Research

Human members of the eukaryotic protein kinase family

Mitch Kostich*, Jessie English†, Vincent Madison‡, Ferdous Gheyas§, Luquan Wang*, Ping Qiu*, Jonathan Greene* and Thomas M Laz*

Addresses: *Discovery Technology, †Tumor Biology, ‡Structural Chemistry, and §Biostatistics, Schering-Plough Research Institute, Kenilworth, NJ 07033, USA.

Correspondence: Mitch Kostich. E-mail: mitchell.kostich@spcorp.com

Published: 22 August 2002

Genome ***Biology*** 2002, **3(9)**:research0043.1–0043.12

The electronic version of this article is the complete one and can be found online at http://genomebiology.com/2002/3/9/research/0043

(Print ISSN 1465-6906; Online ISSN 1465-6914)

Received: 6 June 2002
Revised: 4 July 2002
Accepted: 10 July 2002

Abstract

Background: Eukaryotic protein kinases (EPKs) constitute one of the largest recognized protein families represented in the human genome. EPKs, which are similar to each other in sequence, structure and biochemical properties, are important players in virtually every signaling pathway involved in normal development and disease. Near completion of projects to sequence the human genome and transcriptome provide an opportunity to identify and perform sequence analysis on a nearly complete set of human EPKs.

Results: Publicly available genetic sequence data were searched for human sequences that potentially represent EPK family members. After removal of duplicates, splice variants and pseudogenes, this search yielded 510 sequences with recognizable similarity to the EPK family. Protein sequences of putative EPK catalytic domains identified in the search were aligned, and a phenogram was constructed based on the alignment. Representative sequence records in GenBank were identified, and derived information about gene mapping and nomenclature was summarized.

Conclusions: This work represents a nearly comprehensive census and early bioinformatics overview of the EPKs encoded in the human genome. Evaluation of the sequence relationships between these proteins contributes contextual information that enhances understanding of individual family members. This curation of human EPK sequences provides tools and a framework for the further characterization of this important class of enzymes.

Background

The eukaryotic protein kinase (EPK) family is one of the largest protein families represented in the human genome. The human genome has been estimated to contain between 500 and 1,000 EPK genes [1,2]. EPKs play key roles in many intercellular and intracellular signaling pathways by transducing, amplifying or integrating upstream signals [3-6]. Upstream signaling events modulate the activity of EPKs through a variety of means that often involve alterations in the phosphorylation of key EPK residues or changes in the physical association of regulatory proteins with the EPK. Signals are typically relayed downstream by the EPK through the covalent transfer of the terminal phosphate group from ATP or GTP to serine, threonine or tyrosine residues of substrate proteins [7]. Phosphorylation of the substrate protein then alters its ability to physically interact with other molecules in the cell [2,8].

The key feature that distinguishes EPK family members from other proteins is the sequence of a contiguous stretch

of approximately 250 amino acids that constitutes the catalytic domain [9-11]. Although no residue in this region is absolutely conserved in all family members, the presence of most of the signature EPK residues can be used to determine that a particular sequence belongs in the family. The pattern of residue conservation seen within this core of 250 amino acids is thought to be due to selective evolutionary pressure to preserve the major function of this gene family: catalysis of phosphate transfer from ATP to a protein substrate. The solution of crystal structures for several EPKs, some of which include bound ATP and protein substrate, has clarified the functional role of particular conserved residues in binding different portions of ATP and protein substrate molecules, and in regulating these binding events [12-16]. Interestingly, family members exist that no longer retain the characteristic catalytic activity, even though they retain most of the conserved sequence features of the kinase catalytic domain. In these latter cases, the role of the conserved residues in protein function is not known.

In addition to orthodox EPKs, there are several other proteins that have demonstrated protein kinase activity but share little or no recognizable sequence similarity with the EPK family. Examples include A6 kinases [17], a number of lipid kinase family members [18], aminoglycoside phosphotransferases [19], pyruvate dehydrogenase kinase family members [20], DNA-dependent protein kinase [21], ATM [22], ATR [23], BCR [24,25], a transient receptor potential channel [26] and actin-fragmin kinases [27]. Although some of these non-EPK protein kinases share a similar overall fold with each other and with orthodox EPKs [28], the low sequence similarity between these proteins and EPKs confounds attempts to align the sequences in a single alignment or to perform comparative sequence analysis.

Regulation of EPK function can occur at many levels, including control of synthesis, posttranslational modification, binding of regulatory proteins and subcellular localization. One frequently reported mechanism for regulating EPK activity involves phosphorylation of key residues of the EPK catalytic domain by other upstream EPKs [13,29-32]. *In vitro*, purified EPKs often display broad protein substrate specificity [7], and it is thought that in many cases *in vivo* substrate specificity is limited by a requirement for substrate to associate not only with its cognate EPK, but also with other components of an EPK-containing protein assembly [8,33,34]. These protein complexes can consist of signaling proteins involved in several parallel inputs or several consecutive steps in a signaling cascade [35]. EPKs frequently contain one or more non-catalytic domains, some of which are thought to serve for docking EPKs to various constituents of these complexes.

The presence of multiple potential protein-docking sites in some EPKs allows them to serve as scaffolding molecules around which a protein complex can assemble [4,5,8]. The assembly and activity of the complex can be controlled by altering the ability of these docking sites to be bound. Some well characterized, physiologically important EPKs, such as ErbB3 [36], seem to have lost their enzymatic activity altogether and are thought to function solely as scaffolding proteins.

Appreciation of the central role of EPKs in virtually every signaling pathway involved in normal development and disease [2] has stimulated much work on individual family members, as well as interesting subsets of the entire family. Progress toward sequencing the genome and transcripts of several organisms has allowed the identification of most of the EPK genes present in *Saccharomyces cerevisiae* [37], *Caenorhabditis elegans* [1,38], and *Drosophila melanogaster* [39]. A description of the tyrosine kinase subset of human EPKs has been published [40], and at least one partial list of human EPKs is available on the web [41]. In our study, this progress is extended to include almost all the EPK genes in the human genome. Publicly available sequence data were searched for sequences potentially encoding human EPK family members. For the purposes of this study, the family was defined to include proteins that share a particular previously described pattern of amino-acid sequence conservation [9-11]. As discussed above, this definition includes family members that lack protein kinase activity but may still have important cellular functions that are dependent on retention of a protein-kinase-like structure. Conversely, this definition excludes proteins that have protein kinase activity, but lack substantial sequence similarity to the rest of the family. A conservative mining approach was used to minimize over-counting because of inclusion of splice variants, pseudogenes and sequencing artifacts. In total, 510 known and novel human EPK loci were identified and cross-referenced to publicly available sequence records. Information on nomenclature and genetic mapping was extracted from the sequence records and summarized. The protein sequences of the EPK catalytic domains of the family members were aligned, and the alignment was used to construct a phenogram that illustrates the sequence relatedness between family members. This work represents a nearly comprehensive census and an early bioinformatics overview of this large gene family in *Homo sapiens*.

