PROTEINS: Structure, Function and Bioinformatics

In press (Prot-00261-2006.R1)

TITLE: Classification and Functional Annotation of Eukaryotic Protein Kinases

RUNNING TITLE: Classification and Annotation of Protein Kinases

AUTHORS: Diego Miranda-Saavedra and Geoffrey J. Barton*

School of Life Sciences, University of Dundee, Dow St, Dundee DD1 5EH, Scotland

KEYWORDS: database searching, signal transduction, phosphorylation, hidden Markov model, evolution

*Correspondence to: Geoffrey Barton

School of Life Sciences

University of Dundee

Dow St

Dundee DD1 5EH

Scotland, UK

Tel. +44 1382 385860

Fax: +44 1382 385764

E-mail: geoff@compbio.dundee.ac.uk

E-mail address:

D. Miranda-Saavedra: diego@compbio.dundee.ac.uk

Abstract

Reversible protein phosphorylation by protein kinases and phosphatases is a ubiquitous signalling mechanism in all eukaryotic cells. A multi-level hidden Markov model library is presented which is able to classify protein kinases into one of 12 families, with a mis-classification rate of zero on the characterised kinomes of *H. sapiens, M. musculus, D. melanogaster, C. elegans, S. cerevisiae, D. discoideum,* and *P. falciparum.* The Library is shown to outperform BLASTP and a general Pfam hidden Markov model of the kinase catalytic domain in the retrieval and family-level classification of protein kinases. The application of the Library to the 38 unclassified kinases of yeast enriches the yeast kinome in protein kinases of the families AGC (5), CAMK (17), CMGC (4), and STE (1), thereby raising the family-level classification of yeast conventional protein kinases from 66.96% to 90.43%. The application of the Library to 21 eukaryotic genomes shows 7 families (AGC, CAMK, CK1, CMGC, STE, PIKK, RIO) to be present in all genomes analysed, and so are likely to be essential to eukaryotes. Putative tyrosine kinases (TKs) are found in the plants *A. thaliana* (2), *O. sativa ssp. Indica* (6), and *O. sativa ssp. Japonica* (7), and in the amoeba *E. histolytica* (7). To our knowledge, TKs have not been predicted in plants before. This also suggests that a primitive set of TKs might have predated the radiation of eukaryotes. Putative tyrosine kinase-like kinases (TKLs) are found in the fungi *C. neoformans (*2), *P. chrysosporium* (4), in the Apicomplexans *C. hominis* (4*), P. yoelii* (4), and *P. falciparum* (6), the amoeba *E. histolytica* (109), and the alga *T. pseudonana* (6). TKLs are found to be abundant in plants (776 in *A. thaliana*, 1010 in *O. sativa ssp. Indica*, and 969 in *O. sativa ssp. Japonica*). TKLs might have predated the radiation of eukarytes too and have been lost secondarily from some fungi. The application of the Library facilitates the annotation of kinomes and has provided novel insights on the early evolution and subsequent adaptations of the various protein kinase families in eukaryotes.

Introduction

Reversible protein phosphorylation by protein kinases and phosphatases is thought to regulate virtually every cellular activity (1). Since abnormal levels of phosphorylation are known to be responsible for severe diseases (2), there is considerable therapeutic promise in obtaining a detailed understanding of phosphorylation events, both within specific cell types and in an evolutionary context. A thorough understanding of the evolution of protein kinases will help decipher how signalling events have shaped the development, pathology and biochemistry of eukaryotes and so lead to a better description of the biochemical circuitry of cells which may guide the development of more effective drugs (3).

The KinBase resource (http://www.kinase.com/kinbase/, (4)) reflects the currently accepted classification of eukaryotic protein kinases, which are split into two broad groups: 'conventional' protein kinases (ePKs) and 'atypical' protein kinases (aPKs). ePKs are the largest group, and have been sub-classified into 8 families by examining sequence similarity between catalytic domains, the presence of accessory domains, and by considering any known modes of regulation (5, 6). The 8 ePK families defined in KinBase are: the AGC family (including cyclic-nucleotide and calcium-phospholipid-dependent kinases, ribosomal S6-phosphorylating kinases, G protein-coupled kinases, and all close relatives of these groups); the CAMKs (calmodulin-regulated kinases); the CK1 family (casein kinase 1, and close relatives); the CMGC family (including cyclin-dependent kinases, mitogen-activated protein kinases, CDK-like kinases, and glycogen synthase kinase); the RGC family (receptor guanylate cyclase kinases, which are similar in domain sequence to tyrosine kinases); the STE family (including many kinases functioning in MAP kinase cascades); the TK family (tyrosine kinases); and the TKL family (tyrosine kinase-like kinases, a diverse group resembling TK but which are in fact serine-threonine kinases). A ninth group, called the 'Other'

group, consists of a mixed collection of kinases that could not be classified easily into the previous families.

The aPKs are a small set of protein kinases that do not share clear sequence similarity with ePKs, but have been shown experimentally to have protein kinase activity (6) and comprise the following families: Alpha (exemplified by myosin heavy chain kinase of *Dictyostelium discoideum*); PIKK (phosphatidyl inositol 3' kinase-related kinases); PHDK (pyruvate dehydrogenase kinases); RIO (‘*ri*ght *o*pen reading frame’ as it was one of two adjacent genes that were found to be transcribed divergently from the same intergenic region (7); BRD (bromodomain-containing kinases); ABC1 (ABC1 domain-containing kinases); and TIF1 (transcriptional intermediary factor 1). The aPKs also include H11 (a homolog of gene ICP10 of Herpes simplex virus), FASTK (Fas-activated serine-threonine kinase), G11 (reported kinase activity against alpha casein and histones), BCR (fused with Abl in chronic myologenous leukemia), TAF (TATA binding factor-associated factor), and A6 (2 genes in human, A6 and A6r). However, only the Alpha, PIKK, PHDK and RIO families have strong experimental evidence for kinase activity. For the RIO and Alpha families, X-ray crystallography has revealed clear similarities to the ePK kinase fold (8, 9).

Entries in KinBase are filtered by stringent criteria, including verification by cDNA cloning, in order to reduce the possibility of incorrect classification. As a consequence, KinBase is favoured by experimentalists working on kinases and signal transduction pathways. A disadvantage of KinBase is that it is only cross-referenced to a variety of genome-specific databases such as the *Saccaromyces* Genome Database rather than to universal databases such as UniProt (10).

In order to obtain a clearer picture of the spectrum of kinase genes and so understand the events that have shaped the evolution of protein kinase function in various lineages, it is necessary to identify

the protein kinase complements (the 'kinomes') of as many organisms as possible. The kinomes of several organisms have previously been determined by a combination of *in silico* and wet-lab studies, including those for *C. elegans* (11), *S. cerevisiae* (12), *D. melanogaster* (4, 13), *H. sapiens* (6), *M. musculus* (14), *P. falciparum* (15), and *D. discoideum* (16). Identification of protein kinase sequences in these organisms has typically relied on BLAST searches (17) and HMM profiles designed to identify kinase catalytic domains (18). Finer-grained classification into the 12 sub-families has been achieved by the application of consensus sequences for the catalytic and activation loops (subdomains VIB and VIII, (19)), and by clustering on sequence similarity to previously classified protein kinases (6, 15). However, classifying all protein kinases in a newly sequenced genome by these methods has proved very time-consuming. The growth in eukaryotic genome sequencing has highlighted the need for more efficient and reliable methods of kinase classification.

Profile hidden Markov models (HMMs) are statistical descriptions of sequence conservation from multiple sequence alignments (20), and have been shown to outperform standard pairwise sequence comparison methods, both in terms of sensitivity and specificity (21). HMMs form the basis of protein family and domain description libraries such as SUPERFAMILY (22) and Pfam (18), which have been used successfully in the automatic annotation of many genomes (23). Brown *et al* have described formally the division of protein families into subfamilies for the purpose of achieving finer classification with HMMs (24). Brown *et al* found that a better separation between sub-family members is possible by using sub-family HMMs rather than a single HMM for the entire family, and that homologues are recognised with stronger scores, and non-homologues are rejected with larger E-values.

In this paper the development of a multi-level HMM library from functional sub-families of

eukaryotic protein kinases is described and applied to the classification of protein kinases into the 12 sub-families for 21 eukaryotic genomes.

Methods

Source of kinase sequences

KinBase (http://www.kinase.com/kinbase/, (4)) was the source for the annotated protein kinase complement of *H. sapiens, M. musculus, D. melanogaster, C. elegans, S. cerevisiae, and D. discoideum* (Table I). The protein kinase complement of *P. falciparum* was kindly made available by Dr. Christian Doerig (Wellcome Trust Centre for Molecular Parasitology, Glasgow, UK).

Since KinBase does not map its entries onto Swiss-Prot/TrEMBL, this mapping was performed for sequences from *H. sapiens, D. melanogaster*, *S. cerevisiae*, and *C. elegans* (Swiss-Prot/TrEMBL Release 47, July 2004). The total number of protein kinases for *D. melanogaster, C. elegans, H. sapiens,* and *S. cerevisiae* in KinBase was 1487, which included 107 pseudogenes. Although it has been noted that many kinase pseudogenes are transcribed and could have a residual or scaffolding function (6), kinase pseudogenes were not mapped onto Swiss-Prot/TrEMBL since many of them are partial transcripts or have non-amino acid characters in their sequences that indicate STOP codons.

The mapping of KinBase sequences onto Swiss-Prot/TrEMBL was performed by BLASTP-searching each KinBase sequence against Swiss-Prot/TrEMBL (Release 47, July 2004). Each of the 1380 result files was parsed to retrieve the top hit belonging to the same species as the query sequence. Three human sequences (SgK110, SgK424, and Slob) and two worm sequences (D1073.1 and aSWk029) were found to have no corresponding sequence in Swiss-Prot/TrEMBL. 993 (75.4%) of KinBase ePK sequences could be mapped directly to identical sequences in Swiss-Prot/TrEMBL, whereas 162 (12.3%) of ePKs were mapped to sequences of non-identical length but whose alignment contained no gaps. A further 161 (12.2%) of KinBase ePK sequences matched to

Swiss-Prot/TrEMBL polypeptides of non-identical length but where the alignment contained gaps. This group encompasses pairs of alternatively spliced forms of kinase sequences, or pairs of very similar sequences with major DNA sequencing errors. All aPK sequences from KinBase were found to map to identical sequences in Swiss-Prot/TrEMBL.

KinBase contains full-length kinase sequences, but does not annotate the location of the catalytic domain in all sequences. In order to retrieve the catalytic domains of ePKs from *H. sapiens, D. melanogaster, C. elegans,* and *S. cerevisiae*, HMMs were generated from known human kinase catalytic domains to represent each family of kinases, one HMM per family. These HMMs were used to scan the full-length protein kinase sequences from each species. All kinase catalytic domains retrieved in this way were inspected and found to have the sequence features typical of the kinase catalytic domain (5).

The catalytic domains of the aPKs, PIKK, Alpha, and RIO were retrieved by scanning the full-length protein sequences with the Pfam HMMs PF00454, PF02816, and PF01163, respectively. The catalytic domains of PDHKs were manually extracted from a multiple alignment by following the description given in (25).

Table II summarises the aPK complement of *H. sapiens, M. musculus, S. cerevisiae, D. melanogaster,* and *C. elegans* , and the mapping of aPKs to the Swiss-Prot database.

Library 1: Single HMM per family, and validation

A set of 12 profile Hidden Markov Models (HMMs) (20) was built, one for each of the families of 8 ePK and 4 aPK kinases from an organism (the 'training set'), and tested to see if it could classify correctly the kinases of a different organism (the 'test set'). Models from *H. sapiens, C. elegans and D. melanogaster* were used as both training and test sets; however, since *S. cerevisiae* does not

possess kinases of the RGC, TK, or TKL families, the ePKs of *S. cerevisiae* were only used as a test set. For the aPKs, only aPK catalytic domains from *H. sapiens* and *C. elegans* were used to build HMMs, since these organisms contain all four families of aPKs (RIO, Alpha, PDHK, and PIKK), whereas *D. melanogaster* and *S. cerevisiae* lack alpha kinases (4).

The HMMs representing the families of ePKs and aPKs were created from multiple sequence alignments of all the protein kinase catalytic domain sequences in a family, generated by the AMPS suite of programs (26). Since an HMM profile is only as reliable as the underlying alignment from which it is derived the quality of the alignment was assessed by Z-score analysis. In benchmarks, alignments that gave Z-scores above 5.0S.D. had better than 70% accuracy within the core secondary structural regions of the proteins. All 32 ePK and aPK alignments gave Z-scores of >19 (data not shown), suggesting that the alignments are likely to be of high quality and so suitable for the derivation of HMMs. As a further test of quality, each alignment was inspected by eye to verify conservation of the core kinase catalytic domain motifs (5) HMMs were built with the HMMER suite of programs (http://hmmer.wustl.edu, (27), HMMER version 2.1.1) . The hmmbuild program was set to build HMMs that could identify one or more non-overlapping alignments to the complete model (multiple global alignments with respect to the model, and local with respect to the sequence) and the HMMs were calibrated by the hmmcalibrate program.

The classification performance of the family-specific models was tested by searching against the full-length protein kinase sequences of the characterised kinomes from KinBase. All kinases gave E-values better than 1e-05. The classification was regarded as correct whenever the model that aligned with the best E-value to a given query sequence belonged to the same family as the sequence being classified. Table III summarises the classification performance of Library 1 built to represent the families of protein kinase catalytic domains from *H. sapiens.* Library 1 correctly

classified 394/395 of human ePK domains. The mis-classified human kinase turned out to be RSK2, an AGC kinase. RSK kinases harbour two functional kinase catalytic domains, and whereas the first domain was correctly classified as 'AGC', the second domain would rather be classified as 'CAMK'. This appears to be a typical case of neo-functionalisation following domain duplication. Since the classification of RSK kinases is based on the identity and behaviour of the first kinase catalytic domain, this mis-classified kinase may be ignored.

Similarly good performance was obtained for the classification of *S. cerevisiae* (75/77) and *D. melanogaster* (175/177) ePKs. However, the HMM library performed less well on the ePKs of *C. elegans*, where only 338/359 ePKs (94.15%) were classified correctly. The mis-classification rate was particularly marked in the TKL family of kinases from *C. elegans* where only 6/15 (40%) were classified correctly. In contrast, the HMM library derived from human aPKs correctly classified all atypical kinases of *C. elegans*, *H. sapiens*, *D. melanogaster, and S. cerevisiae,* reflecting the high degree of sequence conservation of aPKs across a large evolutionary distance.

The classification performance of the HMM libraries built from *D. melanogaster* and *C. elegans* kinase catalytic domains were assessed in a similar way. The *D. melanogaster*-derived library had a performance similar to the *H. sapiens*-derived library, yielding a correct self-classification rate of 176/177 (99.44%), and cross-species correct classification rates of 77/77 (100%), 389/395 (98.48%), and 347/359 (96.66%) for *S. cerevisiae*, *H. sapiens*, and D. *melanogaster*. Self-classification by the *C. elegans*-derived library yielded a correct classification rate of only 161/359 (44.84%), and similarly poor rates on *H. sapiens* (185/395, 53.17%), *D. melanogaster* (99/177, 55.93%) and *S. cerevisiae* (50/77, 64.93%) (data not shown).

Library 2: Multiple HMMs per family, and assessment of performance

With the exception of *C. elegans*, the classification performed by Library 1 HMMs appeared good. In an attempt to raise the accuracy of classification still further, and in particular for *C. elegans*, the larger sequence families were sub-divided to generate multiple HMMs for each family to create Library 2. Sub-dividing large families can raise recognition accuracy for HMMs by allowing the unique features of each sub-family to be captured more effectively. It can also give benefits by eliminating the need to align divergent sequences and so raise the accuracy of the alignments from which the HMMs are derived (23).

