KiNext: a portable and scalable workflow for the identification and classification of protein kinases

Elisabeth Hellec, Flavia Nunes, Charlotte Corporeau, Alexandre Cormier

BMC Bioinformatics

SOFTWARE

Open Access

KiNext: a portable and scalable workflow for the identification and classification of protein kinases

Elisabeth Hellec[1,2,3], Flavia Nunes[2], Charlotte Corporeau[3] and Alexandre Cormier[1*]

*Correspondence:
alexandre.cormier@ifremer.fr

[1] Ifremer, IRSI-SeBiMER, Plouzané, France
[2] Ifremer, DYNECO-LEBCO, Plouzané, France
[3] Ifremer, Université de Brest, CNRS, IRD, LEMAR, F -29280 Plouzané, France

Abstract

Background: Protein kinases are a diverse superfamily of proteins common to organisms across the tree of life that are typically involved in signal transduction, allowing organisms to sense and respond to biotic or abiotic environmental factors. They have important roles in organismal physiology, including development, reproduction, acclimation to environmental stress, while their dysregulation can lead to disease, including several forms of cancer. Identifying the complement of protein kinases (the kinome) of any organism is useful for understanding its physiological capabilities, limitations and adaptations to environmental stress. The increasing availability of genomes makes it now possible to examine and compare the kinomes across a broad diversity of organisms. Here we present a pipeline respecting the FAIR principles (findable, accessible, interoperable and reusable) that facilitates the search and identification of protein kinases from a predicted proteome, and classifies them according to group of serine/threonine/tyrosine protein kinases present in eukaryotes.

Results: *KiNext* is a Nextflow pipeline that regroups a number of existing bioinformatic tools to search for and classify the protein kinases of an organism in a reproducible manner, starting from a set of amino acid sequences. Conventional eukaryotic protein kinases (ePKs) and atypical protein kinases (aPKs) are identified by using Hidden Markov Models (HMMs) generated from the catalytic domains of kinases. Furthermore, *KiNext* categorizes ePKs into the eight kinase groups by employing dedicated Hidden Markov Models (HMMs) tailored for each group. The performance of the *KiNext* pipeline was validated against previously identified kinomes obtained with other tools that were already published for two marine species, the Pacific oyster *Crassostrea gigas* and the unicellular green alga *Ostreoccocus tauri*. *KiNext* outperformed previous results by finding previously unidentified kinases and by attributing a large proportion of previously unclassified kinases to a group in both species. These results demonstrate improvements in kinase identification and classification, all while providing traceability and reproducibility of results in a FAIR pipeline. The default HMM models provided with *KiNext* are most suitable for eukaryotes, but the pipeline can be easily modified to include HMM models for other taxa of interest.

Conclusion: The *KiNext* pipeline enables efficient and reproducible identification of kinomes based on predicted amino acid sequences (*i.e.* proteomes). *KiNext* was designed to be easy to use, automated, portable and scalable.

Keywords: Kinase, Genome annotation, Workflow

Background

The kinome of a specie is the set of genes encoding protein kinases. Protein kinases are enzymes which catalyze the transfer of a phosphate group from adenosine triphosphate (ATP) to the hydroxyl group (–OH) of side chains on some amino acids. Protein kinases are intracellular proteins, while a very small number of kinases can be secreted, which activate signaling pathways that enable rapid cellular responses crucial for the survival and adaptation of organisms to their environment [1, 2]. They are sensing proteins that transduce signals inside the cell via the phosphorylation of downstream targets, modifying their protein structure, and thereby regulating their activity [1, 3]. Phosphorylation induced by protein kinases regulate numerous cell functions such as DNA replication, cell cycle control, cytoskeletal rearrangement, cell movement, gene transcription, protein translation, apoptosis, differentiation and cell energy metabolism [4]. These processes are vital for numerous physiological functions including development, reproduction, defense and survival.

The structure of protein kinases

The majority of protein kinases possess a catalytic domain involved in the regulation of the kinase activity, the eukaryotic protein kinases (ePK) domain, where the phosphate-donating nucleotide ATP binds to a phosphate-accepting protein as substrate [3, 5]. The ePK domain is highly conserved across species, and is usually 250–300 amino acids in length. Outside the ePK catalytic domain, the amino acid sequences of kinases are often highly divergent making it difficult to identify kinases based on full amino acid sequences [6]. A second category of protein kinases called atypical protein kinases (aPK) are also an important component of the kinome. While aPKs possess biochemical kinase activity, they lack the amino-acid sequence similarity to the ePK domain [7]. Both ePKs and aPKs have both essential roles in cell signaling pathways [8]. Despite their low amino acid sequence similarity with ePKs, most atypical kinases share the same characteristic eukaryotic protein kinase fold and often conserve some level of structural similarity to ePKs [9, 10]. This structural similarity induces similar protein kinase activity, hence the importance of considering aPKs in kinome search.