Results and discussion

Searching sequence databases

A previously published alignment of EPK catalytic domains [11], which is available on the web [42], includes approximately 300 sequences from a variety of organisms, along with links to the corresponding GenBank records. The protein sequences from this alignment were used as bait for BLAST [43] searches of GenBank [44] nucleotide and protein sequence datasets. Human and non-human sequences from the alignment were used as bait to reduce the probability of missing human members of subfamilies

that are poorly represented in humans. Hits identified using different bait sequences were consolidated, and duplicate records, including those representing splice and allelic variants, were removed (see Materials and methods).

Each hit was manually evaluated for the presence of conserved residues known to be distinctive for the EPK family [9-11], and approved human hits were added to the EPK collection. Non-human sequences were added to the collection only if the corresponding protein sequence was less than 50% identical to any other protein sequence represented in the collection. The sequence-searching process was repeated five times, with the augmented EPK collection resulting from one iteration being used as bait for the next iteration. After removal of probable pseudogenes and sequencing artifacts (see below), the final EPK collection contained sequences representing 510 distinct putative human loci, 12 of which are thought to encode proteins that contain two separate EPK domains. Several distantly related sequences, which often lack some residues thought to be critical for EPK enzymatic activity, were included in the collection, including members of the ABC1, RIO1, C8FW, ILK and guanylate cyclase subfamilies. The hit set was also found to contain a number of sequences with even more remote similarity to EPKs, such as lipid kinases and antibiotic-resistance genes. To simplify subsequent analysis, members of these more distantly related families were not added to the EPK collection.

Removing pseudogenes and sequencing artifacts

The presence of pseudogene sequences and poor-quality sequences in target data sets tends to cause over-prediction of the number of EPKs present in the genome. Poor-quality sequences often present as novel singleton hits that closely resemble known genes, but encode potential proteins that are missing key residues, or contain apparent stop codons or frameshift mutations within functionally important regions. Removing hits with these features carries with it the risk of filtering poor-quality singleton sequences that represent real novel EPKs. In cases involving poor-quality singletons and other questionable sequences, we implemented consistent curation rules to help discriminate between sequences that probably represent novel functional family members, and sequences that probably represent pseudogenes and poor-quality sequences of known genes.

Sequences that appeared to be of very poor quality (three or more internal stop codons or frameshifts observed in any 60-amino-acid stretch) were rejected because these sequences either are derived from pseudogenes or, if they represent functional loci, novel sequence information is obscured by the high levels of noise present in the sequences. In addition, any sequences which were found to contain poly(A) tracts within the genomic sequence were filtered out, because such sequences almost certainly represent processed pseudogenes [45]. Processed pseudogenes are thought to arise when mRNA molecules are reverse transcribed and reintegrated into the genome. This mechanism results in the creation of pseudogenes that lack introns and often contain a poly(A) tail in the genomic sequence. The potential significance of pseudogene contamination within the hit set is highlighted by a study of pseudogenes on chromosomes 21 and 22 [46] which showed that approximately 20% of identifiable potential protein-coding regions represent pseudogenes. This same study determined that about half of all pseudogenes are processed pseudogenes.

The presence of internal stop codons or frameshift mutations was used to identify pseudogenes, but only if the feature could be verified in a sequence derived from a different cloning library. The nucleotide sequence of each hit that contained unverifiable stop codons or frameshift mutations was further analyzed by comparing the nucleotide sequence of the hit to the nucleotide sequence of closely related known EPKs. Comparisons that showed a pattern of nucleotide mismatches between the hit and the known EPK that suggested an absence of selective evolutionary pressure on the encoded protein sequence of the hit were used to filter suspect hits. For protein-coding portions of a nucleotide sequence, natural selection usually imposes greater constraints on the encoded protein sequence than on the underlying nucleotide sequence. As a result, comparing two functional genes that belong to the same family usually shows bias in the pattern of observed nucleotide identity between the genes. For example, a greater fraction of nucleotide mismatches in the third position of codons (the wobble position) will be detected than would be expected from a random distribution of nucleotide mismatches [47]. This preference for wobble position and other synonymous (codon preserving) nucleotide mismatches, results in levels of amino-acid identity that are greater than would be expected given the degree of identity between nucleotide sequences and an assumption of randomness in the pattern of nucleotide mismatches [48]. The degree of this bias will depend on the protein family and the corresponding functional constraints on the amino-acid sequence.

By contrast, comparison of a functional gene with a closely related processed pseudogene will often show no evidence of this codon-preserving bias in the pattern of nucleotide mismatches [47]. This is true for processed pseudogenes (which are usually non-functional from their inception) and older unprocessed pseudogenes, because most of the evolution of the nucleotide sequence of these pseudogenes was not constrained by natural selection on the encoded protein sequence. Other pseudogenes may have been inactivated relatively recently in evolutionary history, and may show substantial codon-preserving bias accumulated over the evolutionary period during which the gene was functional. These young non-processed pseudogenes are not readily identifiable on the basis of nucleotide sequence comparisons with known family members, and may pass through this filtering process.

In poor-quality sequences, sequencing artifacts occur more or less at random, and not in a manner that respects the integrity of the encoded protein sequence. As a result, comparisons of poor-quality sequences of known genes with the reference sequences for these genes should show no preference for synonymous nucleotide mismatches, and such hits should also be readily identifiable on the basis of the sort of comparison described here. By contrast, poor-quality sequences of novel functional genes are expected to present an intermediate picture, in which nucleotide mismatches that are due to sequencing artifacts show no preference for synonymous substitution, whereas mismatches that are due to evolution will show a preference for synonymous substitution. The degree of codon-preserving bias in these cases will depend on the relative impact of these two processes on the nucleotide sequence.

The degree of codon-preserving bias present in a hit was estimated by comparing the percent nucleotide identity between the hit and the most similar EPK with the percent amino-acid identity seen over the same region. We empirically determined cut-off scores for these comparisons that lead to rejection of most independently verified pseudogenes and simulated poor-quality sequences, while retaining all known EPK genes and verifiable novel EPK sequences (data not shown). Comparison conditions were chosen that would lead to retention of borderline sequences in order to reduce the probability of rejecting poor-quality singletons that represent novel EPK family members.

Aligning the catalytic domains

The sequences of the catalytic domains of all the human EPKs in the collection were manually aligned. Alignment of this family is difficult, because only a small number of residues are recognizably conserved across all family members. As a result, the full manual alignment contains small blocks of residues that are well aligned throughout the EPK family, punctuated by blocks of residues that are aligned within particular subfamilies but not throughout the rest of the EPK family. Although the alignment can be forced in these latter regions, length heterogeneity and poor residue conservation often make several alignments seem equally reasonable, and it is difficult to ascertain criteria for objectively choosing one over the others. To avoid effects stemming from arbitrary decisions on how to handle these difficult portions of the alignment, the full alignment was trimmed to contain only those regions that are relatively straightforward to align. The resulting partial alignment (see Additional data files), which more clearly delineates the key residues that support inclusion of a particular sequence in the EPK family, was used for subsequent phenogram construction.