Given the wide spectrum of protein kinases in *H. sapiens* and the high classification rates of the human kinase-derived library (Library 1, Table III), it was decided to base all further work on the available human kinases. Attempts were only made to improve the library of HMMs representing human ePKs since the HMM library of aPKs of human origin already gave a mis-classification rate of zero on the aPK complement of all 4 organisms.

Figure 1 outlines the iterative process devised to optimise the representation of each kinase family by multiple sub-family HMMs. Z-scores were calculated for each pairwise alignment of the ePK catalytic domain sequences in each family. Complete linkage clustering was then applied to the Z-scores and clusters selected according to a Z-score cutoff. This is illustrated for the CK1 family in Figure 2. An initial cutoff of 19.2 clustered the twelve kinase catalytic domain sequences of human CK1 protein kinases into a single group. A cutoff of 20 S.D. split the family into two groups of 5 and 7 domains (Fig. 2). For each sub-family a multiple alignment and HMM was generated. As for Library 1, the classification performance of the resulting library of HMMs was tested on the full-length kinases of *H. sapiens, C. elegans, S. cerevisiae* and *D. melanogaster*. It was found empirically that generating models for any family where each model consisted of only a handful of

sequences, resulted in a serious decrease in the rate of correct classification. The trade-off between specificity and number of models for each family led to the number of models summarised in Table IV. 7 models were needed to represent the largest family of kinases from *H. sapiens* (the CAMK family), while the small RGC family and all the atypical kinases continued to be represented by single models from Library 1. Figure 3 provides an overview of the development of Library 2 which has multiple HMMs for larger families from Library 1. The resulting library of sub-family HMMs was referred to as "Library 2".

Table V shows the classification rate for Library 2 on the ePKs from *H. sapiens, D. melanogaster, C. elegans and S. cerevisiae*. With the exception of *C. elegans*, the library gave perfect classification, while for *C. elegans*, the accuracy rose by 14 kinases from the 338/359 achieved by Library 1 to 354/359. The largest improvement for *C. elegans* was seen for the TKL family, where the number of mis-classified kinases dropped from 9 to zero. No change in classification accuracy was seen for the CK1 family (83/85) while for CMGC, a single error was introduced with 48/49 correctly classified kinase domains compared to 49/49 for Library 1. The mis-classified CMGC kinase was C34G6.5 of *C. elegans.* The Library 2 classification suggests that this might be a kinase of the AGC family since the HMM classifying it with the lowest E-value was AGC_sub4.hmm (E-value=5.6e-15), followed by CMGC_sub2.hmm with an E-value of 1.6e-12. Likewise, the AGC kinase that was not classified as an AGC (YL3D4A.6) was, according to Library 2 a CAMK (E-value of CAMK_sub4.hmm 8.7e-40 versus E-value of AGC_sub4.hmm 1.5e-29, the top-matching AGC model). This suggests that CAMK R166.5 might indeed be an AGC (E-value of AGC_sub4.hmm 1.3e-57) and CK1s F26A1.3 and F39F10.2 might be CAMKs (E-values of CAMK_sub7.hmm 3.2e-5 and 4.1e-11, respectively). The division of each family of human kinase catalytic domains into a number of sub-families, followed by their representation in the form of HMMs has been shown to increase the classification rate for KinBase sequences and suggest

possible incorrect annotations for C34G6.5, YL3D4A.6, R166.5, F26A1.3, and F39F10.2 of *C. elegans.*

Table V shows best-case results for Library 2 since they are for organisms against which the sub-families that make up the Library were optimised. In order to obtain a more realistic estimate of the performance of Library 2, it was applied to the kinomes of *M. musculus* (14) and the phylogenetically distant human malaria parasite *P. falciparum* (15) and *D. discoideum* (16). These organisms have had their kinomes annotated and so provide a standard against which to test the approach described here. The results for Library 2 are summarised in Table VI. The *M. musculus* kinome is very similar to that of *H. sapiens* and so the 100% correct classification shown is no surprise. Library 2 was able to classify correctly all protein kinases of *P. falciparum* that had been assigned to families by Ward *et al* (15). In the case of the *Dictyostelium* kinome, Library 2 was able to classify correctly all the kinases of the AGC, CK1, CMGC, and TKL families. The protein kinase DDB0229351 had been classified by Goldberg *et al* as a CAMK kinase (16). However, Library 2 suggests that this sequence would rather belong to the AGC family (E-value for AGC_sub4.hmm of 2.6e-72) than to the CAMK family (E-value for the best hit to a CAMK sub-HMM was 1.6e-67). When the phylogenetic tree provided by Goldberg *et al* was inspected in detail, DDB0229351 was found to be classified as a CAMK because it groups loosely with a cluster of *bona fide* CAMK kinases (the 'FHAK') (16). However, Goldberg *et al* could not provide sound bootstrap support for this association. Eight protein kinases that were classified as STE kinases (16) were not classified as such by Library 2. Examination of the phylogenetic tree provided by Goldberg *et al* (16) showed that all these protein kinases were found to belong to a self-contained group that has some loose relationship to the main group of STE kinases and to which bootstrap support was not provided. The authors of the *Dictyostelium* kinome (16) produced a phylogenetic tree from a multiple alignment of the entire kinome, including a number of representative sequences

from other kinomes. A phylogenetic tree is only as good as the underlying alignment. Given the degree of sequence divergence of the protein kinase superfamily members, the generation of whole kinome multiple alignments is likely to produce trees have little or no bootstrap support for some of the branching patterns observed. Therefore, the generation of protein kinase family-specific multiple alignments is bound to produce more meaningful alignments from which more robust phylogenetic trees can be derived.

In conclusion, Library 2 was able to classify correctly all *bona fide* protein kinases of *M. musculus, P. falciparum,* and *D. discoideum. Dictyostelium* and the malaria parasite represent early branching points in eukaryote evolution (28-30), and the fact that Library 2 can classify their kinomes correctly suggests that it should be generally useful for the classification of eukaryotic protein kinases.

Results and Discussion

A multi-level HMM library (Library 2) has been developed that is able to classify eukaryotic protein kinases into one of the 12 previously established families. The ePK families include the AGC, CAMK, CK1, CMGC, RGC, STE, TK, and TKL, and the *bona fide* aPK families Alpha, PIKK, PDHK, and RIO. As explained in Methods, the reliability of Library 2 has been tested by cross-validation between the kinomes of *H. sapiens, D. melanogaster, C. elegans, S. cerevisiae, M. musculus, D. discoideum,* and *P. falciparum.*

Comparison of HMM Library 2 to BLAST and a general HMM for kinase classification

HMMs are known to be more sensitive and selective than pairwise comparison methods such as BLAST (21). However, since BLAST has been widely applied in kinome analysis, the performance of Library 2 and BLASTP were compared in database searches. For this, a database was assembled that consisted of Uniref100 (10) plus the well-characterised protein kinase complements of *H. sapiens, D. melanogaster, C. elegans,* and *S. cerevisiae* from KinBase. The database was scanned with Library 2 and, for comparison, the domains making up each HMM were also used to search the database by BLASTP (version 2.2.10) with default parameters (matrix=BLOSUM62; cost to extend/open a gap=0; word size=3; but with the 'Filter query sequence' parameter set to F).

For Library 2, the number of true positives (TP) was determined as the number of KinBase sequences returned with scores better than 1e-05. Library 2 was able to retrieve and correctly classify all ePKs of *H. sapiens, C. elegans, D. melanogaster*, and *S. cerevisiae,* respectively (Table VII). The Library 2 database search results illustrate the E-value cutoffs that should be applied to each family of protein kinases in database searches. Over 99% of the proteins from Uniref100 retrieved alongside the KinBase sequences with E-values above the cutoff for each family had been annotated as protein kinases.

As expected, BLASTP searching with the kinase catalytic domains was found to be less sensitive and less specific than the search performed with Library 2. The detailed results of using KinBase human kinase catalytic domains as query sequences to search for KinBase sequences of the four organisms are presented in Table VIII. The typical BLAST behaviour is illustrated by the search performed with the human AGC catalytic domains. Searching with AGC kinase catalytic domains returned an average of 20.44/63 (32.4%) AGC kinases of *H. sapiens* above the default BLAST cutoff E-value of 10.0, including kinases belonging to families other than AGC. The average performance of human AGC kinase domains as query sequences was similarly poor with respect to the AGC kinases of *C. elegans, D. melanogaster,* and *S. cerevisiae*, retrieving on average 6.47/30 (21.57%), 7.13/30 (23.77%) and 4.72/17 (27.76%), respectively, although the average number of false positives was lower in these organisms as smaller families tend to harbour less divergent sequences. The E-values returned by BLAST are higher than those returned by Library 2 for the majority of families, and since BLAST only retrieves a fraction of all existing protein kinases, multiple overlapping BLAST searches would be necessary to arrive at the same result that is provided by Library 2 in a single step. Furthermore, Library 2 produces automatically, a classification of protein kinases that is correct at the family level, whereas the hits returned by BLASTP searches retrieve a mixture of kinases from different families.

In kinome analysis papers, researchers have used a combination of BLAST and general HMMs of the kinase catalytic domain to mine databases for protein kinases. Here, we have compared the database scanning performance of a general, multi-species, HMM of the kinase catalytic domain from Pfam (PF00069, Pfam_fs model) with Library 2. For a full comparison of the database retrieval capacities of the two approaches, we have also included the protein kinases belonging to the 'Other' family of *S. cerevisiae, C. elegans, D. melanogaster* and *H. sapiens.* Table IX shows

the results of retrieving the kinomes of *H. sapiens, C. elegans, D. melanogaster* and *S. cerevisiae* with the Pfam HMM PF00069. Whereas Library 2 could retrieve all protein kinases, PF00069 failed to identify a number of protein kinases, especially those belonging to the 'Other' family. This is not surprising as protein kinases belonging to this family are typically the most divergent ones, making them more difficult to retrieve and classify. In these cases, researchers typically use BLAST for identifying select, remote, homologues (16). Overall PF00069 failed to identify 2.93%, 10.09%, 3.60%, and 2.61% of the kinases of human, worm, fly, and yeast origin. This indicates that Library 2 is superior not only to BLAST but also to a general HMM of the kinase catalytic domain for database searches. In addition, Library 2 is capable of doing automatic classification of protein kinases into families.

Classification of the 'Other' kinases in S. cerevisiae

In any characterised kinome, a group called 'Other' always exists which consists of protein kinase sequences known to belong to the ePKs but which annotators have not been able to classify due to lack of clear similarity to kinases in the 8 ePK families. The current *H. sapiens* kinome annotation (6) contains 83 kinases belonging to the 'Other' group, whereas 45 are found in the *D. melanogaster* kinome (4) , and 38 and 67 in the *S. cerevisiae* (12) and *C. elegans* (11) kinomes, respectively.

In an attempt to classify the 'Other' ePKs of yeast, these were scanned through Library 2, and also their syntenic homologues from the related fungus *Ashbya gossypii*. Although *S. cerevisiae* and *A. gossypii* shared a common ancestor over 100 million years ago, the order and orientation of >95% of *Ashbya's* genes is conserved in the genome of *S. cerevisiae* (31). Therefore, the automatic classification of syntenic homologues from two different species into the same family is an acceptable measure of family membership. Table X shows the classification of the 'Other' kinases of yeast into the main ePK families, together with the classification of the homologous genes from *A. gossypii*. 27/38 yeast kinases of the 'Other' group, together with their corresponding, syntenic,

homologues in *A. gossypii*, were classified by Library 2 automatically into the same family, and above the family-specific E-value cut-off. These 27 kinases of yeast correspond to 23 loci in the ancestral fungal genome. This suggests that, following the whole genome duplication event in the *Saccharomyces* lineage, most duplicate kinase genes, like most duplicate *S. cerevisiae* genes, were lost secondarily (32). However, those pairs of yeast kinases that originated from the same ancestral locus (the pairs are KKQ8 and HAL5 (*Ashbya* AEL118C); PAK1 and TOS3 (*Ashbya ACL053C*); NPR1 and PRR2 (*Ashbya* ABL143C); and PTK1 and PTK2 (*Ashbya* AFR372W), have retained the same family classification for both of the yeast kinases.

The yeast kinase of the 'Other' group VPS15 corresponds with locus ADL316C in *A. gossypii*, and which in *Ashbya* codes for a degenerate, non-functional, form of a kinase catalytic domain. The remaning 10 yeast 'Other' kinases could not be classified with strict confidence into the same ePK family together with their syntenic homologues from *A. gossypii*. Even though these two fungal species shared a common ancestor 100 million years ago, the two have radically different life styles. *A. gossypii* is a fungal pathogen of cotton plants, whereas *S. cerevisiae* has been used as a model organism for decades in the laboratory. The real power of comparative kinomics will come from comparing the kinomes of all yeast species that have been, and are being sequenced. It might be possible that a number of yeast 'Other' kinases constitute the seeds of fungi-specific kinase subfamilies that represent independent clusters themselves.

The application of Library 2 enriched the repertoire of yeast ePKs as follows: AGC (+5), CAMK (+17), CMGC (+4), STE (+1). This has raised the classification rate of yeast ePKs from 77/115 (66.96%) to 104/115 (90.43%). While it is not possible to prove computationally that the suggested new classification for these 27 kinases is correct, the consistency at syntenic loci for the two organisms lends support to the classification. The high degree of selectivity of Library 2 suggests

that it might be useful for reassigning the 'Other' kinases of a number of kinomes. Representative alignments of yeast kinases of the 'Other'group, now re-classified into one of the main ePK families, are provided in Fig. S1.

Classification of the protein kinase complement of 21 eukaryotic genomes

In order to allow a comparison of kinomes between divergent species, Library 2 was used to search for protein kinases in 21 completed and published eukaryotic genomes. These included 2 algal species, *Thalassiosira pseudonana* (33) and *Cyanidioschyzon merolae* (34), the arthropod *Anopheles gambiae* (35), the chordate *Ciona intestinalis* (36), the fishes *Tetraodon negroviridis* (37), and *Takifugu rubripes* (38), 8 fungal species: *Candida glabrata, Kluyveromyces lactis, Debaryomyces hansenii*, and *Yarrowia lipolytica* (39) , *Cryptococcus neoformans* (40), *Neurospora crassa* (41), *Phanerochaete chrysosporium* (42), and *Schyzosaccharomyces pombe* (43); the mammal *Rattus norvegicus* (44); 3 plants: *Arabidopsis thaliana* (45), *Oryza sativa L. ssp. Indica* (46), and *Oryza sativa L. ssp. Japonica* (47); and 3 parasitic protozoans: *Cryptosporidium hominis* (48), *Entamoeba histolytica* (49), and *Plasmodium yoelii* (50) (Table XI). These genomes are representative of a diversity of phylogenetic lineages and display a large variation in their predicted gene numbers, ranging from 3994 for *Cryptosporidium hominis* (an *Apicomplexan* pathogen causing diarrhoea and acute gastroenteritis in humans) to 46000-55000 for *Oryza sativa ssp. indica* (rice).

The set of predicted peptides for each proteome was downloaded from the sources indicated in Table XI. Library 2 was then applied uniformly to the 21 genomes, with an E-value cutoff specific for each family (Table VII). The 'Other' kinases of the characterised kinomes of *H. sapiens, D. melanogaster, C. elegans, S. cerevisiae, M. musculus, P. falciparum, S. cerevisiae* and *A. gossypii* were also classified by Library 2 into the main ePK families. This was done in order to compare the

number of kinases *per* family across organisms.

