.
- Groves MR, Hanlon N, Turowski P, Hemmings BA, Barford D: **The structure of the protein phosphatase 2A PR65/A subunit reveals the conformation of its 15 tandemly repeated HEAT motifs.** *Cell* 1999, **96**:99-110.
- Chen JJ, Crosby JS, London IM: **Regulation of heme-regulated eIF-2 alpha kinase and its expression in erythroid cells.** *Biochimie* 1994, **76**:761-769.
- Thomis, DC, Doohan JP, Samuel CE: **Mechanism of interferon action: cDNA structure, expression, and regulation of the interferon-induced, RNA-dependent P1/eIF-2 alpha protein kinase from human cells.** *Virology* 1992, **188**:33-46.
- Alam MR, Caldwell BD, Johnson RC, Darlington DN, Mains RE, Eipper BA: **Novel proteins that interact with the COOH-terminal cytosolic routing determinants of an integral membrane peptide-processing enzyme.** *J Biol Chem* 1996, **271**:28636-28640.
- Leung T, Chen XQ, Tan I, Manser E, Lim L: **Myotonic dystrophy kinase related Cdc42-binding kinase acts as a Cdc42 effector in promoting cytoskeletal reorganization.** *Mol Cell Biol* 1998, **18**:130-140.
- Ponting CP: **Novel domains in NADPH oxidase subunits, sorting nexins, and PtdIns 3-kinases: binding partners of SH3 domains?** *Protein Sci* 1996, **5**:2353-2357.
- Seipel K, Medley QG, Kedersha NL, Zhang XA, O'Brien SP, Serrano C, Hemler ME, Streluli M: **Trio amino-terminal guanine nucleotide exchange factor domain expression promotes actin cytoskeleton reorganization, cell migration and anchorage-independent cell growth.** *J Cell Sci* 1999, **112**:1825-1834.
- Brotherton DH, Dhanaraj V, Wick S, Brizuela L, Domaille PJ, Volynik E, Xu X, Parisini E, Smith BO, Archer SJ, et al.: **Crystal structure of the complex of the cyclin D-dependent kinase Cdk6 bound to the cell-cycle inhibitor p19INK4d.** *Nature* 1998, **395**:244-250.
- Andrade MA, Petosa C, O'Donoghue SI, Muller CW, Bork P: **Comparison of ARM and HEAT protein repeats.** *J Mol Biol* 2001, **309**:1-18.
- Walden PD, Cowan NJ: **A novel 205-kilodalton testis-specific serine/threonine protein kinase associated with microtubules of the spermatid manchette.** *Mol Cell Biol* 1993, **12**:127625-127635.
- Lumeng C, Phelps S, Crawford GE, Walden PD, Barald K, Chamberlain JS: **Interactions between beta 2-syntrophin and a family of microtubule-associated serine/threonine kinases.** *Nat Neurosci* 1999, **2**:611-617.
- Miyagoe-Suzuki Y, Takeda SI: **Association of neuronal nitric oxide synthase (nNOS) with alpha1-syntrophin at the sarcolemma.** *Microsc Res Tech* 2001, **55**:164-170.
- Bashton M, Chothia C: **The geometry of domain combination in proteins.** *J Mol Biol* 2002, **315**:927-939.
- Sasakura Y, Ogasawara M, Makabe KW: **Maternally localized RNA encoding a serine/threonine protein kinase in the ascidian, *Halocynthia roretzi*.** *Mech Dev* 1998, **76**:161-163.
- Oishi I, Sugiyama S, Otani H, Yamamura H, Nishida Y, Minami Y: **A novel *Drosophila* nuclear protein serine/threonine kinase expressed in the germline during its establishment.** *Mech Dev* 1998, **71**:49-63.
- Gaozza E, Baker SJ, Vora RK, Reddy EP: **A novel tyrosine kinase induced during growth arrest and apoptosis of myeloid cells.** *Oncogene* 1997, **15**:3127-3135.

43. Kearns AE, Donohue MM, Sanya LB, Demay MB: **Cloning and characterization of a novel protein kinase that impairs osteoblast differentiation in vitro.** *J Biol Chem* 2001, **276**:42213-42218.
44. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice.** *Nucleic Acids Res* 1994, **22**:4673-4680.
45. Filippa N, Sable CL, Hemmings BA, Obberghen EV: **Effect of phosphoinositide-dependent kinase I on protein kinase B translocation and its subsequent activation.** *Mol Cell Biol* 2000, **20**:5712-5721.
46. Yue X, Favot P, Dunn TL, Cassady AI, Hume DA: **Expression of mRNA encoding the macrophage colony-stimulating factor receptor (c-fms) is controlled by a constitutive promoter and tissue-specific transcription elongation.** *Mol Cell Biol* 1993, **13**:3191-3201.