The partial alignment shows that even the best conserved amino-acid positions in the EPK catalytic domain are found to vary in some family members (Table 1). To understand better how different residues can be accommodated at these key locations, available structural data were searched for examples of proteins with non-canonical residues in these positions. Structures and sequences were examined for a representative member of each of the 27 kinase families listed in the Structural Classification of Proteins (SCOP) database [49]. In these kinases, 7 out of 10 of the critical residues given in Table 1 are completely conserved. Previously, conserved residues and their structural roles have been discussed within the context of subdomains of the canonical kinase catalytic domain [11,50]. Here, this discussion is expanded on the basis of the known structures of the 27 kinase families and the variations in sequence noted in the entire human protein kinase family. Consensus sequence motifs are specified rather than completely conserved residues. Subdomain I comprises the GXGXϕGXV motif (ϕ = F, Y; single-letter amino-acid code) that forms a β hairpin to enclose one side of the triphosphate group of ATP. Among the known structures only G3 varies. G is preferred because of close steric contacts with an adjacent β strand. A small conformational change permits the A or S side chains to extend into the binding cavity behind the triphosphate; S hydrogen bonds to the β-phosphate. There is less space near G1 and G2, which are directly in contact with ribose and the β-phosphate, respectively. Nevertheless, there is also considerable sequence variation at these two positions. Subdomain II contains a conserved K that contacts the β-phosphate, whereas subdomain III has a conserved E that forms a salt bridge to the K to stabilize its conformation. Subdomain VIB has the motif HRDLKPXN in Ser/Thr kinases and HRDLXARN in Tyr kinases. Both the D and the downstream basic residue (K, R) are directly involved in the catalytic phosphorylation of substrate. Differences in the binding pocket for the phosphate-accepting residue permit the changes in the position in the sequence and in the nature of the basic residue; the N^{ϵ} atoms of the K and R residues interact with the acceptor OH and occupy the same position in space. Subdomain VII has the DFG motif. The D residue ligates Mg^{2+} which in turn binds the β- and γ-phosphates of ATP. Among the known structures, titin has a D→E modification. Unfortunately, only the apo-titin structure has been determined so there is no structure to define how the D→E change is accommodated in binding ATP. Subdomain VIII contains the TXXYXAPE motif in Ser/Thr kinases and PXXWXAPE in Tyr kinases as noted previously [11]. This motif is critical in stabilizing distinct conformations of the activation loop to form a platform for binding to the protein substrate. The first residue (T, P) lies directly underneath the acceptor residue (Ser/Thr, Tyr). In Tyr kinases, the P residue forces the loop to swing out to properly position the accepting Tyr residue. The last residue (E) forms a conserved salt bridge with an R in subdomain XI. Casein kinase 1 has an E→N variant and is missing the salt bridge. The D residue of subdomain IX hydrogen bonds to backbone NHs to stabilize the conformation of the catalytic loop of subdomain VIB.

Of the highly conserved residues (Table 1), the Gs of subdomain I show the most variability. Apparently, other amino

Table 1

Variability tolerated at key residue positions within the EPK catalytic domain

Residue	G1	G2	G3	K	E	D	N	D	E	D
Subdomain	I	I	I	II	III	VIB	VIB	VII	VIII	IX
Number not conserved*	52/473	29/473	135/473	16/482	19/483	29/493	23/493	23/496	34/493	22/493
% conserved	89%	94%	71%	97%	96%	94%	95%	95%	93%	95%
Amino-acid substitution†	A/14; S/10	A/5; D/4	A/55; S/53	C/4; R/3, N/3	A/7; D/3, R/3	N/17; S/5	(S/T)/8; (K/R)/8	E/6; G/5	D/14; N/10	N/7; A, G, T/3 each
SCOP (structure)‡										
Number not conserved	0/27	0/27	4/27	0/27	0/27	0/27	0/27	1/27	1/27	0/27
Amino-acid substitution			S/3, A/1					E/1	N/1	

The partial alignment (see Additional data files) was used to examine conservation among previously identified [52] highly conserved residues in EPK subdomains I-IX. The amino acids listed along the top of the table correspond to residues of PKA-Calpha (NP_002721): G1 = G51, G2 = G53, G3 = G56, K73, E92, D167, N172, D185, E209, D221. *The number of residues in Table 1 that differed from the consensus residue are listed first, along with the number of sequences informative at that residue. †The top two most commonly substituted amino acids at the given position and the number of occurrences of the substituted amino acid at that position. ‡The 27 EPK structural families defined in SCOP were searched for examples of EPKs with non-canonical residues at any of these ten positions. G53 is substitued three times with S (CK2, Cal, Phk) and once with A (Dap). Residue D in subdomain VII is substituted with E in titin. E in subdomain VIII is substituted with N in CK1.

acids with small side chains can replace the glycines while maintaining the β-hairpin conformation of the phosphate-binding loop and avoiding steric interference with ATP binding. Residues in subdomains VIB and VII are directly involved in catalysis and only the most conservative substitutions would be consistent with enzymatic activity. In the entire human protein kinase family these residues are approximately 95% conserved; many of the variants at these positions may lack enzymatic activity. In principle, more variability would be permitted for residues in subdomains III, VIII and IX which play a structural role but are not in direct contact with either ATP or the protein substrate. Nevertheless, these residues are also approximately 95% conserved. Their role in forming linking salt bridges and hydrogen bonds seems to be nearly as critical as that of residues directly involved in catalysis.

Building the phenogram

The sequence relationship between different EPKs was analyzed by estimating the phenetic distance between each possible pair of sequences and building a phenogram to portray the results graphically. A distance matrix representing the sequence similarity between each pair of sequences in the partial alignment was calculated using the Jukes-Cantor distance correction method [51]. A phenogram (see Additional data files) was then built from this matrix, using the neighbor-joining algorithm [52].

A dendrogram (Figure 1) that summarizes the results seen in the phenogram (see Additional data files) was constructed by collapsing branches that were relatively well separated from the rest of the tree, and naming the collapsed branch, guided by previously proposed subfamily nomenclature [9-11]. This earlier work suggests that EPKs can be classified into five families (PTK, AGC, CMGC, CaMK and OPK), which in turn can be split into a total of 55 subfamilies. Our work largely corroborates the validity of this classical naming scheme, but adds several names for branches representing sequences that do not fit into any of the previously described subfamilies, or that fall into one of the catch-all subfamilies (such as OPK_Other) whose members do not form a well defined sequence cluster in the phenogram. Sequences that do not cleanly fall into a cluster (singletons) are indicated in blue in Figure 1. The granularity of the classical subfamily naming scheme is much finer for some families (particularly PTK) than for others (particularly CaMK). In an attempt to provide greater resolution of clusters evident in the phenogram, some of the larger subfamilies were split into sets of smaller individually named branches. To facilitate translation between the branch names and the classical subfamily nomenclature, each human EPK is listed in the EPK data table (see Additional data files) along with the name used to identify its parent branch in Figure 1, and the classical subfamily name most appropriate for that EPK. No branch names are supplied for