If we accept the classification as presented here (Table XII), the 21-genome kinase-family annotation showed that 5 ePK families (AGC, CAMK, CK1, CMGC, STE) and 2 aPK families (RIO, PIKK) were present in all eukaryotic organisms whose genomes have been sequenced. As a consequence, these 7 kinase families are likely to be indispensable to eukaryotic life. The other families (TK, TKL, Alpha, PDHK, RGC) appear to be restricted to specific lines of descent, as discussed below.

TKs

In metazoans, tyrosine phosphorylation plays essential roles in intercellular communication (organ development, tissue homeostasis) as well as intracellular communication (transcriptional control, proliferative/differentiation decisions, cell shape, and cell motility). Abnormal levels of tyrosine phosphorylation result in conditions such as cancer and immune diseases (51-54). TKs were absent from fungi, and were mainly found in metazoan species, in agreement with their role. Interestingly, the three plant species seemed to encode putative tyrosine kinases: 2 in *A. thaliana*, and 6 and 7 in the two rice species. When looked at in detail, the two putative tyrosine kinases of *A. thaliana* lack the conserved lysine residue of subdomain II and the conserved aspartate of the activation loop in subdomain VII (5). This suggests that these two enzymes are likely to be catalytically inactive. The two rice species were found to encode 6 and 7 putative tyrosine kinases. Out of the 6 predicted tyrosine kinases of *O. sativa ssp. Indica* , only 3 of them were found to include all the residues known to be essential for catalytic activity. Out of the 7 putative tyrosine kinases of *O. sativa ssp. Japonica,* 4 were found to harbour all residues known to be required for catalytic activity. To our knowledge, TKs have never been predicted in plants before (55), although the existence of tyrosine phosphorylation in plants has been previously documented. Tyrosine phosphorylation in plants is

thought to be involved in stress-related responses, embryogenesis, and tissue differentiation (56), and in response to movement (57). The small number of putative tyrosine kinases identified in the plant species does not necessarily mean that tyrosine phosphorylation in plants is a limited phenomenon, but that tyrosine phosphorylation as carried out by putative tyrosine kinases is probably more limited than that carried out by dual-specificity kinases in plants. A recent *in silico* survey of the *Arabidopsis* proteome suggested that this plant lacks *bona fide* tyrosine kinases, and that tyrosine phosphorylation in plants might be carried out by dual-specificity protein kinases (55).

The human intestinal parasite *Entamoeba histolytica* encodes 7 putative tyrosine kinases, which had already been noted (49, 58). The authors identified 270 ePKs, including TKLs, and at least 90 putative receptor serine/threonine kinases, which are uncommon in protists and usually only found in plants and animals. Library 2 predicted a total of 295 ePKs for *E. histolytica*. This makes *E. histolytica* the single-celled eukaryote with the second largest kinome after *Trichomonas vaginalis* (manuscript in preparation).

The plant TKs, together with the 7 putative tyrosine kinases of *E. histolytica* identified here, and the absence of TKs among the fungal species surveyed in this study, suggest that a primitive set of tyrosine kinases might have pre-dated the radiation of eukaryotic life forms (28). The number of tyrosine kinases per predicted gene complement is smaller in the two plant species and *E. histolytica* than in the arthropods *Drosophila* and *Anopheles*, which harbour 33 and 32 tyrosine kinases, respectively. The fishes *T. nigroviridis* and *T. rubripes* have ca. 2.5 times as many putative tyrosine kinases as does the chordate *Ciona intestinalis*. Whole genome duplication events are known to have occurred in the teleost fish lineage after its divergence from the lineage leading to mammals (37), hence explaining why the two fish species have more TKs than humans do (despite arguably being less complex phenotypically). Receptor and non-receptor tyrosine kinases have

previously been found in choanoflagellates (59, 60). Choanoflagellates are unicellular protozoa related to metazoans and which have been studied in detail to try to explain the molecular innovations that gave rise to animal multicellularity from an ancestral protozoan. Tyrosine kinases have adopted a prominent role in cellular signalling in metazoans when compared to non-metazoans. The radiation of the tyrosine kinase family suggests that this family has progressively become adapted to fulfilling sophisticated signalling roles in the more complex animals.

Phosphotyrosine signalling in evolution

The presence of putative tyrosine kinases in *Entamoeba* and the plant species prompted us to search for further evidence of tyrosine phosphorylation in these organisms. Strong support for tyrosine phosphorylation mechanisms may be provided by the presence of phosphotyrosine-binding domains, such as SH2 and PTB, and the existence of tyrosine phosphatases. Phosphotyrosines are known to bind target proteins via their SH2 or PTB domains to form multiprotein signalling complexes. SH2 domains have previously been described in plants (2 in *A. thaliana*, 1 in *O. sativa*). Plant SH2 domains were found to be shorter than their animal counterparts, but seemed to fold into a functional phosphotyrosine binding domain (61). A search with a Pfam HMM specific for the SH2 domain (PF00017) (18) identified the same SH2 domains as described before for the plant species. Interestingly, *Entamoeba* was found to harbour 4 polypeptide sequences with SH2 domains, three of which were found to be protein kinases. This comes in contrast with the plant SH2-containing polypeptide chains, where the regions outwith their SH2 domains seem to have no inferable function (61).

PTB domains are also known to bind phosphotyrosines (62). However, a search with a Pfam HMM specific for the PTB domain (PF08416) returned no plausible candidates.

A third type of phosphotyrosine-binding domain has recently been characterised, the C2 domain of human protein kinase C delta (63). Crystallographic analysis showed that the phosphotyrosine binding ability of the C2 domain is dependent on a number of key residues. A search with an HMM specific for the C2 domain (PF00168) returned 25 proteins containing C2 domains in *Entamoeba* (2 of which have double C2 domains). *A. thaliana, O. sativa ssp. Indica* and *O. sativa ssp. Japonica* were found to possess 85, 103, and 106 C2 domain-containing proteins, respectively. However, whereas no *Arabidopsis* protein contains more than one C2 domain, 19 and 20 proteins in the *O. sativa Indica* and *Japonica* species harbour two, three, or four C2 domains. None of the C2 domains of the plant species were found in the predicted protein kinases, whereas one C2 domain of *Entamoeba* was found N-terminal to the kinase domain of the *Dictyotelium* homologue of Myosin Light Chain Kinase (MLCK). *Dictyostelium's* MLCK (UniProt Q54W26) lacks the C2 domain N-terminal to the kinase catalytic domain.

The phosphotyrosine binding ability of the C2 domain of human PKCdelta was found to be dependent on a number of key residues (63). These residues are found in an internal extension of the C2 domain that appears to be particular to PKCs. This extension was not found in the plant or amoeba C2 domains retrieved. It remains to be determined whereas these residues are critical and confer the phosphotyrosine binding ability to PKCdeltas or whether the C2 domain has a general phosphotyrosine binding ability that can be carried out by alternative residues.

Dual-specificity phosphatases and, most importantly, tyrosine phosphatases are strong indicators of a phosphotyrosine signalling system. A search with an HMM specific for dual-specificity phosphatases (PF00782) retrieved 15 candidates in *Entamoeba* and 6 in each of the plant species which could align meaningfully with previously characterised dual-specificity phosphatases (Fig.

4). A search with an HMM specific for tyrosine phosphatases (PF00102) returned 2 candidates in *Entamoeba*, 1 in *A. thaliana* and 2 in each of the *O. sativa* species (Fig. 5).

The presence of putative tyrosine kinases, phosphotyrosine-binding domains as well as tyrosine phosphatases in *Entamoeba* and the 3 plant species lends support to the concept that phosphotyrosine signalling in eukaryotes appeared early in evolution, and is not a phenomenon exclusive to metazoans (61). The ancestors of modern vascular plants diverged before the lineage leading to animals, fungi and *Dictyostelids* did (29, 30). However, plants are known to carry out many of their signalling needs by means of serine/threonine receptor kinases, whereas receptor tyrosine kinase signalling appears to be more widespread in metazoans. It seems that plants and animals have solved their signalling needs according to similar yet distinct strategies. Still, the early appearance of putative phosphotyrosine signalling in evolution suggests that this might be an important, conserved, mechanism in plants and a number of plant-related species of ecological and economic importance such as red algae, chlorophytan green algae, and mosses.

TKLs

The TKLs are present in only 2 of the fungal species: *C. neoformans* (2 TKLs), and *P. chrysosporium* (4 TKLs). The new observation made on the existence of putative TKLs in only 2 fungal species: *C. neoformans* (2 TKLs) and *P. chrysosporium* (4 TKLs); in the Apicomplexa *C. hominis* (4 TKLs), *P. yoelii* (4 TKLs), *P. falciparum* (6 TKLs); the amoeba *E. histolytica* (109); and in the algal species *C. merolae* (9 TKLs) and *T. pseudonana* (6 TKLs), suggests that TKLs might have been present at the very beginning of eukaryotic evolution. Although TKLs appear to be absent from most fungal species, it is reasonable to think that they were secondarily lost upon divergence and speciation. TKLs were found to be especially abundant in plants, with 776 TKLs in *A. thaliana* and around one thousand TKLs being encoded in the genomes of the two rice

species, although its significance remains unknown. The unusually large number of TKLs found in the *E. histolytica* genome (109) relative to the number of genes encoded in its genome (9938), probably represents an extreme metabolic adaptation of this single-celled organism to enhance its virulence in the environments it encounters.

The understanding, both functional and evolutionary, of the primitive TKs and TKLs should shed light on their early functions and mechanisms of regulation, and will aid understanding of their subsequent diversification and incorporation into more sophisticated control circuits in animals.

aPKs

Among the four families of aPKs, PIKKs appear in all the organisms examined here with no more than 6 PIKKs being found per genome. Their presence is indicative of their central importance in sensing DNA and RNA damage (64, 65). The PDHKs are present in all the organisms examined here with the exception of the Apicomplexan species, which possess divergent, tubular-shaped, mitochondria (66). The PDHKs are known to regulate the pyruvate dehydrogenase complex (PDC), which catalyses the oxidative decarboxylation of pyruvate, linking the degradation of intracellular glycogen and extracellular glucose via glycolysis to the Krebs cycle (67). Genes encoding enzymes of the TCA cycle have been predicted *in* silico for *P. falciparum*, but the parasite mitochondrion is likely to be divergent from mammalian mitochondria in other respects (66). The RIO kinases, which seem to share the basic structural elements of the ePK fold (8) and whose function remains unknown, are present in all the eukaryotic lineages examined here, suggesting an essential function in eukaryotes. The copy number of RIO kinases per genome is never greater than three. The Alpha kinases, which also seem to share the ePK fold (68), were absent from plants, one of the algal species (*C. merolae)*, most fungi (except for 2 members in *N. crassa*) and from the Apicomplexa, but were present in *Entamoeba histolytica*.

The results presented in Table XII also allow the comparison of protein kinase families in organisms that are related phylogenetically. Fungi are believed to have appeared approximately one billion years ago, and the divergence of the Basidiomycota and Ascomycota probably occurred ca. 968 million years ago (30). The group of fungi examined here encompasses species from different lineages. Although the predicted gene number of the various fungal species ranges from 5283 for *C. glabrata* to 11777 for *P. chrysosporium*, it is remarkable that the number of protein kinases in most families of ePKs is rather similar in most instances: around 20 for the AGCs, 2-5 for the CK1s, around 25 for the CMGCs, and around 14 for the STEs. The CAMKs of fungi present a wider range from 16 for *C. neoformans* to 44 for *S. cerevisiae.* Some fungal lineages are known to have undergone whole genome duplication events, followed by massive gene loses (32). Therefore, direct comparisons between the protein kinase complements of the various fungal species is not strictly correct. However, the fact that some ePK families in fungi tend to be more compact than others in numbers might be an indication that in fungi some protein kinase families tolerate gene duplications better than others, and that novel genes might be allowed to co-exist with the old copies, either contributing to the function of the former by being incorporated to existing signal transduction pathways without detrimental effects, or by evolving new functions (neo-functionalisation). It will be interesting to perform phylogenetic analysis of the fungal kinomes to determine what extent of each kinome is present in all fungi, which kinases are lineage-specific, and also whether the homologous kinase genes have different combinations of transcription factor binding sites that would promote differential transcription of homologous kinase genes, thereby accounting for the vast range of metabolic adaptations displayed by the fungal species considered here.

The difficulty in producing a catalogue of the major families of eukaryotic protein kinases for an

organism is highlighted by the lack of discussion of kinase-mediated signal transduction pathways in all but the genome papers of *E. histolytica, A. thaliana, N. crassa,* and *F. rubripes.* Even though the genome of *A. thaliana* has been carefully annotated, the knowledge website (http://www.arabidopsis.org) features only 1031 open-reading frames annotated automatically as protein kinases. Application of Library 2 predicts that this plant contains 1301 putative protein kinases. The plant species examined here were found to have an unusually large number of kinases in comparison with, for example, the human genome. It must be taken into account that reported whole genome duplication events in plants (45) prevent the meaningful comparison between genome size, phenotypic complexity, and kinome size.

At the time of the publication of the *Arabidopsis* genome, the *ab initio* gene models were validated with known ESTs and protein sequences, the researchers finding that 93% of ESTs of *Arabidopsis* matched gene models, with less than 1% of ESTs matching non-coding regions (45). With the rice species, only 61% of *ab initio* gene models were found to have a high identity match with rice ESTs or full-length cDNAs. Whereas 71% of predicted rice proteins were found to have a homologue in the *Arabidopsis* proteome, 89.8% of the proteins from the *Arabidopsis* genome have a homologue in the rice proteome (69). This suggests that caution must be exercised when considering the gene models of the rice species that do not have EST coverage or cannot be verified otherwise. Still, the predicted putative ePKs for the rice species (1276 and 1330, Table XII) agree with those predicted by the authors (between 1075-1425, depending on the InterPro signature used, (69)).

The use of Library 2 provides a starting point for characterising the kinomes of organisms that are representative of key points in the evolutionary history of eukaryotes. Detailed kinase family phylogenetic analyses will help identify orthologous sub-families across the vast spectrum of eukaryotic life. This should illuminate the early stages of protein kinase evolution and help

understanding of how gene loss, duplication, and innovation within each eukaryotic lineage correlate with specific functions. Protein kinase orthologue and paralogue identification is also an essential step in the process of drug target identification for establishing assay and animal models. Paralogues are known to introduce complications such as selectivity issues, pleiotropy, and functional redundancy of targets, all of which are critical to assessing the druggability of a particular protein kinase. This wealth of detailed kinomes will help extend Library 2 to classify automatically and reliably to the sub-family and sub-sub-family level. This will be particularly important for the larger families (e.g. CAMK), where it could help identify further sub-families in their own right which might have been obscured hitherto by the limited number of kinomes available. Having enough information on what the key residues/regions of a particular sub-family are, coupled to structural information, will help highlighting those regions of the structure that are important for the sub-family, but which might not be captured appropriately by the HMM.

Library 2 has been used to assist in the analysis of the kinomes of the human parasitic protozoan *Trichomonas vaginalis* and the human parasitic nematode *Brugia malayi* (manuscripts in preparation). A server is being developed that will integrate the ability to scan peptide sequences with Library 2, and display precalculated analyses of the kinomes of a number of species of biological, medical, and economic interest. This facility will be available under the URL: http://www.compbio.dundee.ac.uk/kinomes/. Library 2 is available from the authors upon request.

Conclusions

In this paper, a sensitive library of HMMs (Library 2) has been developed to identify and sub-classify protein kinase catalytic domains into one of the 12 accepted families. The main conclusions of the work are:

1. Library 2 showed a protein kinase family misclassification rate of zero on the characterised kinomes of *H. sapiens, M. musculus, D. melanogaster, C. elegans, S. cerevisiae, P. falciparum,* and *D. discoideum.* Library 2 was able to retrieve all protein kinases from KinBase in a database search. In contrast, BLASTP could only retrieve on average 1/3 of protein kinases from the same database. The general kinase domain HMM PF00069 failed to identify between 2.61-10.09% of the ePKs of characterised kinomes. These properties make Library 2 a useful tool for database searching of kinases and their automatic classification into families.