47. Nocka K, Buck J, Levi E, Besmer P: **Candidate ligand for the c-kit transmembrane kinase receptor: KL, a fibroblast derived growth factor stimulates mast cells and erythroid progenitors.** *EMBO J* 1990, **9**:3287-3294.
48. Xu Z, Williams BR: **Genomic features of human PKR: alternative splicing and a polymorphic CGG repeat in the 5'-untranslated region.** *J Interferon Cytokine Res* 1998, **18**:609-616.
49. Clemens MJ, Elia A: **The dsRNA-dependent protein kinase PKR: structure and function.** *J Interferon Cytokine Res* 1997, **17**:503-524.
50. Stanger BZ, Leder P, Lee TH, Kim E, Seed B: **RIP: a novel protein containing a death domain that interacts with Fas/APO-1 (CD95) in yeast and causes cell death.** *Cell* 1995, **81**:513-523.
51. **National Center for Biotechnology Information: genomes** [<ftp://ftp.ncbi.nlm.nih.gov/genomes>]
52. **Ensembl: pep** [[ftp://ftp.ensembl.org/pub/current\\_human/data/fasta/pep](ftp://ftp.ensembl.org/pub/current_human/data/fasta/pep)]
53. **The C. elegans Protein Database: Wormpep** [[http://www.sanger.ac.uk/Projects/C\\_elegans/wormpep/](http://www.sanger.ac.uk/Projects/C_elegans/wormpep/)]
54. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25**:3389-3402.
55. Schaffer AA, Wolf YI, Ponting CP, Koonin EV, Aravind L, Altschul SF: **IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position specific score matrices.** *Bioinformatics* 1999, **15**:1000-1011.
56. Eddy SR: **Profile hidden Markov models.** *Bioinformatics* 1998, **14**:755-763.
57. Muller A, MacCallum RM, Sternberg MJ: **Benchmarking PSI-BLAST in genome annotation.** *J Mol Biol* 1999, **293**:1257-1271.
58. Lindahl E, Elofsson A: **Identification of related proteins on family, superfamily and fold level.** *J Mol Biol* 2000, **295**: 613-625.
59. Bateman A, Birney E, Durbin R, Eddy SR, Howe KL, Sonnhammer EL: **The Pfam protein families database.** *Nucleic Acids Res* 2000, **28**:263-266.
60. Krogh A, Larsson B, von Heijne G, Sonnhammer EL: **Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes.** *J Mol Biol* 2001, **305**:567-580.
61. Xu B, English JM, Wilsbacher JL, Stippec S, Goldsmith EJ: **WNK1, a novel mammalian serine/threonine protein kinase lacking the catalytic lysine in subdomain II.** *J Biol Chem.* 2000, **275**:16795-16801.
62. Verissimo F, Jordan P: **WNK kinases, a novel protein kinase subfamily in multi-cellular organisms.** *Oncogene* 2001, **20**:5562-5569.
63. Falquet L, Pagni M, Bucher P, Hulo N, Sigrist CJ, Hofmann K, Bairoch A: **The PROSITE database, its status in 2002.** *Nucleic Acids Res* 2002, **30**:235-238.
64. Kostich M, English J, Madison V, Gheys F, Wang L, Qiu P, Greene J, Laz TM: **Human members of the eukaryotic protein kinase family.** *Genome Biol* 2002, **3**:research0043.1-0043.12.
65. Walker EH, Perisic O, Ried C, Stephens L, Williams RL: **Structural insights into phosphoinositide 3-kinase catalysis and signalling.** *Nature* 1999, **402**:313-320.
66. Hon WC, McKay GA, Thompson PR, Sweet RM, Yang DS, Wright GD, Berghuis AM: **Structure of an enzyme required for aminoglycoside antibiotic resistance reveals homology to eukaryotic protein kinases.** *Cell* 1997, **89**:887-895.
67. Krupa A, Srinivasan N: **Lipopolysaccharide-phosphorylating enzymes encoded in the genomes of Gram-negative bacteria are related to the eukaryotic protein kinases.** *Protein Sci* 2002, **11**:1580-1584.
68. Smith CM, Shindyalov IN, Veretnik S, Gribskov M, Taylor SS, ten Eyck LS, Bourne PE: **The protein kinase resource.** *Trends Biochem Sci* 1997, **22**:444-446.
69. **The Protein Kinase Resource** [<http://pkr.sdsc.edu/html/index.shtml>]
70. Johnson MS, Overington JP, Blundell TL: **Alignment and searching for common folds using a database of structural templates.** *J Mol Biol* 1993, **231**:735-752.
71. Manning G, Plowman G, Hunter T, Sudarsanam S: **Evolution of protein kinase signaling from yeast to man.** *Trends Biochem Sci* 2002, **27**:514-520.