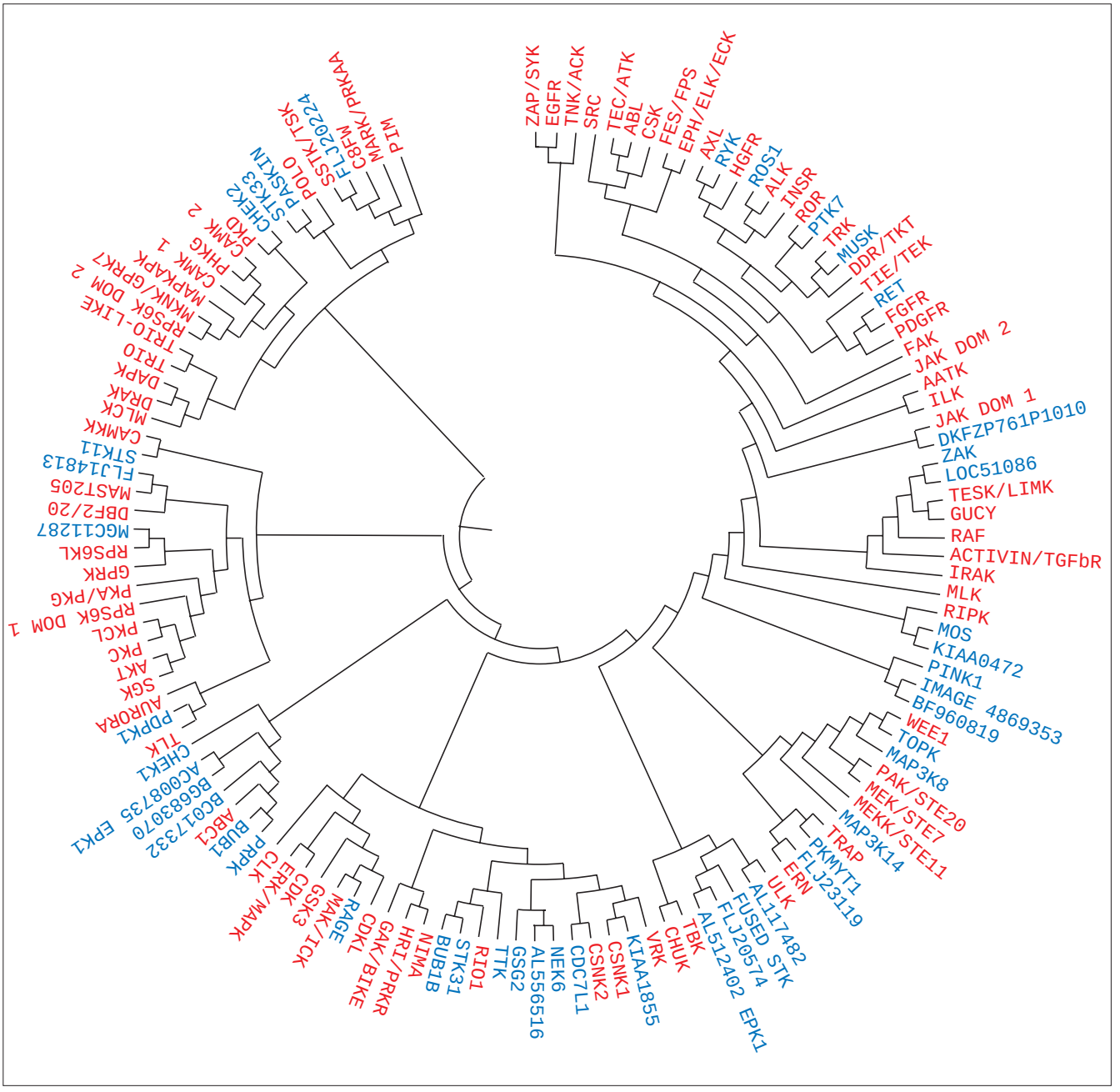

Figure 1
Dendrogram summarizing sequence groupings and branching patterns of EPKs derived from the phenogram. The program TreeExplorer was used to edit the main phenogram (see Additional data files), collapsing branches composed entirely of sequences that cluster with one another. Branch nomenclature was guided by previously accepted subfamily nomenclature, but differs from the classical subfamily nomenclature where justified by the structure of the phenogram. Sequences that do not clearly belong to a particular cluster were left in the figure as singletons. Because no branch lengths were specified for partial sequences present in the phenogram, and because only a portion of the catalytic domain of these sequences is available for comparison, assigning partial sequences to clusters would have been subject to significant error, and was not attempted. Partial sequences are therefore not included in this figure. Subfamilies are labeled in red, while singletons are labeled in blue. The relationship between each EPK, the assigned branch names, and the classical subfamily nomenclature is presented in the EPK data table (see Additional data files).

singletons or partial sequences, and classical names are not assigned to sequences that do not clearly fall into any of the classical subfamilies.

Nearly all previously recognized EPK subfamilies are represented in the human genome. No representatives were found for four of the classical subfamilies, and one of the subfamilies

is represented in humans by a single member. Human members of the AGC V (budding yeast AGC-related protein kinases) subfamily were not identified, but human EPKs assigned to the AKT and SGK branches are approximately 50% identical to yeast AGC V family members. The AGC VIII (flowering plant PVPK1 protein kinase homology) subfamily was also not represented in humans, although the DBF2/DBF20 branch contains members that are approximately 30% identical to plant AGC VIII subfamily members. Obvious human members of the OPK XIII (PKN prokaryotic protein kinases) subfamily were not found, although AJ336398_EPK1 may represent a distant family member, and human members of the ULK branch are approximately 40% identical to bacterial OPK XIII family members. The PTK XXII (nematode kin15/16 related kinases) subfamily was not identified in the human genome, although members of the PDGFR and FGFR branches are about 30-35% identical to known nematode PTK XXII subfamily members. In addition to these unrepresented families, only one member (Ros1) of the PTK XVIII (Ros/sevenless family) subfamily was detected. Each of the other 50 classical subfamilies was represented by two or more human sequences.

The distance calculations described above were carried out using only those sequences that extend across at least 95% of the alignment. As the presence of sequence fragments in an alignment is known to skew trees built from the alignment, fragment removal before distance calculation was essential. Unfortunately, this process had the undesired effect of excluding a sizable fraction of family members from tree construction. To provide some information on the subfamily membership of these partial sequences, fragments excluded from the initial construction of each tree were added back into the tree based on BLASTP similarity to the more complete sequences that were used to construct the tree (see Materials and methods). These partial sequences appear in the phenogram (see Additional data files) in parentheses, next to the more complete sequence with which they share the greatest degree of sequence identity. This scheme indicates which branch of the tree the fragmentary sequence probably belongs in, without attempting to assign a branch length from the fragment to the rest of the tree.