2. 27/38 'Other' kinases of yeast, and their syntenic homologues in the fungus *A. gossypii,* were classified using Library 2 into the main ePK families; AGC (5), CAMK (17), CMGC (4), and STE (1), raising the family-level classification rate of the yeast ePKs from 66.96% to 90.43%.

3. The application of Library 2 to 21 eukaryotic genomes showed that 5 five ePK (AGC, CAMK, CK1, CMGC, STE) and 2 aPK (PIKK, RIO) protein kinase families were present in all eukaryotic genomes analysed. These seven families are likely to be indispensable to all eukaryotic life forms. The families present in specific lines of descent (RGC, TK, TKL, Alpha, PDHK) are likely to be late innovations. Alternatively, their absence from a particular phylogenetic group might indicate that they have been lost secondarily.

4. No TKs were found in fungi. The plants *A. thaliana, O. sativa ssp. Indica,* and *O. sativa ssp. Japonica* were found to encode 2, 6, and 7 putative TKs, respectively. To our knowledge, TKs have

never been predicted in plants before, although a tyrosine phosphorylation system is known to operate in *A. thaliana.*

5. The confirmation of the presence of 7 putative TKs in *Entamoeba histolytica*, together with the putative TKs in plants, suggests that a primitive set of TKs might have predated the radiation of eukaryotic organisms.

6. The fungi *C. neoformans* and *P. chrysosporium* were found to encode 2 and 4 putative TKLs, respectively. Putative TKLs were also found in the Apicomplexa *C. hominis* (4), *P. yoelii* (4), *P. falciparum* (6), the amoeba *E. histolytica* (109), and in the algae *T. pseudonana* (6) and *C. merolae* (9). TKLs were found to be especially abundant in plants: *A. thaliana* (776), *O. sativa ssp. Indica* (1010), *O. sativa ssp. Japonica* (969). These data suggest that a) TKLs predated the radiation of eukaryotic life forms; b) TKLs in plants probably carry out important and specific functions to plants; c) TKLs have been lost secondarily in some fungi.

7. The atypical protein kinase family PIKK is ubiquitous in all eukaryotic genomes examined here, with never more than 6 PIKKs per genome. Their universal presence reflects their importance in sensing DNA and RNA damage. RIO kinases were also found to be ubiquitous, although their biological substrates remain unknown.

Acknowledgements

The authors thank Professor Grahame Hardie and Dr. David Martin for many insightful discussions, and Dr. Jonathan Monk for computer support. The authors also wish to thank the reviewers for their insightful comments. This work has been funded by the Wellcome Trust under a Prize Studentship to D. M.-S. and G.J.B.

Tables

Table I.

Family	*H. sapiens*	*C. elegans*	*D. melanogaster*	*S. cerevisiae*	*M. musculus*	*D. discoideum*	*Total*
AGC	63	30	30	17	60	22	222
CAMK	74	40	32	21	96	21	284
CK1	12	85	10	4	11	3	125
CMGC	61	49	33	21	60	30	254
RGC	5	27	6	N/A	7	N/A	45
STE	47	25	18	14	47	45	196
TK	90	88	31	N/A	90	N/A	299
TKL	43	15	17	N/A	43	68	186
Total	**395**	**359**	**177**	**77**	**414**	**189**	**1611**

Table II.

Organism/Family	*Protein length*	*KinBase Accession Nos.*	*Swiss-Prot Accession Nos.*	*Notes*
***S. cerevisiae*/PIKK**	2368	6319612_MEC1	P38111	ESR1 protein
***S. cerevisiae*/PIKK**	2787	TEL1	P38110	Telomer length regulation protein TEL1
***S. cerevisiae*/PIKK**	2470	TOR1	P35169	PI3K-related kinase TOR1
***S. cerevisiae*/PIKK**	2473	TOR2	P32600	PI3K-related kinase TOR2
***S. cerevisiae*/PIKK**	3744	TRA1	P38811	Transcription-associated protein 1
***D. melanogaster*/PIKK**	3218	CG2905	Q8I8U7	DTRA1 (Transcription-associated protein 1). CG2905 and Q8I8U7 are isoforms.
***D. melanogaster*/PIKK**	2429	CG4549	Q9W3V6	Smg-1
***D. melanogaster*/PIKK**	2470	CG6535	Q9VFB1	CG6535-PA
***D. melanogaster*/PIKK**	2354	Tor	Q9VK45	Tor
***C. elegans*/PIKK**	2697	mei-41	Q9VXG8	mei-41
***C. elegans*/PIKK**	3944	B0261.2	Q95Q95	Target of rapamycin homolog (CeTOR), gene let-363
***C. elegans*/PIKK**	2531	C47D12.1	Q6A4L2	Hypothetical protein C47D12.1b, gene trr-1
***C. elegans*/PIKK**	649	atl-1	Q22258	Hypothetical protein T06E4.3a
		atm-1	Q9N3Q4	ATM family protein 1

Organism/Family	***Protein length***	***KinBase Accession Nos.***	***Swiss-Prot Accession Nos.***	***Notes***
***C. elegans*/PIKK**	2327	Smg-1	O01510	Smg-1 (suppressor with morphological effect on genitalia protein 1). KinBase smg-1 and O01510 are isoforms.
***M. musculus*/PIKK**	2635	ATR	Q9JKK8	No direct correspondence. Q9JKK8 is a fragment of KinBase ATR.
***M. musculus*/PIKK**	4128	DNAPK	P97313	DNA-dependent protein kinase.
***M. musculus*/PIKK**	2549	FRAP	Q9JLN9	FKBPI2-rapamycin complex-associated protein
***M. musculus*/PIKK**	3658	SMG1	Q8BLU4	smg-1
***M. musculus*/PIKK**	3877	TRRAP	Q80YV3	Transformation/Transcription-associated protein. TRRAP and Q80YV3 are isoforms.
***H. sapiens*/PIKK**	3056	ATM	Q13315	ATM ser/thr protein kinase
***H. sapiens*/PIKK**	2644	ATR	Q13535	ATR
***H. sapiens*/PIKK**	4128	DNAPK	P78527	DNA-dependent protein kinase
***H. sapiens*/PIKK**	2549	FRAP	P42345	FKBP12-rapamycin complex-associated protein
***H. sapiens*/PIKK**	3657	SMG-1	Q96Q15	PI3K-related protein kinase smg-1
***H. sapiens*/PIKK**	3877	TRRAP	Q9Y4A5	Transformation/Transcription domain-associated protein
S. cerevisiae/RIO	484	6324693_RIO1	Q12196	Ser/Thr protein kinase RIO1
S. cerevisiae/RIO	425	6324122_RIO2	P40160	Ser/Thr protein kinase RIO2
D. melanogaster/RIO	585	CG11660	Q9VTL5	CG11660-PA, isoform A
D. melanogaster/RIO	538	CG11859	Q9UBU2	CG11859-PA
D. melanogaster/RIO	603	CG3008	Q9VR42	CG3008-PA
C. elegans/RIO	548	M01B12.5	O44959	Hypothetical protein M01B12.5
C. elegans/RIO	510	ZK632.3	P34649	Putative RIO-type ser/thr

Organism/Family	***Protein length***	***KinBase Accession Nos.***	***Swiss-Prot Accession Nos.***	***Notes***
				protein kinase
C. elegans/RIO	529	aSWK440	Q95Q34	Hypothetical protein Y105E8B.3
M. musculus/RIO	567	RIOK1	Q9CU84	RIOK1 ser/thr protein kinase
M. musculus/RIO	547	RIOK2	Q9CQS5	RIOK2 ser/thr protein kinase
M. musculus/RIO	519	RIOK3	Q9DBU3	RIOK3 ser/thr protein kinase
H. sapiens/RIO	568	RIOK1	Q9BRS2	RIO1 ser/thr protein kinase
H. sapiens/RIO	552	RIOK2	Q9BVS4	RIO2 ser/thr protein kinase
H. sapiens/RIO	519	RIOK3	O14730	RIO3 ser/thr protein kinase
***S. cerevisiae*/PDHK**	445	6321379_YGL059W	P53170	Hypothetical 51.9 Kda protein in PYC-UBC2 intergenic region
***S. cerevisiae*/PDHK**	394	6322147_YIL042C	P40530	Hypothetical 45.4 Kda protein in CBR5-NOT3 intergenic region
***D. melanogaster*/PDHK**	413	PDK	P91622	Pyruvate dehydrogenase kinase
***C. elegans*/PDHK**	401	ZK370.5	Q02332	Probable PDHK, mitochondrial precursor
***M. musculus*/PDHK**	412	BCKDK	O55028	BCKDK, mitochondrial precursor
***M. musculus*/PDHK**	434	PDHK1	Q8BFP9	PDHK, isozyme 1, mitochondrial precursor
***M. musculus*/PDHK**	407	PDHK2	Q9JK42	PDHK2, mitochondrial precursor
***M. musculus*/PDHK**	415	PDHK3	Q922H2	PDHK3, mitochondrial precursor
***M. musculus*/PDHK**	412	PDHK4	O70571	PDHK4, mitochondrial precursor
***H. sapiens*/PDHK**	412	BCKDK	O14874	BCKDK, mitochondrial precursor

Organism/Family	***Protein length***	***KinBase Accession Nos.***	***Swiss-Prot Accession Nos.***	***Notes***
***H. sapiens*/PDHK**	436	PDHK1	Q15118	PDHK, isozyme 1, mitochondrial precursor
***H. sapiens*/PDHK**	407	PDHK2	Q15119	PDHK, isozyme 2, mitochondrial precursor
***H. sapiens*/PDHK**	406	PDHK3	Q15120	PDHK, isozyme 3, mitochondrial precursor
***H. sapiens*/PDHK**	411	PDHK4	Q16654	PDHK, isozyme 4, mitochondrial precursor
C. elegans/Alpha	760	efk-1	O01991	Elongation factor-2 kinase
M. musculus/Alpha	1862	ChaK1	Q923J1	TRMP7
M. musculus/Alpha	2028	ChaK2	Q8CIR4	TRMP6
M.musculus/Alpha	724	eEF2K	O08796	Elongation factor-2 kinase
*H. sapien*s/Alpha	1907	AlphaK1	Q96L96	Muscle Alpha-kinase
H. sapiens/Alpha	1531	AlphaK2	Q96L95	Heart Alpha-kinase
H. sapiens/Alpha	1244	AlphaK3	Q96QP1	Lymphocyte Alpha-kinase
H. sapiens/Alpha	1865	ChaK1	Q96QT4	TRPM7 human
H. sapiens/Alpha	2012	ChaK2	Q9BX84-2	TRPM6
H. sapiens/Alpha	725	eEF2K	O00418	Elongation factor-2 kinase

Table III.

H. sapiens-based models	***C elegans***			***H. sapiens***			***D. melanogaster***			***S. cerevisiae***		
Family	**No.K**	**No.CC**	**%CC**	**No.K**	**No.CC**	**%CC**	**No.K**	**No.CC**	**%CC**	**No.K**	**No.CC**	**%CC**
AGC	30	27	90.00	63	62	98.41	30	30	100	17	16	94.12
CAMK	40	36	90.00	74	74	100	32	31	96.87	21	20	95.24
CK1	85	83	97.65	12	12	100	10	10	100	4	4	100
CMGC	49	49	100	61	61	100	33	33	100	21	21	100
RGC	27	27	100	5	5	100	6	6	100	N/A	N/A	N/A
STE	25	22	88.00	47	47	100	18	18	100	14	14	100
TK	88	88	100	90	90	100	31	31	100	N/A	N/A	N/A
TKL	15	6	40.00	43	43	100	17	16	94.12	N/A	N/A	N/A
Total (ePKs)	**359**	**338**	**94.15**	**395**	**394**	**99.77**	**177**	**175**	**98.87**	**77**	**75**	**97.40**
Alpha	1	1	100	6	6	100	N/A	N/A	N/A	N/A	N/A	N/A
PIKK	5	5	100	6	6	100	5	5	100	5	5	100
RIO	1	1	100	5	5	100	1	1	100	2	2	100
PDHK	3	3	100	3	3	100	3	3	100	2	2	100
Total (aPKs)	**10**	**10**	**100**	**20**	**20**	**100**	**9**	**9**	**100**	**9**	**9**	**100**

Table IV.

Family	*Number of catalytic domain sequences*	*Z-score cutoff*	*Number of groups generated for each family*
AGC	78	19.5	5
CAMK	79	19.5	7
CK1	12	20	2
CMGC	61	19	4
RGC	5	N/A	1
STE	47	19	4
TK	102	19	6
TKL	43	19	5
Total ePKs	**407**		**34**
Alpha	6	N/A	1
PIKK	5	N/A	1
PDHK	5	N/A	1
RIO	3	N/A	1
Total aPKs	**20**		**4**

Table V.

Family	*H. sapiens*			*D. melanogaster*			*C. elegans*			*S. cerevisiae*		
	No.K	**No.CC**	**%CC**	**No.K**	**No.CC**	**%CC**	**No.K**	**No.CC**	**%CC**	**No.K**	**No.CC**	**%CC**
AGC	63	63	100	30	30	100	30	29	96.67	17	17	100
CAMK	74	74	100	32	32	100	40	39	97.50	21	21	100
CK1	12	12	100	10	10	100	85	83	97.65	4	4	100
CMGC	61	61	100	33	33	100	49	48	97.96	21	21	100
RGC	5	5	100	6	6	100	27	27	100	N/A	N/A	N/A
STE	47	47	100	18	18	100	25	25	100	14	14	100
TK	90	90	100	31	31	100	88	88	100	N/A	N/A	N/A
TKL	43	43	100	17	17	100	15	15	100	N/A	N/A	N/A
Total	395	395	100	177	177	100	359	354	98.61	63	63	100

Table VI.

Family	*M. musculus*			*P. falciparum*			*D. discoideum*		
	No.K	**No.CC**	**%CC**	**No.K**	**No.CC**	**%CC**	**No.K**	**No.CC**	**%CC**
AGC	60	60	100	5	5	100	22	22	100
CAMK	97	97	100	13	13	100	21	20	95.24
CK1	11	11	100	1	1	100	3	3	100
CMGC	60	60	100	18	18	100	30	30	100
RGC	7	7	100	N/A	N/A	N/A	N/A	N/A	N/A
STE	47	47	100	N/A	N/A	N/A	45	37	82.22
TK	91	91	100	N/A	N/A	N/A	N/A	N/A	N/A
TKL	44	44	100	5	5	100	68	68	100
Total	417	417	100	42	42	100	189	180	95.24

Table VII.

	H. sapiens			*C. elegans*			*D. melanogaster*			*S. cerevisiae*		
Family	No.K	TP	E-value	No.K	TP	E-value	No.K	TP	E-value	No.K	TP	E-value
AGC	63	63	8.50e-014	30	30	2.7e-07	30	30	5.4e-09	17	17	1.6e-58
CAMK	74	74	2.2e-16	40	40	5.6e-24	32	32	9e-15	21	21	3.2e-14
CK1	12	12	5.8e-165	85	85	3.2e-05	10	10	1.2e-08	4	4	4.3 e-126
CMGC	61	61	5.6e-108	49	49	6.7e-12	33	33	7.1e-39	21	21	1.2e-07
RGC	5	5	7.5e-168	27	27	4.8e-05	6	6	4.9e-59	0	0	N/A
STE	47	47	1.6e-98	25	25	1.4e-06	18	18	2.6e-13	14	14	8.2e-56
TK	90	90	7.3e-89	88	88	1.1e-09	31	31	7.7e-13	0	0	N/A
TKL	43	43	1.7e-60	15	15	1.7e-12	17	17	1.4e-14	0	0	N/A
Total	395	395/395 (100%)		359	359/359 (100%)		177	177/177 (100%)		77	77/77 (100%)	

Table VIII.

	H. sapiens			*C. elegans*			*D. melanogaster*			*S. cerevisiae*		
Kinase family	No.K	TP / FP	E-value	No.K	TP / FP	E-value	No.K	TP / FP	E-value	No.K	TP / FP	E-value
AGC	63	20.44 / 3.88	6e-15	30	6.47 / 0.68	5e-15	30	7.13 / 0.96	4e-15	17	4.72 / 0.26	2e-15
CAMK	74	20.03 / 1.32	2e-1	40	6.26 / 0.16	5e-16	32	6.70 / 0.16	1e-15	21	2.36 / 0.03	1e-17
CK1	12	11.08 / 0.00	2e-04	85	48.00 / 0.08	2e-04	10	8.83 / 0.00	7e-08	4	4.00/ 0.00	5e-08
CMGC	61	18.02 / 0.00	4e-07	49	7.89 / 0.00	3e-07	33	8.15 / 0.00	1e-08	21	3.84 / 0.00	1e-07
RGC	5	5.00 / 11.40	2e-20	27	21.80 / 2.20	1e-11	6	6.00 / 1.60	1e-23	0	0.00 / 0.00	N/A
STE	47	22.38 / 0.36	2e-16	25	6.40 / 0.04	5e-18	18	7.55 / 0.09	3e-16	14	3.83 / 0.04	5.00E-015
TK	90	19.34 / 2.19	4e-15	88	2.99 / 0.68	3e-15	31	5.15 / 0.69	2e-15	0	0.00 / 0.07	N/A
TKL	43	10.42 / 6.98	2e-09	15	2.98 / 2.16	2e-09	17	3.86 / 2.21	2e-09	0	0.00 / 0.23	N/A
Total	395			359			177				77	

Table IX.

	H. sapiens			*C. elegans*			*D. melanogaster*			*S. cerevisiae*		
Kinase Family	*No.K*	*No.missed*	*% missed*	*No.K*	*No.missed*	*% missed*	*No.K*	*No.missed*	*% missed*	*No.K*	*No.missed*	*% missed*
AGC	63	0	0	30	0	0	30	0	0	17	0	0
CAMK	74	0	0	40	1	2.50	32	0	0	21	0	0
CK1	12	1	8.33	85	13	15.29	10	1	10	4	0	0
CMGC	61	0	0	49	0	0	33	0	0	21	0	0
RGC	5	0	0	27	4	14.81	6	0	0	0	0	0
STE	47	0	0	25	0	0	18	0	0	14	0	0
TK	90	0	0	88	1	1.14	31	0	0	0	0	0
TKL	43	0	0	15	0	0	17	0	0	0	0	0
Other	83	13	15.66	67	24	35.82	45	7	15.55	38	3	7.89
Total	478	14	2.93%	426	43	10.09%	222	8	3.60%	115	3	2.61%

Table X.

Yeast protein kinase of the 'Other' group	*Locus ID on the genome of S. cerevisiae*	*Syntenic locus on the genome of A. gossypii*	*New family(re-classification)*
BUB1	YGR188C	AGR315C	AGC
CDC5	YMR001C	ACL006W	AGC
IRE1	YHR079C	ADR293C	AGC
IPL1	YPL209C	AFL101C	AGC
YKL171W	YKL171W	AEL120W	AGC
APG1	YGL180W	ACL054W	CAMK
KKQ8	YKL168C	AEL118C	CAMK
HAL5	YJL165C	AEL118C	CAMK
IKS1	YJL057C	AEL173W	CAMK
ISR1	YPR106W	AEL330C	CAMK
PAK1	YER129W	ACL053C	CAMK
TOS3	YGL179C	ACL053C	CAMK
NPR1	YNL183C	ABL143C	CAMK
PRR2	YDL214C	ABL143C	CAMK
KSP1	YHR082C	ADR300C	CAMK
PTK1	YKL198C	AFR372W	CAMK
PTK2	AFR372W	AFR372W	CAMK
SAT4	YCR008W	AER195C	CAMK
YDL025C	YDL025C	ADR313W	CAMK
YGR052W	YGR052W	AEL284C	CAMK
YOR267C	YOR267C	ADR174C	CAMK
YPL236C	YPL236C	AFL143C	CAMK
CDC7	YDL017W	AER216C	CMGC
CKA1	YIL035C	ADL102C	CMGC
CKA2	YOR061W	ADR204W	CMGC
MPS1	YDL028C	ADR317C	CMGC
SWE1	YJL187C	AEL149C	STE

Table XI.

Taxonomy and Species (a)	***Common name/description***	***Source Database (b)***	***Release (c)***	***Haploid Genome Size (Mb)***	***Predicted number of genes***
Fungi					
C. glabrata	Pathogen causing human candidiasis	Genolevures	N/A	12.3	5283
C. neoformans	Basidiomycetous yeast	TIGR	N/A	19	6572
D. hansenii	Halotolerant yeast	Genolevures	N/A	12.2	6906
K. lactis	Hemiascomycete fungus	Genolevures	N/A	10.6	5329
N. crassa	Multic. Filamentous fungus	BROAD-MIT	7	40	10082
P. chrysosporium	White rot fungus	JGI	1.0	30	11777
S. pombe	Fission yeast	Sanger	N/A	13.8	4940
Y. lipolytica	Alkane-using yeast	Genolevures	N/A	20.5	6703
Animals					
A. gambiae	Malaria mosquito	TIGR	8	278	14000
C. intestinalis	Ascidian tadpole	JGI	1.0	160	15852
T. nigroviridis	Freshwater pufferfish	BROAD-MIT	Data version 10/31/01	381	27918
T. rubripes	Pufferfish	JGI	3.0	365	40000
R. norvegicus	'Lab rat'	TIGR	10	2.75 Gb	~25000
Plants					
A. thaliana	Flowering plant	Arabidopsis.org	Data version 31/01/03	125	25426
O. sativa indica	Rice	NCBI	2.1	466	46000-55000
O. sativa jap.	Rice	NCBI	N/A	420	32000-50000
Apicomplexa					
C. hominis	Apicomplexan protozoan pathogen	VCU	N/A	9.2	3994
P. yoelii	Rodent malaria parasite	TIGR	5	23.1	5878
Amoebozoa					
E. histolytica	Intestinal protozoan parasite	TIGR	N/A	18-20	9938
Red algae					
C. merolae	Unicellular red alga	Tokyo Univ.	N/A	16	5331
Diatoms					
T. pseudonana	Unicellular alga (diatom)	JGI	1.0	34.5	11242

Table XII.

Taxonomy and Species	*Gene No.*	*AGC*	*CAMK*	*CK1*	*CMGC*	*RGC*	*STE*	*TK*	*TKL*	*Total ePKs*	*PIKK*	*PDHK*	*RIO*	*Alpha*	*Total aPKs*
Fungi															
C. glabrata	5283	22	29	4	26	0	14	0	0	95	5	2	1	0	8
C. neoformans	6572	22	16	4	24	0	11	0	2	79	4	3	2	0	9
D. hansenii	6906	18	20	3	26	0	14	0	0	81	4	3	1	0	8
K. lactis	5329	20	25	3	25	0	12	0	0	85	4	3	1	0	8
N. crassa	10082	20	20	2	22	0	14	0	0	78	4	3	1	2	10
P. chrysosporium	11777	31	24	5	27	0	17	0	4	108	5	2	1	0	8
S. pombe	4940	20	29	5	26	0	17	0	0	97	6	1	1	0	8
Y. lipolytica	6703	21	19	2	22	0	13	0	0	77	4	3	1	0	8
S. cerevisiae	**5885**	**22**	**44**	**4**	**29**	**0**	**16**	**0**	**0**	**115**	**5**	**2**	**2**	**0**	**9**
Animals															
C. elegans	**~19000**	**35**	**65**	**92**	**56**	**28**	**36**	**94**	**20**	**426**	**5**	**1**	**3**	**1**	**10**
D. melanogaster	**13600**	**42**	**45**	**10**	**39**	**6**	**24**	**33**	**23**	**222**	**5**	**1**	**3**	**0**	**9**
A. gambiae	14000	42	41	9	33	6	24	32	18	205	6	1	3		
C. intestinalis	15852	47	58	6	36	5	27	49	29	257	6	1	3	2	12
T. nigroviridis	27918	105	109	14	90	14	62	120	56	570	5	5	3	6	19
T. rubripes	40000	125	106	17	110	14	76	135	50	633	3	8	3	6	20
R. norvegicus	~25000	96	135	26	84	6	73	106	52	578	5	6	4	6	21
M. musculus	**~30000**	**81**	**125**	**11**	**67**	**7**	**59**	**95**	**54**	**499**	**6**	**5**	**3**	**6**	**20**
H. sapiens	**~30000**	**84**	**98**	**12**	**70**	**5**	**61**	**93**	**55**	**478**	**6**	**5**	**3**	**6**	**20**
Plants															
A. thaliana	25426	67	158	28	149	0	104	2	776	1284	10	3	4	0	17
O. sativa indica	46000-55000	54	93	15	97	0	55	6	1010	1330	5	2	1	0	8
O. sativa jap.	32000-50000	52	88	15	91	0	54	7	969	1276	5	2	1	0	8
Apicomplexa															
C. hominis	3994	11	14	2	18	0	6	0	4	55	2	0	1	0	3
P. yoelii	5878	14	11	1	18	0	0	0	4	48	3	0	2	0	5
P. falciparum	**5268**	**14**	**18**	**1**	**23**	**0**	**3**	**0**	**6**	**65**	**3**	**0**	**2**	**0**	**5**
Amoebozoa															
E. histolytica	9938	38	49	9	48	0	35	7	109	295	6	0	3	4	13
Red algae															
C. merolae	5331	10	9	2	16	0	7	0	9	53	3	1	1	0	5
Diatoms															
T. pseudonana	11242	33	41	3	21	0	8	0	6	112	4	3	2	2	11

Table headings

Table I. The ePK complement of *H. sapiens, C. elegans, D. melanogaster, M. musculus, S. cerevisiae,* and *D. discoideum* as described in KinBase (http://www.kinase.com/kinbase/). Protein kinase families: AGC (including cyclic-nucleotide and calcium-phospholipid-dependent kinases, ribosomal S6-phosphorylating kinases, G protein-coupled kinases, and all close relatives of these groups); the CAMKs (calmodulin-regulated kinases); the CK1 family (casein kinase 1, and close relatives); the CMGC family (including cyclin-dependent kinases, mitogen-activated protein kinases, CDK-like kinases, and glycogen synthase kinase); the RGC family (receptor guanylate cyclase kinases, which are similar in domain sequence to tyrosine kinases); the STE family (including many kinases functioning in MAP kinase cascades); the TK family (tyrosine kinases); and the TKL family (tyrosine kinase-like kinases, a diverse group resembling TK but which are in fact serine-threonine kinases). This classification excludes the 'Other' group, which consists of kinases apparently not easy to classify into any of the groups below. *S. cerevisiae* lacks kinases from the groups: receptor guanylate cyclase kinase (RGC), tyrosine kinase (TK), and tyrosine kinase-like (TKL).

Table II. The revised classification of eukaryotic atypical protein kinases (aPKs). Four families -PIKK, RIO, PDHK, and Alpha- are present in *H. sapiens* and *C. elegans*, whereas *S. cerevisiae* and *D. melanogaster* only possess members of the RIO, PIKK, and PDHK families.

Table III. Classification performance of Library 1 from *H. sapiens*. The classification was regarded as correct whenever the model aligning with the best E-value to a given query sequence belonged to the same family as the sequence being classified. No.K: Number of protein kinases; No.CC: Number of Correctly Classified protein kinases; %CC: Percentage of Correctly Classified protein kinases.

Table IV. Make-up of HMM Library 2. Each family used in the construction of Library 1 was split into sub-

families as shown in this table. For example, the AGC family was represented by 5 sub-family HMMs rather than a single HMM for the entire family.

Table V. Classification performance of Library 2. The classification was regarded as correct whenever the model aligning with the best E-value to a given query sequence belonged to the same family as the sequence being classified. No.K: Number of protein kinases; No.CC: Number of CorrectlyClassified protein kinases; %CC: Percentage of Correctly Classified protein kinases.

Table VI. Performance of Library 2 in the classification of the kinomes of *M. musculus,* the malaria parasite *P. falciparum,* and *D. discoideum*. No.K: Number of protein kinases; No.CC: Number of Correctly Classified protein kinases; %CC: Percentage of Correctly Classified protein kinases.

Table VII. Result of using the Library 2 to search Uniref100 for annotated protein kinases. Library 2 retrieves all of the KinBase kinases in a database search. No.K: Number of KinBase sequences; TP indicates the number of KinBase sequences of that family retrieved by the library of models; E-value: Indicates the worst E-value obtained for a protein kinase member of the family.

Table VIII. Result of using BLAST with human kinase catalytic domains as query sequences to mine Uniref 100 KinBase sequences. BLAST was not found to be as good as Library 2, only returning a fraction of the KinBase sequences. No.K: Number of KinBase sequences; TP / FP: average number of True Positives (TP), and False Positives (FP). A FP is defined as a KinBase sequence that has been retrieved in the search but which belongs to a distinct ePK family as the query sequence; E-value: E-value of lowest-scoring match.

Table IX. Result of using the general HMM of the kinase catalytic domain PF00069 for searching Uniref100 KinBase sequences. The HMM failed to identify between 2.61-10.09% of the ePKs from characterised

kinomes, especially the more divergent ‘Other’ ePKs. No.K: Number of KinBase sequences.

Table X. Classification of the yeast protein kinases of the 'Other' group into the ePK families AGC, CAMK, CMGC, and STE.

Table XI Source and statistics on the 21 eukaryotic genomes whose protein kinase complement was determined by Library 2, and which represent a variety of taxonomic groups. (a) Taxonomy follows Baldauf *et al* (2000). (b) The URLs for the source databases are as follows: JGI, http://genome.jgi-psf.org/; Tokyo University, http://merolae.biol.s.u-tokyo.ac.jp/; NCBI, ftp://ftp.ncbi.nih.gov/genbank/genomes/; BROAD-MIT, ftp://ftp.broad.mit.edu/pub/annotation/; Genolevures, http://cbi.labri.fr/Genolevures/download.php#codingseq ; TIGR, http://www.tigr.org/; Sanger, http://www.sanger.ac.uk/Projects/S_pombe/; VCU, www.parvum.mic.vcu.edu.; http://www.arabidopsis.org. For full references to the genomes, see the text. (c) N/A: version number not available

Table XII. The protein kinase complement of the eukaryotic organisms with completed genomes divided into 8 ePK and 4 aPK families. The peptide complement of each genome was scanned through Library 2. The assignment of kinase family was performed by retrieving the HMM that classifyied a peptide with the best E-value.

Figure 1. Schematic representation of the process for arriving at a library of human-derived HMMs that would optimise the family-specific classification of protein kinases across a large evolutionary distance. In step 1, all human kinase catalytic domain sequences are compared pairwise, and the chosen Z-score cutoff splits the family into a number of sub-families. The sequences in each sub-family are compared pairwise, followed by their multiple alignment and their translation into an HMM. At this stage a new HMM library has

been created where each human kinase family is represented by a number of HMMs. The classification performance of the HMM library is tested on the four characterised kinomes of *H. sapiens* (self-classification), *D. melanogaster, C. elegans*, and *S. cerevisiae*. The Z-score cutoff is adjusted empirically in this iterative process until the best classification rate has been achieved on the four kinomes.

Figure 2. The 12 human kinase catalytic domains of the CK1 family were compared pairwise, and a Z-score cutoff of 20.0 led to their division into 2 subfamilies. The sequences in each subfamily were then compared pairwise, followed by their multiple alignment and the generation of an HMM from each alignment.