The overall accuracy of the phenogram is suggested by the clustering of similarly named proteins, the relatively good agreement with previously published categorization schemes [9-11], and the high level of congruence with the previously published tree of human tyrosine kinases [40]. In addition, alternative phenograms were constructed using different portions of the alignment and different algorithms (data not shown). In general, these alternative trees are similar to the main phenogram in their gross topology, although they often differ in their details. Most discrepancies involve partial sequences, poor-quality sequences or sequences representing outlying members of the family. The phenogram presented here was chosen because we believe it to have the greatest overall accuracy, even though a few outliers (notably BUB1 and BUB1B) are not correctly clustered.

The groupings suggested by the phenogram are based on sequence similarity across the entire alignment, which may suggest categories different from those suggested by considerations of much smaller stretches of residues known to be important for the characteristic functional features of a particular subfamily. Similarly, some EPKs have traditionally been categorized on the basis of residue segments that are important for the distinctive function of the subfamily, but lie completely outside the catalytic domain. For instance, PRKCM and PRKCN have traditionally been included in the protein kinase C (PKC) subfamily, largely because of the presence of a characteristic diacylglycerol-binding cysteine-rich zinc-finger-like domain [53]. This domain, which lies outside the EPK catalytic domain of certain members of the PKC subfamily, mediates the modulatory effects of diacylglycerol and phorbol esters on the function of sensitive PKCs. The trees presented here show that consideration of the EPK catalytic domain sequence alone does not lead to tight association of PRKCM and PRKCN with the rest of the PKC subfamily.

Cross-referencing to GenBank records

GenBank records were associated with each EPK record in order to provide supporting data for the existence, sequence, and transcriptional status of each locus. Readers may also find these records useful for retrieving supplementary information such as links to available literature, genetic mapping, and nomenclature for a particular EPK. Cross-references are provided to GenBank protein, transcript and genomic sequence records, because these data sources tend to contain mutually complementary information. For instance, information about exon-intron organization, non-coding control elements and genetic mapping information are best obtained from genomic records, whereas transcript records can suggest that the gene is transcriptionally active, confirm the predicted splicing pattern, and provide information about the tissue distribution of the gene product.

BLAST was used to identify GenBank records that were 100% identical in a 100-residue stretch to EPK sequences in the collection. Matching GenBank records were placed into the most appropriate of three sequence categories: protein, transcript or genomic. Although sequence records in all three categories were found for most EPK family members, occasionally representative sequence records for only one or two sequence categories could be identified. The finding that many EPKs were only represented in genomic or transcript data, but not both, suggests that each of these datasets is incomplete, or that mining one can give rise to artifacts not found in the other. When more than one representative sequence for an EPK was found within a particular category, one of the sequences was chosen as the primary cross-reference for that category. If a RefSeq [54] sequence was among

the choices, it was chosen as the primary cross-reference, otherwise the sequence with the greatest degree of overlap with the EPK reference sequence was chosen. For each EPK, primary cross-references for each sequence category are provided in the EPK data table (see Additional data files).

Retrieving data on chromosomal mapping

The extensive annotation present in many GenBank records frequently includes chromosomal mapping information. This information was extracted and is listed in the EPK data table (see Additional data files). Often, when multiple genomic sequence records were associated with a particular EPK, these records contained inconsistent mapping information. This occurred more frequently if the associated records were derived from the HTG portion of the GenBank sequence database. These discrepancies between different records were resolved as described in the Materials and methods.

EPK nomenclature

GenBank records often contain a list of names that have been used to identify the corresponding locus in the literature, have been chosen by a nomenclature committee, or have been suggested by the record submitter. This information was gathered and reconciled (see Materials and methods). If no accepted name could be found in any associated GenBank records, the EPK was given an interim name that was based on the accession string of associated GenBank records. Permanent names could have been assigned; this task is, however, best left to scientists engaged in more detailed characterization of these novel sequences. The names arrived at through this process served as identifiers in the alignments and trees, and are listed in the EPK data table (see Additional data files).

Estimating the number of novel EPKs in the collection is complicated by the incremental nature of gene characterization and imprecision in the definition of the term 'novel'. Because nomenclature and characterization often go hand-in-hand, evaluating the state of nomenclature can provide a rough estimate of the extent to which members of the collection have been previously characterized. Four hundred EPKs could be associated with a Human Genome Organization (HUGO) Gene Nomenclature Committee (HGNC) [55] name (such as MAP3K11, RAF1 or PRKCM) or description that implies previous knowledge about the potential function or subfamily membership of the sequence. In addition, 50 EPKs were associated with non-descriptive names (such as FLJ20574, LOC51086 or KIAA0175), that give little information regarding potential function or family ties, but indicate that the submitter believed that the sequence encoded a protein. For 60 EPKs, no name could be found for the corresponding gene or gene product, and an interim name was assigned. This method for estimating the novelty within the collection ignores the occasional unnamed family member whose GenBank annotation suggests probable membership in the EPK family, and the exceptional named family member (RNasel) whose GenBank annotation overlooks similarity to the EPK family. The level of previous characterization for each EPK sequence is summarized in the 'status' column of the EPK data table (see Additional data files). The 400 relatively well characterized sequences described above were assigned a status of '1'. The 50 somewhat less well characterized sequences were assigned a status of '2', and the 60 least well characterized sequences were assigned a status of '3'.

EPKs are frequently known by multiple names in the literature, and sometimes the HGNC-approved name for an EPK is not recognizable to researchers familiar with the corresponding literature. To aid readers in locating kinases of interest within the collection, alternative gene names were gathered from LocusLink [54] and Online Mendelian Inheritance in Man (OMIM) [56] records referred to in GenBank annotations associated with each EPK. These aliases are listed in the EPK data table (see Additional data files).

Conclusions

Reversible protein phosphorylation was discovered almost 40 years ago [3]. Subsequent work has shown that this covalent modification of cellular proteins is involved in the regulation of virtually all cellular functions. Most enzymes that mediate protein phosphorylation are members of a large and diverse evolutionarily conserved gene family. Further evaluation of similarities and differences in the sequences encoding protein kinases will provide significant scientific insights. These include information relevant to structure-function relationships, specificity of therapeutic agents targeting protein kinases, and potential function of uncharacterized family members. The work described here constitutes a summary and classification of the sequences in *H. sapiens* encoding these enzymes. This census of protein kinases in the human genome provides a tool and framework for further investigation of this important gene family.

Materials and methods

Sequence comparisons

Throughout this study, pairs of sequences were compared to determine whether or not the loci they represent were similar or identical to one another. This was done by performing searches with BLASTP (for protein comparisons and for searching protein datasets with protein queries) or BLASTN (for nucleotide comparisons) or TBLASTN (for searching nucleotide datasets with protein query sequences) using one sequence as bait against a BLAST-formatted database containing the other sequence. NCBI BLAST [57] implemented on a variety of UNIX platforms (Sun, SGI, Compaq) was used to carry out all the BLAST comparisons used in this study. BLAST parameters were set to retain low-scoring hits in the output (retain 500 top hits with *E*-values up to 10,000). Perl scripts were used to parse sequence alignments from the BLAST output and identify the 100-residue section

of each alignment that contained the maximum percentage identity obtainable in a window of that size for that alignment. The percentage identity between the two sequences within this optimal window was used to score the similarity between sequences for the purposes of creating the query set used for database searching, database searching itself, duplicate/splice variant filtering, assigning partial sequences to positions within the phenogram and database cross-referencing.