Figure 3. Graphic summary of the process for arriving at Library 2 from Library 1.

Figure 4. Multiple alignment of representative dual-specificity phosphatase catalytic domains with those identified for *E. histolytica* (15 members, sequence ids in green), *A. thaliana* (6 members, sequence ids in red), *O. sativa ssp. Indica* (6 members, sequence ids in light blue) and *O. sativa ssp. Japonica* (6 members, sequence ids in dark blue). The sequence ids in black are representative examples of the Pfam signature PF00782. Dual-specificity phosphatases carry out the dephosphorylation of phosphoserine, phosphothreonine, and phosphotyrosine residues. The alignment was generated and displayed with Jalview (70).

Figure 5. Alignment of representative tyrosine phosphatase sequences with those identified in *Entamoeba* (2 members, sequence ids in green), *A. thaliana* (1 member, sequence ids in red), and the two rice species (2 members each, sequence ids in light blue for *O. sativa ssp. Indica,* and dark blue for *O. sativa ssp. japonica*). The sequence ids in black are representative examples of the Pfam signature PF00102. The presence of tyrosine phosphatases in the amoeba and plant species suggests that tyrosine phosphorylation signalling systems were present before the lineages leading to plants, animals, and amoebozoa split. The alignment was generated and displayed with Jalview (70).

S1. Three representative examples of ePKs of *S. cerevisiae* that belong to the 'Other' group but which have been re-classified by Library 2 into one of the main ePK families. The examples include IPL1 (Aurora kinase), a putative member of the AGC family, and PAK1 and CDC7, putative members of the CAMK and CMGC families, respectively.

References

1. Cohen P. The regulation of protein function by multisite phosphorylation--a 25 year update. Trends Biochem Sci 2000;25(12):596-601.
2. Cohen P. The role of protein phosphorylation in human health and disease. The Sir Hans Krebs Medal Lecture. Eur J Biochem 2001;268(19):5001-10.
3. Cohen P. Protein kinases--the major drug targets of the twenty-first century? Nat Rev Drug Discov 2002;1(4):309-15.
4. Manning G, Plowman GD, Hunter T, Sudarsanam S. Evolution of protein kinase signaling from yeast to man. Trends Biochem Sci 2002;27(10):514-20.
5. Hanks SK, Hunter T. Protein kinases 6. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification. Faseb J 1995;9(8):576-96.
6. Manning G, Whyte DB, Martinez R, Hunter T, Sudarsanam S. The protein kinase complement of the human genome. Science 2002;298(5600):1912-34.
7. Angermayr M, Bandlow W. The general regulatory factor Reb1p controls basal, but not Gal4p-mediated, transcription of the GCY1 gene in yeast. Mol Gen Genet 1997;256(6):682-9.
8. LaRonde-LeBlanc N, Wlodawer A. Crystal structure of A. fulgidus Rio2 defines a new family of serine protein kinases. Structure 2004;12(9):1585-94.
9. 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(5):1047-57.
10. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, et al. The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res 2006;34(Database issue):D187-91.
11. 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 U S A 1999;96(24):13603-10.
12. Hunter T, Plowman GD. The protein kinases of budding yeast: six score and more. Trends Biochem Sci 1997;22(1):18-22.
13. Morrison DK, Murakami MS, Cleghon V. Protein kinases and phosphatases in the Drosophila genome. J Cell Biol 2000;150(2):F57-62.
14. Caenepeel S, Charydczak G, Sudarsanam S, Hunter T, Manning G. The mouse kinome: discovery and comparative genomics of all mouse protein kinases. Proc Natl Acad Sci U S A 2004;101(32):11707-12.
15. Ward P, Equinet L, Packer J, Doerig C. Protein kinases of the human malaria parasite Plasmodium falciparum: the kinome of a divergent eukaryote. BMC Genomics 2004;5(1):79.
16. Goldberg JM, Manning G, Liu A, Fey P, Pilcher KE, Xu Y, et al. The dictyostelium kinome--analysis of the protein kinases from a simple model organism. PLoS Genet 2006;2(3):e38.
17. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25(17):3389-402.
18. Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, et al. The Pfam protein families database. Nucleic Acids Res 2004;32(Database issue):D138-41.
19. Hanks SK. Genomic analysis of the eukaryotic protein kinase superfamily: a perspective. Genome Biol 2003;4(5):111.
20. Krogh A, Brown M, Mian IS, Sjolander K, Haussler D. Hidden Markov models in computational biology. Applications to protein modeling. J Mol Biol 1994;235(5):1501-31.
21. Park J, Karplus K, Barrett C, Hughey R, Haussler D, Hubbard T, et al. Sequence comparisons using multiple sequences detect three times as many remote homologues as pairwise methods. J Mol Biol

1998;284(4):1201-10.
22. Madera M, Vogel C, Kummerfeld SK, Chothia C, Gough J. The SUPERFAMILY database in 2004: additions and improvements. Nucleic Acids Res 2004;32(Database issue):D235-9.
23. 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(4):903-19.
24. Brown D, Krishnamurthy N, Dale JM, Christopher W, Sjolander K. Subfamily hmms in functional genomics. Pac Symp Biocomput 2005:322-33.
25. Gudi R, Bowker-Kinley MM, Kedishvili NY, Zhao Y, Popov KM. Diversity of the pyruvate dehydrogenase kinase gene family in humans. J Biol Chem 1995;270(48):28989-94.
26. Barton GJ, Sternberg MJ. Evaluation and improvements in the automatic alignment of protein sequences. Protein Eng 1987;1(2):89-94.
27. Eddy SR. Profile hidden Markov models. Bioinformatics 1998;14(9):755-63.
28. Baldauf SL. The deep roots of eukaryotes. Science 2003;300(5626):1703-6.
29. Baldauf SL, Doolittle WF. Origin and evolution of the slime molds (Mycetozoa). Proc Natl Acad Sci U S A 1997;94(22):12007-12.
30. Hedges SB, Blair JE, Venturi ML, Shoe JL. A molecular timescale of eukaryote evolution and the rise of complex multicellular life. BMC Evol Biol 2004;4:2.
31. Dietrich FS, Voegeli S, Brachat S, Lerch A, Gates K, Steiner S, et al. The Ashbya gossypii genome as a tool for mapping the ancient Saccharomyces cerevisiae genome. Science 2004;304(5668):304-7.
32. Kellis M, Birren BW, Lander ES. Proof and evolutionary analysis of ancient genome duplication in the yeast Saccharomyces cerevisiae. Nature 2004;428(6983):617-24.
33. Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, et al. The genome of the diatom Thalassiosira pseudonana: ecology, evolution, and metabolism. Science 2004;306(5693):79-86.
34. Matsuzaki M, Misumi O, Shin IT, Maruyama S, Takahara M, Miyagishima SY, et al. Genome sequence of the ultrasmall unicellular red alga Cyanidioschyzon merolae 10D. Nature 2004;428(6983):653-7.
35. Holt RA, Subramanian GM, Halpern A, Sutton GG, Charlab R, Nusskern DR, et al. The genome sequence of the malaria mosquito Anopheles gambiae. Science 2002;298(5591):129-49.
36. Dehal P, Satou Y, Campbell RK, Chapman J, Degnan B, De Tomaso A, et al. The draft genome of Ciona intestinalis: insights into chordate and vertebrate origins. Science 2002;298(5601):2157-67.
37. Jaillon O, Aury JM, Brunet F, Petit JL, Stange-Thomann N, Mauceli E, et al. Genome duplication in the teleost fish Tetraodon nigroviridis reveals the early vertebrate proto-karyotype. Nature 2004;431(7011):946-57.
38. Aparicio S, Chapman J, Stupka E, Putnam N, Chia JM, Dehal P, et al. Whole-genome shotgun assembly and analysis of the genome of Fugu rubripes. Science 2002;297(5585):1301-10.
39. Dujon B, Sherman D, Fischer G, Durrens P, Casaregola S, Lafontaine I, et al. Genome evolution in yeasts. Nature 2004;430(6995):35-44.
40. Loftus BJ, Fung E, Roncaglia P, Rowley D, Amedeo P, Bruno D, et al. The genome of the basidiomycetous yeast and human pathogen Cryptococcus neoformans. Science 2005;307(5713):1321-4.
41. Galagan JE, Calvo SE, Borkovich KA, Selker EU, Read ND, Jaffe D, et al. The genome sequence of the filamentous fungus Neurospora crassa. Nature 2003;422(6934):859-68.
42. Martinez D, Larrondo LF, Putnam N, Gelpke MD, Huang K, Chapman J, et al. Genome sequence of the lignocellulose degrading fungus Phanerochaete chrysosporium strain RP78. Nat Biotechnol 2004;22(6):695-700.
43. Wood V, Gwilliam R, Rajandream MA, Lyne M, Lyne R, Stewart A, et al. The genome sequence of Schizosaccharomyces pombe. Nature 2002;415(6874):871-80.
44. Gibbs RA, Weinstock GM, Metzker ML, Muzny DM, Sodergren EJ, Scherer S, et al. Genome sequence of the Brown Norway rat yields insights into mammalian evolution. Nature

2004;428(6982):493-521.
45. Initiative AG. Analysis of the genome sequence of the flowering plant Arabidopsis thaliana. Nature 2000;408(6814):796-815.
46. Yu J, Hu S, Wang J, Wong GK, Li S, Liu B, et al. A draft sequence of the rice genome (Oryza sativa L. ssp. indica). Science 2002;296(5565):79-92.
47. Goff SA, Ricke D, Lan TH, Presting G, Wang R, Dunn M, et al. A draft sequence of the rice genome (Oryza sativa L. ssp. japonica). Science 2002;296(5565):92-100.
48. Xu P, Widmer G, Wang Y, Ozaki LS, Alves JM, Serrano MG, et al. The genome of Cryptosporidium hominis. Nature 2004;431(7012):1107-12.
49. Loftus B, Anderson I, Davies R, Alsmark UC, Samuelson J, Amedeo P, et al. The genome of the protist parasite Entamoeba histolytica. Nature 2005;433(7028):865-8.
50. Carlton JM, Angiuoli SV, Suh BB, Kooij TW, Pertea M, Silva JC, et al. Genome sequence and comparative analysis of the model rodent malaria parasite Plasmodium yoelii yoelii. Nature 2002;419(6906):512-9.
51. Alonso A, Sasin J, Bottini N, Friedberg I, Friedberg I, Osterman A, et al. Protein tyrosine phosphatases in the human genome. Cell 2004;117(6):699-711.
52. Hunter T. A thousand and one protein kinases. Cell 1987;50(6):823-9.
53. Mustelin T, Feng GS, Bottini N, Alonso A, Kholod N, Birle D, et al. Protein tyrosine phosphatases. Front Biosci 2002;7:d85-142.
54. Mustelin T, Abraham RT, Rudd CE, Alonso A, Merlo JJ. Protein tyrosine phosphorylation in T cell signaling. Front Biosci 2002;7:d918-69.
55. Rudrabhatla P, Reddy MM, Rajasekharan R. Genome-wide analysis and experimentation of plant serine/ threonine/tyrosine-specific protein kinases. Plant Mol Biol 2006;60(2):293-319.
56. Barizza E, Lo Schiavo F, Terzi M, Filippini F. Evidence suggesting protein tyrosine phosphorylation in plants depends on the developmental conditions. FEBS Lett 1999;447(2-3):191-4.
57. Kameyama K, Kishi Y, Yoshimura M, Kanzawa N, Sameshima M, Tsuchiya T. Tyrosine phosphorylation in plant bending. Nature 2000;407(6800):37.
58. Shiu SH, Li WH. Origins, lineage-specific expansions, and multiple losses of tyrosine kinases in eukaryotes. Mol Biol Evol 2004;21(5):828-40.
59. King N, Carroll SB. A receptor tyrosine kinase from choanoflagellates: molecular insights into early animal evolution. Proc Natl Acad Sci U S A 2001;98(26):15032-7.
60. King N, Hittinger CT, Carroll SB. Evolution of key cell signaling and adhesion protein families predates animal origins. Science 2003;301(5631):361-3.
61. Williams JG, Zvelebil M. SH2 domains in plants imply new signalling scenarios. Trends Plant Sci 2004;9(4):161-3.
62. Schlessinger J, Lemmon MA. SH2 and PTB domains in tyrosine kinase signaling. Sci STKE 2003;2003(191):RE12.
63. Benes CH, Wu N, Elia AE, Dharia T, Cantley LC, Soltoff SP. The C2 domain of PKCdelta is a phosphotyrosine binding domain. Cell 2005;121(2):271-80.
64. Abraham RT. The ATM-related kinase, hSMG-1, bridges genome and RNA surveillance pathways. DNA Repair (Amst) 2004;3(8-9):919-25.
65. Abraham RT, Tibbetts RS. Cell biology. Guiding ATM to broken DNA. Science 2005;308(5721):510-1.
66. Krungkrai J. The multiple roles of the mitochondrion of the malarial parasite. Parasitology 2004;129(Pt 5):511-24.
67. Holness MJ, Sugden MC. Regulation of pyruvate dehydrogenase complex activity by reversible phosphorylation. Biochem Soc Trans 2003;31(Pt 6):1143-51.
68. Drennan D, Ryazanov AG. Alpha-kinases: analysis of the family and comparison with conventional protein kinases. Prog Biophys Mol Biol 2004;85(1):1-32.
69. Project IRGS. The map-based sequence of the rice genome. Nature 2005;436(7052):793-800.

70. Clamp M, Cuff J, Searle SM, Barton GJ. The Jalview Java alignment editor. Bioinformatics 2004;20(3):426-7.

Supplementary Material and Figures

S1.