Searching public sequence data

An initial set of EPK amino-acid sequences was downloaded from Hanks's and Quinn's alignment of EPK catalytic domains [11,42]. This alignment includes approximately 300 sequences from a variety of organisms, along with links to GenBank [44] records representing each sequence. The links were used to retrieve the corresponding GenBank records. The GenBank records were parsed in order to obtain the protein sequence, taxonomic information and additional accessions that were used to retrieve the corresponding GenBank nucleotide records. This collection of EPK data was used to create a query protein set that contained EPK catalytic domain sequences that are less than 50% identical to each other within any window 100 amino acids long.

Each protein sequence in this query set was used to search the NT, NR, EST, HTG, GSS and STS divisions of GenBank [44] release 126.0 using BLASTP (for NR) or TBLASTN (for the remainder of the GenBank divisions). Hits produced by different query sequences were combined and duplicate records, which were identified on the basis of accession, were removed. Taxonomic information for each hit was retrieved from the corresponding complete GenBank record. Additional duplicates, along with splice variants, were identified by comparing all the sequences in the hit set to each other and looking for sequences that were 100% identical in a 100-residue stretch. Sequences meeting this criterion were selectively removed so as to leave only the longest representative from each set of duplicates and splice variants.

The resulting filtered hit set was manually reviewed by aligning the protein sequence of each hit with the catalytic domains of known EPKs and looking for the presence of a loose pattern of conserved residues known to be distinctive for the family [9-11]. Approved hits were added back to the EPK collection, and the expanded collection was used for additional rounds of searches. Non-human sequences were included in the EPK query set if the corresponding protein sequence was less than 50% identical in any 100-residue window to any other protein sequence in the collection.

Identifying pseudogenes and sequencing artifacts

Probable pseudogenes were identified on the basis of the presence of a poly(A) tail at the 3′ end of corresponding genomic sequences, the presence of internal stop codons or the presence of frameshift mutations within the catalytic domain. The presence of internal stop codons and frameshift mutations was used to classify a sequence as a pseudogene only if the feature could be confirmed in another sequence derived from an independent cDNA or genomic library, or if the original sequence contained three or more such features in any 60-amino-acid stretch.

Comparison of the levels of synonymous (amino-acid preserving) and non-synonymous (amino-acid altering) nucleotide mismatches, between hits containing unverifiable stop codons or frameshift mutations and the most similar registered EPK, was used to identify additional probable pseudogenes and sequencing artifacts. This was done by comparing the percentage identity of corresponding regions between the BLASTN (nucleotide versus nucleotide) and TBLASTX (virtual translation versus virtual translation) alignments. If the BLASTN alignment spanned at least 240 nucleotides, and the percentage identity in the BLASTN comparison exceeded that for the corresponding segment of the TBLASTX comparison by more than 10%, this was taken as evidence that the protein potentially encoded by the novel sequence was under less selective pressure than is seen for known EPK family members, and is therefore probably a pseudogene or sequencing artifact. The length and percentage cut-offs used were determined empirically so as not to eliminate any well supported EPKs from the collection.

Aligning catalytic domains and phenogram construction

The EPK catalytic domains of all the human amino-acid sequences in the collection were hand-aligned using the Wisconsin Package Version 10 [58] SeqLab alignment editor. Phenograms were constructed from these alignments on the basis of distance calculations performed by the Wisconsin Package implementation of the Distances program. The Jukes-Cantor method was used to correct the distance calculation for possible multiple substitutions at a single site. The resulting distance matrix was used to construct a phenogram using the neighbor-joining method as implemented by the Wisconsin Package GrowTree program. Only sequences spanning at least 95% of the partial alignment were included in the construction of the tree. The Nexus format tree file output by the GrowTree program was converted into Newick format using the TreeView program [59]. The Newick format tree file was then imported into TreeExplorer [60], which was used to build the summary dendrogram (Figure 1) by manually collapsing branches that represented sequence clusters evident in the phenogram (see Additional data files).

Placing partial sequences in the trees

Sequences that were excluded from phenogram construction because of their short length or high gap content were added back into the tree guided by comparison of the partial sequence with the more complete sequences utilized to construct the tree. The fragmentary sequences appear in the phenogram (see Additional data files) in parentheses, next to the more complete sequence to which they share the greatest

degree of protein sequence identity in a 100-residue window. This convention marks the inferred approximate location of the fragment within the tree, but does not assign a branch length from the sequence fragment to the rest of the tree.

Cross-referencing to GenBank records

Representative records in GenBank which correspond to each human EPK were identified by using EPK protein sequences or virtual translations for BLASTP searches against the NR division of GenBank, and EPK nucleotide sequences were used for BLASTN search against the NT, EST and HTG divisions of GenBank. The BLAST output was automatically searched for hits that showed 100% identity in a 100-residue window. Hits meeting this criterion were associated with the query EPK sequence and placed into one of three sequence categories: protein, transcript or genomic sequence. Hits discovered in NR were categorized protein, EST hits were categorized as transcript, and HTG hits were categorized as genomic. Whether a particular NT hit should be categorized as a transcript or genomic sequence was determined by examining the annotation of the corresponding GenBank record.

Nomenclature

Names for the EPKs were derived from the 'FEATURES' table [61] of GenBank records associated with each EPK. Values associated with the 'gene' qualifiers of the 'gene' and 'CDS' entries were parsed. If no information could be found in these fields, values associated with the 'note' qualifiers of the 'gene' and 'CDS' entries were examined for possible names. If multiple identifiers were found associated with the 'note' qualifier, the first one listed was given the highest priority. When naming information was present in multiple records, data were derived from records according to the following precedence: RefSeq NM, RefSeq NP, other non-XP protein, RefSeq XP, non-EST transcripts, and finally dbEST records. Nomenclature information was not derived from genomic sequence records.

If no name could be identified for a particular novel EPK, an interim name was assigned on the basis of one of the associated GenBank accession strings. If a reference transcript record had been associated with the EPK, the accession (minus the version number) of the reference transcript was chosen as the interim name for the EPK. If no transcript records were associated with the EPK, but a protein record existed for the EPK, the accession of the protein record was used as the interim name. If no GenBank transcript or protein records were associated with the sequence, the interim name was formed by taking the accession of the reference genomic record, and appending the suffix '_EPK1'. This latter convention was adopted because a genomic record may contain more than one protein kinase gene or gene fragment, in which case the name of the gene closest to the 5´ end of the published sequence would bear the suffix '_EPK1', the next would end in '_EPK2', and so on. In practice, it was never necessary to assign a suffix beyond '_EPK1'.

Alternative names for the EPKs were retrieved from LocusLink [54], and OMIM [56] records referred to from GenBank records that had been associated with the EPK collection.

Mapping

Mapping data was derived from annotation contained in GenBank records associated with each EPK. Values associated with the 'map' and 'chromosome' attributes of the 'source' entry of the 'FEATURES' table of each associated record were retrieved. If mapping information was found in multiple records, source precedence was assigned in the following descending order: RefSeq nucleotide records, RefSeq protein records, NR protein records, NT mRNA records, NT genomic records, and HTG genomic records. If multiple records tied for highest precedence, the map position most frequently indicated in those records was used.

Additional data files

Additional data files available with the online version of this article include the following files.

Partial alignment of human EPK catalytic domains

Well conserved portions of the catalytic domains of the various human EPKs were aligned by hand. The sequences are listed in the same order as they appear in the phenogram (see below), and are numbered in the same order as they appear in the EPK data table (see below). Identifiers ending in '_DOM2' indicate the second EPK domain from a protein which contains two separate EPK domains. Gaps in the alignment are represented either by '.' or '~' characters. This alignment is in PDF format and should be viewed with a PDF-capable reader such as Adobe Acrobat Reader.

Phenogram based on the partial alignment

Distances between each pair of protein sequences in the partial alignment (see above) were calculated using the Jukes-Cantor method for correcting for multiple substitutions at a single site. The tree was built using the neighbor-joining algorithm. Sequences not spanning at least 95% of the partial alignment were excluded from initial tree-building and then added back into the final tree using BLASTP based similarity estimates (see Materials and methods). Identifiers for these partial sequences appear in parentheses in the tree, next to the more complete EPK to which they are most similar. No attempt was made to assign branch lengths between the partial sequences and the rest of the tree. Identifiers ending in '_DOM2' or '_DM2' indicate the second EPK domain from a protein that contains two separate EPK domains.

EPK data table

EPK nomenclature, associated GenBank records, and information regarding subfamily membership, novelty and genetic

mapping are provided. Each EPK was associated with corresponding records from GenBank that represent protein, transcript or genomic sequences. One representative in each sequence category is provided if available. RefSeq record accessions are provided whenever available. Mapping and nomenclature were parsed from the GenBank records, or formed from representative accession strings as described in the Materials and methods section. Aliases are derived from LocusLink and OMIM records referenced in the corresponding GenBank transcript record. The novelty of each sequence was estimated from the associated GenBank nomenclature and description, and summarized in the 'status' column. EPK family members which have been previously named as EPKs, categorized into an EPK subfamily, or whose description clearly suggests they are EPK family members were assigned a status of 1. EPKs whose names or description suggest that the annotator recognized the existence of the protein without giving a clear indication of its similarity to EPKs beyond a similarity score calculated by automated annotation processes were assigned a status of 2. EPKs whose GenBank annotation did not clearly delineate the encoded protein or its potential function were assigned a status of 3. The table is presented in comma separated values (csv) format, and is best viewed with a spreadsheet program such as Microsoft Excel.

Acknowledgements

We thank Melanie Cobb (UT Southwestern Medical Center, Dallas, TX) for useful discussions about the manuscript and suggestions regarding EPK nomenclature.

References

1. 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.
2. Cohen P: **The regulation of protein function by multisite phosphorylation - a 25 year update.** *Trends Biochem Sci* 2001, **25:**596-601.
3. Graves JD, Krebs EG: **Protein phosphorylation and signal transduction.** *Pharmacol Ther* 1999, **82:**111-121.
4. Blume-Jensen P, Hunter T: **Oncogenic kinase signalling.** *Nature* 2001, **411:**355-365.
5. Schlessinger J: **Cell signaling by receptor tyrosine kinases.** *Cell* 2000, **103:**211-225.
6. Chang L, Karin M: **Mammalian MAP kinase signalling cascades.** *Nature* 2001, **410:**37-40.
7. Pearson RB, Kemp BE: **Protein kinase phosphorylation site sequences and consensus specificity motifs: Tabulations.** *Methods Enzymol* 1991, **200:**62-81.
8. Pawson T, Scott JD: **Signaling through scaffold, anchoring, and adapter proteins.** *Science* 1997, **278:**2075-2080.
9. Hanks SK, Quinn AM, Hunter T: **The protein kinase family: conserved features and deduced phylogeny of the catalytic domains.** *Science* 1988, **241:**42-52.
10. Hanks SK, Hunter T: **Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification.** *FASEB J* 1995, **9:**576-596.
11. Hanks S, Quinn AM**: Protein kinase catalytic domain sequence database: identification of conserved features of primary structure and classification of family members.** *Methods Enzymol* 1991, **200:**38-62.
12. Morgan DO, De Bondt HL: **Protein kinase regulation: insights from crystal structure analysis.** *Curr Opin Cell Biol* 1994, **6:**239-246.
13. Canagarajah BJ, Khokhlatchev A, Cobb MH, Goldsmith EJ: **Activation mechanism of the MAP kinase ERK2 by dual phosphorylation.** *Cell* 1997, **90:**859-869.
14. Bossemeyer D: **Protein kinases - structure and function.** *FEBS Lett* 1995, **369:**57-61.
15. Smith CM, Radzio-Andzelm E, Madhusudan, Akamine P, Taylor SS: **The catalytic subunit of cAMP-dependent protein kinase: prototype for an extended network of communication.** *Prog Biophys Mol Biol* 1999, **71:**313-341.
16. Sowadski JM, Epstein LF, Lankiewicz L, Karlsson R: **Conformational diversity of catalytic cores of protein kinases.** *Pharmacol Ther* 1999, **82:**157-164.
17. Beeler JF, LaRochelle WJ, Chedid M, Tronick SR, Aaronson SA: **Prokaryotic expression cloning of a novel human tyrosine kinase.** *Mol Cell Biol* 1994, **14:**982-988.
18. Wymann MP, Pirola L: **Structure and function of phosphoinositide 3-kinases.** *Biochim Biophys Acta* 1998, **1436:**127-150.
19. Daigle DM, McKay GA, Thompson PR, Wright GD: **Aminoglycoside antibiotic phosphotransferases are also serine protein kinases.** *Chem Biol* 1999, **6:**11-18.
20. Popov KM, Zhao Y, Shimomura Y, Kuntz MJ, Harris RA: **Branched-chain alpha-ketoacid dehydrogenase kinase.** *J Biol Chem* 1992, **267:**13127-13130.
21. Hartley KO, Gell D, Smith GC, Zhang H, Divecha N, Connelly MA, Admon A, Lees-Miller SP, Anderson CW, Jackson SP: **DNA-dependent protein kinase catalytic subunit: a relative of phosphatidylinositol 3-kinase and the ataxia telangiectasia gene product.** *Cell* 1995, **82:**849-856.
22. Banin S, Moyal L, Shieh S, Taya Y, Anderson CW, Chessa L, Smorodinsky NI, Prives C, Reiss Y, Shiloh Y, Ziv Y: **Enhanced phosphorylation of p53 by ATM in response to DNA damage.** *Science* 1998, **281:**1674-1677.
23. Cimprich KA, Shin TB, Keith CT, Schreiber SL: **cDNA cloning and gene mapping of a candidate human cell cycle checkpoint protein.** *Proc Natl Acad Sci USA* 1996, **93:**2850-2855.
24. Harihan IK, Adams JM: **cDNA sequence for human *bcr*, the gene that translocates to the *abl* oncogene in chronic myeloid leukaemia.** *EMBO J* 1987, **6:**115-119.
25. Maru Y, Witte ON: **The *BCR* gene encodes a novel serine/threonine kinase activity within a single exon.** *Cell* 1991, **67:**459-468.
26. Runnels LW, Yue L, Clapham DE: **TRP-PLIK, a bifunctional protein with kinase and ion channel activities.** *Science* 2001, **291:**1043-1047.
27. Steinbacher S, Hof P, Eichinger L, Schleicher M, Gettemans J, Vandekerckhove J, Huber R, Benz J: **The crystal structure of the *Physarum polycephalum* actin-fragmin kinase: an atypical protein kinase with a specialized substrate-binding domain.** *EMBO J* 1999, **18:**2923-2929.
28. Yamaguchi H, Matsushita M, Nairn AC, Kuriyan J: **Crystal structure of the atypical protein kinase domain of a TRP channel with phosphotransferase activity.** *Mol Cell* 2001, **7:**1047-1057.
29. Russo AA, Jeffrey PD, Pavletich N: **Structural basis of cyclin-dependent kinase activation by phosphorylation.** *Nat Struct Biol* 1996, **3:**696-700.
30. Huse M, Kuriyan J: **The conformational plasticity of protein kinases.** *Cell* 2002, **109:**275-282.
31. Hubbard SR, Till JH: **Protein tyrosine kinase structure and function.** *Annu Rev Biochem* 2000, **69:**373-398.
32. Johnson LN, Noble ME, Owen DJ: **Active and inactive protein kinases: structural basis for regulation.** *Cell* 1996, **85:**149-158.
33. Moscat J, Diaz-Meco MT: **The atypical protein kinase Cs. Functional specificity mediated by specific protein adapters.** *EMBO Rep* 2000, **1:**399-403.
34. Faust M, Montenarh M: **Subcellular localization of protein kinase CK2. A key to its function?** *Cell Tissue Res* 2000, **301:**329-340.
35. Dan I, Watanabe NM, Kusumi A: **The Ste20 group kinases as regulators of MAP kinase cascades.** *Trends Cell Biol* 2001 **11:**220-30.
36. Kimm HH, Vijapurkar U, Hellyer NJ, Bravo D, Koland JG: **Signal transduction by epidermal growth factor and heregulin via the kinase-deficient ErbB3 protein.** *Biochem J* 1998, **334:**189-195.
37. Hunter T, Plowman GD: **The protein kinases of budding yeast: six score and more.** *Trends Biochem Sci* 1997, **22:**18-22.
38. Popovici C, Roubin R, Coulier F, Pontarotti P, Birnbaum D: **The family of *Caenorhabditis elegans* tyrosine kinase receptors:**