Example 1: Kinase IPL1_Sc (Aurora kinase)

Scores against the sub-family HMMs of Library 2:

```
Scores for sequence family classification (score includes all domains):
Model                Description                          Score    E-value  N
--------             -----------                          -----    ------- ---
AGC_group4_multiple                                       249.3    3.1e-74   1
AGC_group1_multiple                                       223.7    1.6e-66   1
AGC_group3_multiple                                       223.1    2.4e-66   1
AGC_group2_multiple                                       218.7      5e-65   1
AGC_group5_multiple                                       218.3    6.6e-65   1
CAMK_group5_multiple                                      205.8      4e-61   1
CAMK_group6_multiple                                      185.7    4.5e-55   1
CAMK_group4_multiple                                      181.1    1.1e-53   1
CAMK_group3_multiple                                      169.5    3.3e-50   1
CAMK_group7_multiple                                      167.6    1.2e-49   1
CAMK_group2_multiple                                      166.4    2.8e-49   1
CAMK_group1_multiple                                      152.1    5.8e-45   1
TKL_group4_multiple                                        85.7    5.5e-25   1
CMGC_group2_multiple                                       66.8    5.6e-22   1
STE_group1_multiple                                        64.9    9.9e-21   1
STE_group3_multiple                                        69.7    3.6e-20   1
CMGC_group3_multiple                                       55.0      1e-19   1
TKL_group3_multiple                                        62.0    7.7e-18   1
CMGC_group1_multiple                                       37.1    3.8e-17   1
STE_group4_multiple                                        56.7    9.6e-17   1
STE_group2_multiple                                        56.6    1.9e-16   1
TKL_group2_multiple                                        43.5    4.1e-15   1
TK_group5_multiple                                        -29.4    2.7e-14   1
TK_group2_multiple                                        -17.7    2.5e-13   1
TKL_group5_multiple                                        28.7    3.3e-13   1
TK_group1_multiple                                        -45.1    9.2e-13   1
TK_group4_multiple                                        -15.1    1.8e-12   1
TKL_group1_multiple                                        29.9    2.7e-12   1
TK_group3_multiple                                        -42.0      4e-11   1
CK1_group1_multiple                                       -82.7    1.9e-08   1
CK1_group2_multiple                                       -85.8    5.8e-08   1
TK_group6_multiple                                       -178.0    8.9e-07   1
RGC_group1_multiple                                      -142.4     0.0003   1
```

Alignment to the top-scoring sub-family HMM of Library 2:

```
AGC_group4_multiple: domain 1 of 1, from 108 to 355: score 249.3, E = 3.1e-74
                   *->lkLGkGsFGkVflcrkcvhkaTgklYAvKvLkKkrivKkev..qtmt
                      +kLGkG FGkV+ +r+    +Tg   A Kv +K+ i+K + ++q ++
     IPL1_Sc   108    KKLGKGKFGKVYCVRH---RSTGYICALKVMEKEEIIKYNLqkQFRR 151

                   ErdiLsrvrtgHPFiVkLhyaFqTkekLylVmeymnGGeLfthiYvskev
                   E +i + ++  HP+++k +  F++++++yl mey+ +Ge  + +  + +
     IPL1_Sc   152 EVEIQTSLN--HPNLTKSYGYFHDEKRVYLLMEYLVNGEMYKLL--RLHG 197

                   .FsEpearFYaaeIvlALeyLHslgIVYRDLKPENILYLDedGHlEsikI
```

```
                +F    a++Y ++I+ AL+y+H+++I++RD+KPENIL     G+   ik+
IPL1_Sc   198 pFNDILASDYIYQIANALDYMHKKNIIHRDIKPENIL----IGFNNVIKL 243

                tDFGLaKvEdknartFLDktfCGTpeYmAPEVLqrqgYdksvDWWsLGVl
                tDFG + + + + r    kt+CGT++Y++PE ++ ++Yd++ D W+LGVl
IPL1_Sc   244 TDFGWSIINPPENRR---KTVCGTIDYLSPEMVESREYDHTIDAWALGVL 290

                lYEMLaGqsPFraddekLCEiferIlkdevsyPWdtlSpeAkdLlekLLh
                 +E+L+G +PF+    +   +++rI   ++++P  ++S++A+dL+ kLL+
IPL1_Sc   291 AFELLTGAPPFEEEMKD--TTYKRIAALDIKMP-SNISQDAQDLILKLLK 337

                kDPhkRLGkegadeiKqHpFF<-*
                 DP+ R   +   ++K Hp++
IPL1_Sc   338 YDPKDRMRLG---DVKMHPWI    355
```

Example 2: Kinase PAK1_Sc

Scores against the sub-family HMMs of Library 2:

Scores for sequence family classification (score includes all domains):

Model	Description	Score	E-value	N
CAMK_group3_multiple		121.7	8e-36	1
CAMK_group4_multiple		120.6	1.7e-35	1
CAMK_group7_multiple		118.8	6.2e-35	1
CAMK_group5_multiple		111.3	1.1e-32	1
AGC_group4_multiple		107.4	1.6e-31	1
CAMK_group6_multiple		106.1	4e-31	1
CAMK_group2_multiple		87.4	1.7e-25	1
CAMK_group1_multiple		83.9	1.9e-24	1
STE_group3_multiple		78.7	7.1e-23	1
AGC_group2_multiple		49.8	5.7e-20	1
AGC_group1_multiple		69.1	5.7e-20	1
AGC_group3_multiple		68.8	6.7e-20	1
CMGC_group2_multiple		15.0	1.9e-18	1
STE_group1_multiple		32.4	2.2e-18	1
AGC_group5_multiple		59.9	1.2e-17	1
CMGC_group3_multiple		20.4	2.3e-17	1
STE_group4_multiple		28.9	3.5e-15	1
STE_group2_multiple		30.2	6.6e-15	1
CMGC_group1_multiple		-30.1	5.9e-13	1
TK_group2_multiple		-31.9	1.8e-12	1
TKL_group5_multiple		-2.5	3e-11	1
TK_group5_multiple		-84.5	2.2e-10	1
TKL_group4_multiple		-34.2	3.5e-10	1
TKL_group3_multiple		-12.0	7.8e-10	1
TK_group4_multiple		-63.8	1.3e-09	1
TK_group1_multiple		-89.4	1.6e-09	1
TKL_group2_multiple		-38.9	3.5e-09	1
TK_group3_multiple		-83.5	1.7e-08	1
TKL_group1_multiple		-49.4	3.7e-07	1
RGC_group1_multiple		-127.7	3.6e-05	1
CK1_group1_multiple		-135.7	6.3e-05	1
TK_group6_multiple		-235.7	0.0017	1
CK1_group2_multiple		-150.8	0.0018	1
CMGC_group4_multiple		-179.4	0.0019	1

```
Alignment to the top-scoring sub-family HMM of Library 2:

CAMK_group3_multiple: domain 1 of 1, from 133 to 448: score 121.7, E = 8e-36
                    *->YelketLGkGqfsvVkkCh.kaTGrevAiKvmE.............ik
                       Ye  ++LG Gq+++Vk++++  +++ vAiK++++++++++  + ik
     PAK1_Sc   133    YEIIKELGHGQHGKVKLARdILSKQLVAIKIVDrhekkqrkfftfIK 179

                    kkklslevdrekirREieIlrqlrHPnIitLhdvi..EnCesklyLVfEl
                      k+s    ++ki REi+I++   H ++++L++v +++  + k+yLV+E+
     PAK1_Sc   180 SSKIS---ENDKIKREIAIMKKCHHKHVVQLIEVLddLK-SRKIYLVLEY 225

                    vsgGE........LFdyivererHLsEkEArhffrQIlsAVhylHqnrIV
                     s+GE +  +++ + ++       Ls  E+r ++r  + ++ ylH   I+
     PAK1_Sc   226 CSRGEvkwcppdcMESDAKGPSL-LSFQETREILRGVVLGLEYLHYQGII 274

                    HrDLKPENiLLddvkniKLaDFGla.............eleggdkletfG
                    HrD KP N+L +    +K++DFG +   ++++++++++l   +  +t G
     PAK1_Sc   275 HRDIKPANLLISGDGTVKISDFGVSlaasstnssdsseSLDELELAKTVG 324

                    tPaYlSPEvlngdEtH..............PyGkkvDvWScGVILYiLLv
                    tPa+ +PE  +g++  ++ + ++++  +++ ++   D+W+ GV LY LL
     PAK1_Sc   325 TPAFFAPEMCLGEDAFtrynltkenlfrgsCISFMIDIWAVGVTLYCLLF 374

                    GamPFddeDqkeLlrkI...................krGaFPeSpeAKdL
                    G +PF  +   +L++kI +++ + ++ ++ ++++ +k   ++e + AKdL
     PAK1_Sc   375 GMLPFFSDFELKLFEKIvndplkfptfkeiqsnkvsKVSCEEEYEMAKDL 424

                    IrrLLqvdPsrRlTaeqvLkHpWl<-*
                    + +LL+++P +R+T++ + kHp++
     PAK1_Sc   425 LLKLLEKNPQKRMTIPAIKKHPFV    448
```

Example 3: Kinase CDC7_Sc

Scores against the sub-family HMMs of Library 2:

Scores for sequence family classification (score includes all domains):

Model	Description	Score	E-value	N
CMGC_group3_multiple		-5.1	1.3e-15	1
STE_group3_multiple		-44.0	8.9e-11	1
CMGC_group1_multiple		-65.9	1e-10	1
CMGC_group2_multiple		-111.5	8.5e-10	1
AGC_group4_multiple		-109.8	2.2e-09	1
CAMK_group1_multiple		-105.0	2.5e-08	1
STE_group1_multiple		-113.5	7.2e-08	1
AGC_group2_multiple		-135.9	7.5e-08	1
CAMK_group5_multiple		-97.5	1.2e-07	1
AGC_group1_multiple		-115.3	1.3e-07	1
CAMK_group2_multiple		-120.9	7e-07	1
AGC_group3_multiple		-111.3	7.9e-07	1
CAMK_group4_multiple		-103.5	1.7e-06	1
CK1_group2_multiple		-111.9	3.7e-06	1
STE_group2_multiple		-120.4	3.9e-06	1
STE_group4_multiple		-135.1	5.8e-06	1
CAMK_group6_multiple		-113.4	8.4e-06	1

```
TKL_group2_multiple                                    -86.9    9.9e-06   1
CAMK_group7_multiple                                  -121.3    1.4e-05   1
AGC_group5_multiple                                   -135.4    2.1e-05   1
CAMK_group3_multiple                                  -129.5      4e-05   1
CK1_group1_multiple                                   -133.7    4.7e-05   1
TKL_group5_multiple                                   -106.0    9.3e-05   1
TK_group2_multiple                                    -163.0    0.00016   1
TKL_group4_multiple                                   -113.2    0.00017   1
TK_group1_multiple                                    -165.5    0.00058   1
TK_group5_multiple                                    -182.9     0.0022   1
TKL_group1_multiple                                   -111.7     0.0039   1
TK_group4_multiple                                    -176.2      0.005   1
TKL_group3_multiple                                   -137.2      0.023   1
TK_group3_multiple                                    -182.5      0.028   1
RGC_group1_multiple                                   -191.5       0.34   1
TK_group6_multiple                                    -282.9        0.8   1
CMGC_group4_multiple                                  -254.2        1.9   1


Alignment to the top-scoring sub-family HMM of Library 2:

CMGC_group3_multiple: domain 1 of 1, from 33 to 469: score -5.1, E = 1.3e-15
                    *->YeklgCklGeGTYGvVykardkktg..............elVAlKki
                       Y  ++ k+GeGT+++Vyka+d  ++ +++  ++  + ++++VAlKki
     CDC7_Sc    33    YKLID-KIGEGTFSSVYKAKDITGKitkkfashfwnygsNYVALKKI 78

                    pkta.....lRElklLKkLkHaNkddvVnLldvfhtkrkCGLyLVFEyle
                      t ++++ ++El lL  +  ++   v  L d  +  ++  +  V++y +
     CDC7_Sc    79 YVTSspqriYNELNLLYIM--TGSSRVAPLCDAKRVRDQ--VIAVLPYYP 124

                    hdLyellgkllplhlVksfmyQlLqgLaylHknkvlHRDLKPeNiLvnLY
                    h  ++   + lp++ +k++ + lL++L++ H+ +++HRD+KP N L n
     CDC7_Sc   125 HEEFRTFYRDLPIKGIKKYIWELLRALKFVHSKGIIHRDIKPTNFLFN-- 172

                    NHKCergelKlaDFGlA..seppY.........................
                        e g+  l DFGlA+   ++++  +++++ ++  +++++++ + +++
     CDC7_Sc   173 ---LELGRGVLVDFGLAeaQMDYKsmissqndydnyantnhdggysmrnh 219

                    ..................................................
                    ++  +   +++ +++++++++  + ++++  + ++ ++ + +++ ++++
     CDC7_Sc   220 eqfcpcimrnqyspnshnqtppmvtiqngkvvhlnnvngvdltkgypkne 269

                    .........VvTRWYRAPEiLLs..tYstpIDvWsvGCIlAEll.tGrPL
                    +++ ++ ++  TR  RAPE+L + +  st+ID+WsvG Il  ll++++P
     CDC7_Sc   270 trrikranrAGTRGFRAPEVLMKcgAQSTKIDIWSVGVILLSLLgRRFPM 319

                    FPGkse.......................................
                    F+  ++ ++  +  +  + ++ ++    ++ + + ++   +++++ +++
     CDC7_Sc   320 FQSLDDadsllelctifgwkelrkcaalhglgfeasgliwdkpngysngl 369

                    .................DQLalIfkvlGktPtpelwpgvkpefknnLvsf
                    ++   +  +++ + ++  +   +f++ G ++ +el+  ++ e       +
     CDC7_Sc   370 kefvydllnkectigtfPEYSVAFETFG-FLQQELHDRMSIEP-----QL 413

                    Pqvtvpnd.L............kklfrldsdaidLlkkmLelDPakRita
                    P+   p+++ +  +  + ++ ++  ++  +  ++ l+++ e+DP kR +a
     CDC7_Sc   414 PD---PKTnMdavdayelkkyqEEIWSDHYWCFQVLEQCFEMDPQKRSSA 460

                    eqALrHPFF<-*
                    e+ L+ PFF
     CDC7_Sc   461 EDLLKTPFF    469
```

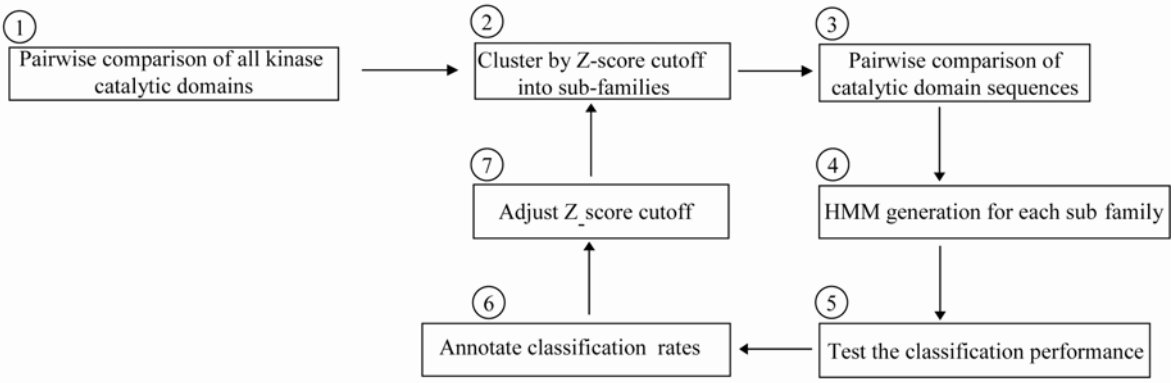
1
Pairwise comparison of all kinase catalytic domains
2
Cluster by Z-score cutoff into sub-families
3
Pairwise comparison of catalytic domain sequences
4
HMM generation for each sub family
5
Test the classification performance
6
Annotate classification rates
7
Adjust Z_score cutoff

Fig. 1

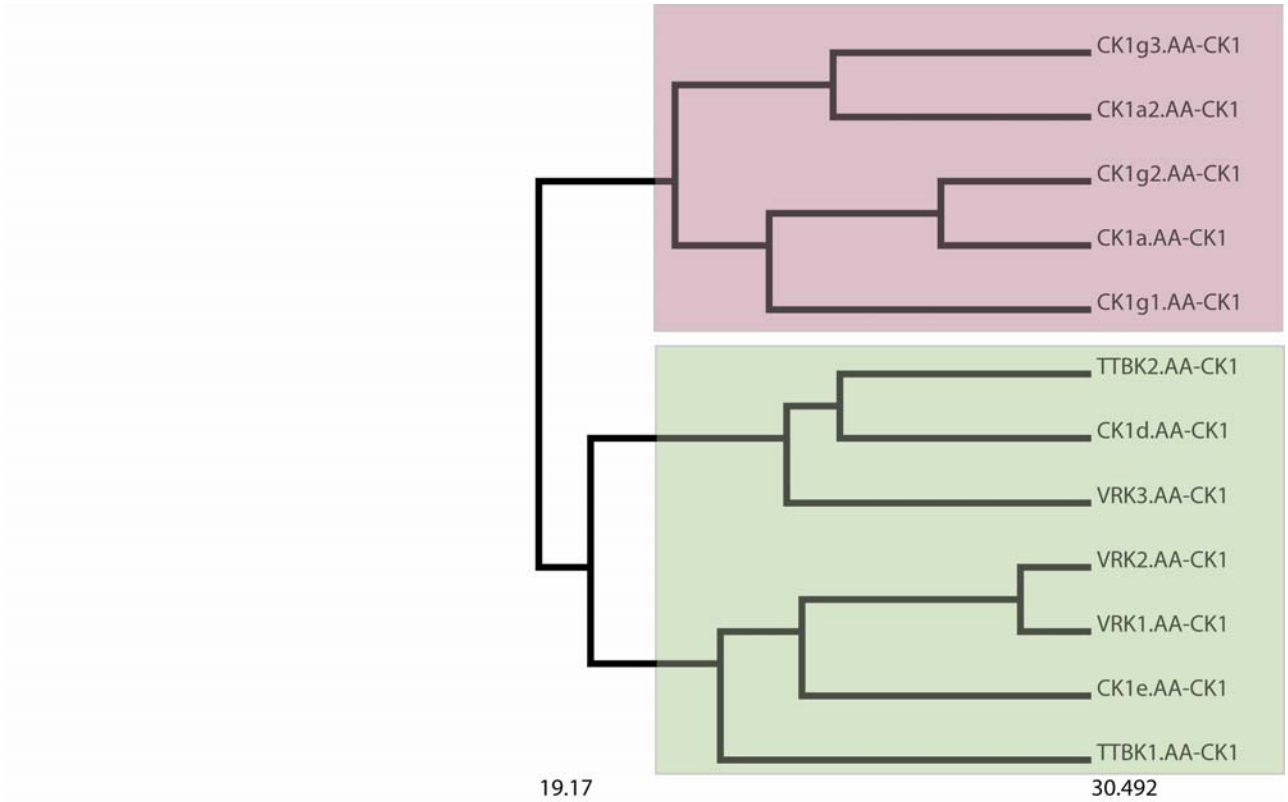
CK1g3.AA-CK1
CK1a2.AA-CK1
CK1g2.AA-CK1
CK1a.AA-CK1
CK1g1.AA-CK1
TTBK2.AA-CK1
CK1d.AA-CK1
VRK3.AA-CK1
VRK2.AA-CK1
VRK1.AA-CK1
CK1e.AA-CK1
TTBK1.AA-CK1
19.17
30.492

Fig. 2.

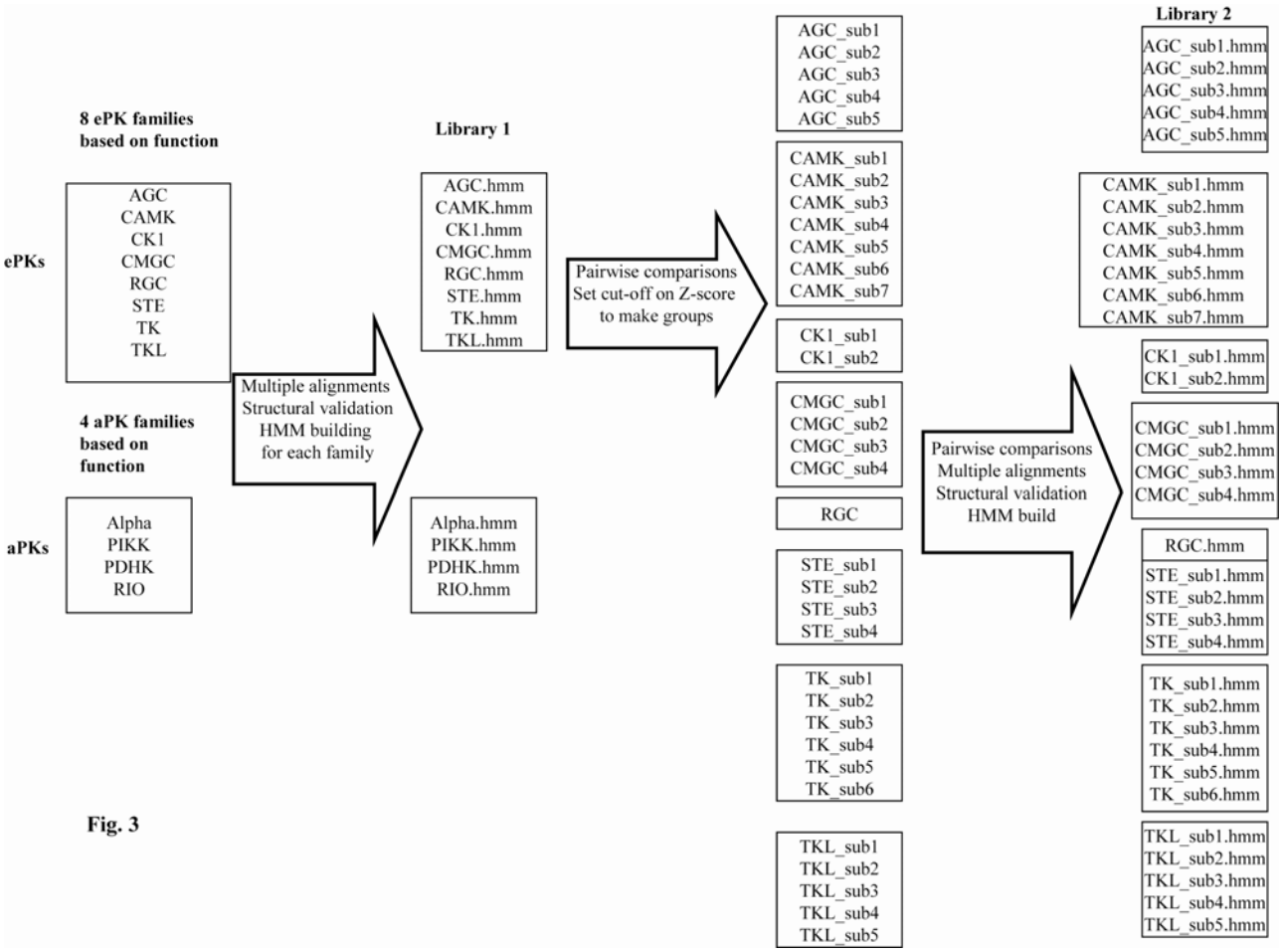
8 ePK families based on function
ePKs
AGC
CAMK
CK1
CMGC
RGC
STE
TK
TKL
4 aPK families based on function
aPKs
Alpha
PIKK
PDHK
RIO
Multiple alignments
Structural validation
HMM building for each family
Library 1
AGC.hmm
CAMK.hmm
CK1.hmm
CMGC.hmm
RGC.hmm
STE.hmm
TK.hmm
TKL.hmm
Alpha.hmm
PIKK.hmm
PDHK.hmm
RIO.hmm
Pairwise comparisons
Set cut-off on Z-score to make groups
AGC_sub1
AGC_sub2
AGC_sub3
AGC_sub4
AGC_sub5
CAMK_sub1
CAMK_sub2
CAMK_sub3
CAMK_sub4
CAMK_sub5
CAMK_sub6
CAMK_sub7
CK1_sub1
CK1_sub2
CMGC_sub1
CMGC_sub2
CMGC_sub3
CMGC_sub4
RGC
STE_sub1
STE_sub2
STE_sub3
STE_sub4
TK_sub1
TK_sub2
TK_sub3
TK_sub4
TK_sub5
TK_sub6
TKL_sub1
TKL_sub2
TKL_sub3
TKL_sub4
TKL_sub5
Pairwise comparisons
Multiple alignments
Structural validation
HMM build
Library 2
AGC_sub1.hmm
AGC_sub2.hmm
AGC_sub3.hmm
AGC_sub4.hmm
AGC_sub5.hmm
CAMK_sub1.hmm
CAMK_sub2.hmm
CAMK_sub3.hmm
CAMK_sub4.hmm
CAMK_sub5.hmm
CAMK_sub6.hmm
CAMK_sub7.hmm
CK1_sub1.hmm
CK1_sub2.hmm
CMGC_sub1.hmm
CMGC_sub2.hmm
CMGC_sub3.hmm
CMGC_sub4.hmm
RGC.hmm
STE_sub1.hmm
STE_sub2.hmm
STE_sub3.hmm
STE_sub4.hmm
TK_sub1.hmm
TK_sub2.hmm
TK_sub3.hmm
TK_sub4.hmm
TK_sub5.hmm
TK_sub6.hmm
TKL_sub1.hmm
TKL_sub2.hmm
TKL_sub3.hmm
TKL_sub4.hmm
TKL_sub5.hmm

Fig. 3

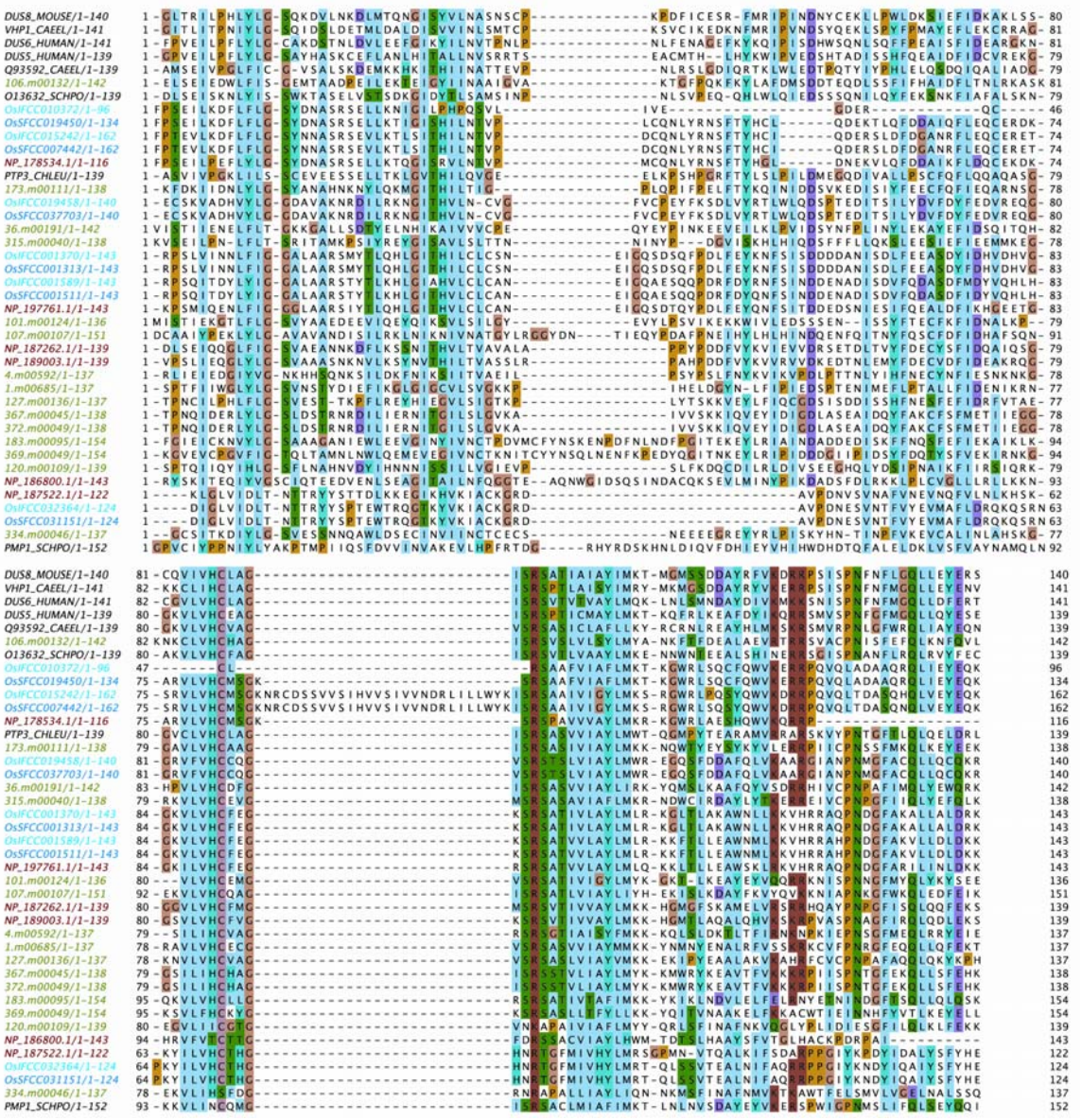

Fig. 4

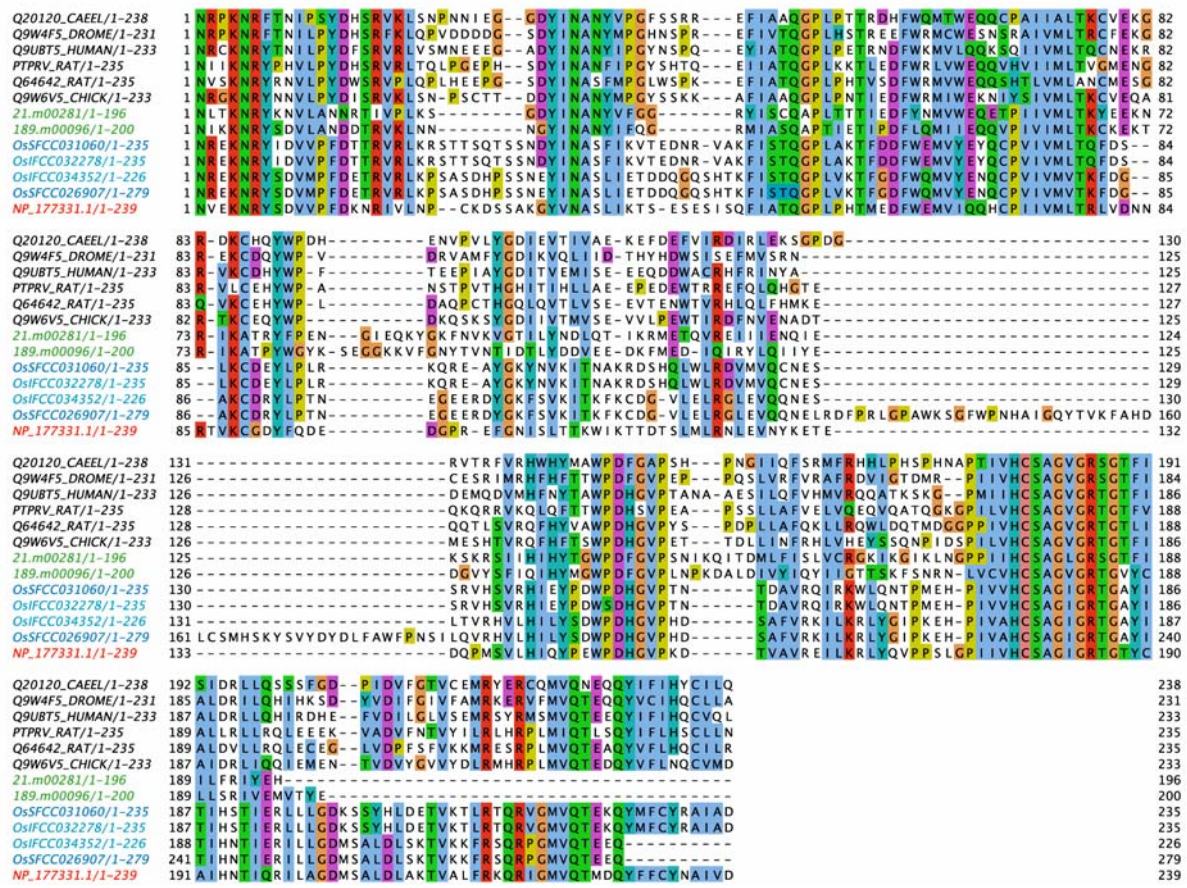
Q20120_CAEEL/1-238
Q9W4F5_DROME/1-231
Q9UBT5_HUMAN/1-233
PTPRV_RAT/1-235
Q64642_RAT/1-235
Q9W6V5_CHICK/1-233
21.m00281/1-196
189.m00096/1-200
OsSFCC031060/1-235
OsIFCC032278/1-235
OsIFCC034352/1-226
OsSFCC026907/1-279
NP_177331.1/1-239

Fig. 5