similarities and differences with mammalian receptors. *Genome Res* 1999, **9:**1026-1039.
39. Morrison DK, Murakami MS, Cleghon V: **Protein kinases and phosphatases in the *Drosophila* genome.** *J Cell Biol* 2000, **150:**F57-F62.
40. Robinson DR, Wu YM, Lin SF: **The protein tyrosine kinase family of the human genome.** *Oncogene* 2000, **19:**5548-5557.
41. **Mammalian Protein Kinases** [http://www.kinase.com/mammalian/mam_kinase.htm]
42. **Protein Kinase Catalytic Domain Sequence Alignments** [http://pkr.sdsc.edu/html/pk_classification/pk_catalytic/pk_hanks_seq_align_long.html]
43. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
44. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Rapp BA, Wheeler DL: **GenBank.** *Nucleic Acids Res* 2002, **30:**17-20.
45. Vanin EF: **Processed pseudogenes: characteristics and evolution.** *Annu Rev Genet* 1985, **19:**253-272.
46. Harrison PM, Hegyi H, Balasubramanian S, Luscombe NM, Bertone P, Echols N, Johnson T, Gerstein M: **Molecular fossils in the human genome: identification and analysis of the pseudogenes in chromosomes 21 and 22.** *Genome Res* 2002, **12:**272-280.
47. Gojobori T, Li WH, Graur D: **Patterns of nucleotide substitution in pseudogenes and functional genes.** *J Mol Evol* 1982, **18:**360-369.
48. Kimura M: **Preponderance of synonymous changes as evidence for the neutral theory of molecular evolution.** *Nature* 1977, **267:**275-276.
49. Murzin AG, Brenner SE, Hubbard T, Chothia C: **SCOP: a structural classification of proteins database for the investigation of sequences and structures.** *J Mol Biol* 1995, **247:**536-540.
50. Hanks S, Hunter T: **Chapter 2.** In *The Protein Kinase Facts Book*. Edited by Hardie G, Hanks S. New York: Academic Press; 1995: 7-47.
51. Jukes TH, Cantor CR: **Evolution of protein molecules.** In *Mammalian Protein Metabolism*. Edited by Munro HN. New York: Academic Press; 1969: Vol III 21-132.
52. Saitou N, Nei M: **The neighbor-joining method: a new method for reconstructing phylogenetic trees.** *Mol Biol Evol* 1987, **4:**406-425.
53. Hurley JH, Newton AC, Parker PJ, Blumberg PM, Nishizuka Y: **Taxonomy and function of C1 protein kinase C homology domains.** *Protein Sci* 1997, **6:**477-480.
54. Maglott DR, Katz KS, Sicotte H, Pruitt KD: **NCBI's LocusLink and RefSeq.** *Nucleic Acids Res* 2000, **28:**126-128.
55. Povey S, Lovering R, Bruford E, Wright M, Lush M, Wain H: **The HUGO Gene Nomenclature Committee (HGNC).** *Hum Genet* 2001, **109:**678-680.
56. **Online Mendelian Inheritance in Man** [http://www.ncbi.nlm.nih.gov/omim/]
57. **NCBI BLAST** [ftp://ftp.ncbi.nlm.nih.gov/blast/]
58. Genetics Computer Group: *Wisconsin Package User Release Notes Version 10.0*. Madison, WI: Genetics Computer Group; 1999.
59. Page RD: **TreeView: an application to display phylogenetic trees on personal computers.** *Comput Appl Biosci* 1996, **12:**357-358.
60. **TreeExplorer** [http://evolgen.biol.metro-u.ac.jp/pub/MolEvol/TE212.zip]
61. **The DDBJ/EMBL/GenBank Feature Table Definition** [http://www.ncbi.nlm.nih.gov/collab/FT/index.html]