Eukaryotic protein kinases (ePKs) are separated into 8 distinct groups according to the Manning classification [1], depending on the type of downstream protein target they phosphorylate: the AGC group (cAMP-dependent kinase/protein kinase G/protein kinase C), the CaMK group (Ca2 +/calmodulin-dependent protein kinase), the CK1 group (Casein Kinase I), the CMGC group (comprising several sub-families such as Cyclin-Dependent Kinases (CDK), the Mitogen-Activated Protein Kinases (MAPK) group, Glycogen Synthase Kinase 3 (GSK3) group, the Cyclin-Dependent Kinase-Like (CKL) group, the "Sterile" (STE) group, containing homologs of sterile yeast kinases, the Protein Tyrosine Kinase (TK) group, the Tyrosine Kinase-Like (TKL) group and

the Receptor Guanylate Cyclase (RGC) group [1]. These groups are present in nearly all eukaryotes, and most groups are present in Archaea and Bacteria (although some groups, such as the TK group, have not yet been demonstrated for some prokaryotes) [7].

Identifying the kinome from a genome annotation

Given the importance of protein kinases to vital cellular and physiological processes, there has been growing interest over the past few decades in identifying the full complement of kinases (i.e. the kinome) in the genomes of model species. As genome sequencing and annotation becomes increasingly accessible, the search for protein kinases can be expanded to non-model species, facilitating the understanding of mechanisms of physiological adaptation at the cellular level. Identifying the kinome in species with sequenced and annotated genomes is thus useful for studying the physiology, development and disease beyond model organisms.

In the field of bioinformatics, various tools have been developed for identifying and classifying protein kinases, such as Kinannote [11] or the works of Stroehlein et al. (2018) [12] and KinFams [13]. However, some of these tools do not fully adhere to FAIR principles [14] which are essential for ensuring transparency, reproducibility, and accessibility in scientific research. The kinomes of several species have been published but the associated methodologies employed often differed. Furthermore, the details of the analyses were not always explicitly reported, making it difficult and at times impossible to reproduce results in further studies. Sadly, without this information, comparative studies are likely to incur inaccuracies in the results, generated from biases associated with differences in methodology. These issues motivated the development of a pipeline which would adhere to the FAIR principles [14], so that future kinomes could be produced in a standardized and comparable manner. Moreover, scripting languages such as Python or R are not sufficient to support the development of large-scale pipelines that are shareable, maintainable and reusable, capable of processing large volumes of data and running on high-performance computing clusters [15]. Thus, the Nextflow [16] workflow manager was chosen to simplify the development of the *KiNext* pipeline, optimize the use of its resources, manage the installation and versions of the software with the capabilities of operating on a personal computer as well as on a computing cluster.

In order to guarantee the reproducibility of results and the portability of *KiNext*, containers were built to group together the bioinformatic tools used and to encapsulate their dependencies in an isolated environment using Singularity [17]. The containerization of its software thus guarantees that analyses run in a controlled environment, independently of the differences in operating systems or machine configurations. An important advantage of the *KiNext* pipeline is to provide a predefined and reusable model for identifying and classifying protein kinases based on the similarity of their protein sequences. The kinomes produced by *KiNext* can then be used to study the structural similarity among kinases and kinase groups and improve the construction of their evolutionary trees [8]. *KiNext* is currently accessible from a Gitlab repository made public to the scientific community (Gitlab: https://gitlab.ifremer.fr/bioinfo/workflows/kinext).

Implementation

The *KiNext* pipeline consists of 7 processes for kinome identification and protein kinase classification followed by phylogenetic analysis of the recovered kinases (Fig. 1).

Kinome identification

The search for protein kinases from an annotated genome (*i.e.* proteome) of a given species requires one or more probabilistic models called "Hidden Markov Models" or HMMs profiles. This type of search enables homologous protein sequences to be identified using a probability algorithm. Two separate models were collated for protein kinase identification, one for the identification of ePKs and a second for the identification of aPKs. The catalytic domains of ePKs from *Homo sapiens*, *Caenorhabditis elegans* and *Drosophila melanogaster* were downloaded from Kinbase (The kinase Database. http://www.kinase.com/kinbase/FastaFiles/. Accessed 20 November 2023) and added to a fasta file used to create the HMM model, which was subsequently be used as a reference to the search for ePKs. At the same time, the catalytic domains of human, *D. melanogaster* and *C. elegans* aPKs were added to a separate fasta file, this time to create the HMM model for identifying aPKs.

In order to run, the *KiNext* pipeline requires (1) the proteome (predicted amino acid sequences obtained from automatic genome annotation) in fasta format, (2) files containing the HMM profiles for ePKs and aPKs (here based on kinases from *H. sapiens*, *C. elegans* and *D. melanogaster*, but different HMM models could be provided) and (3) HMM models containing separate HMMs profiles for each kinase groups. The HMM profiles used in the test dataset presented here are available in the *KiNext* Gitlab repository (https://gitlab.ifremer.fr/bioinfo/workflows/kinext).

The **first process** searches for protein sequences homologous to the HMMs profiles of ePKs and aPKs (Fig. 1, step 1) run on the proteome of the given species with HMMSEARCH [18] v3.3.252 and the following options: "–E 0.05" to set the E-value to its standard value, "–tblout" to summarize the results in a table and "–noali" option to remove the alignment from the main output in order to reduce its volume. The output

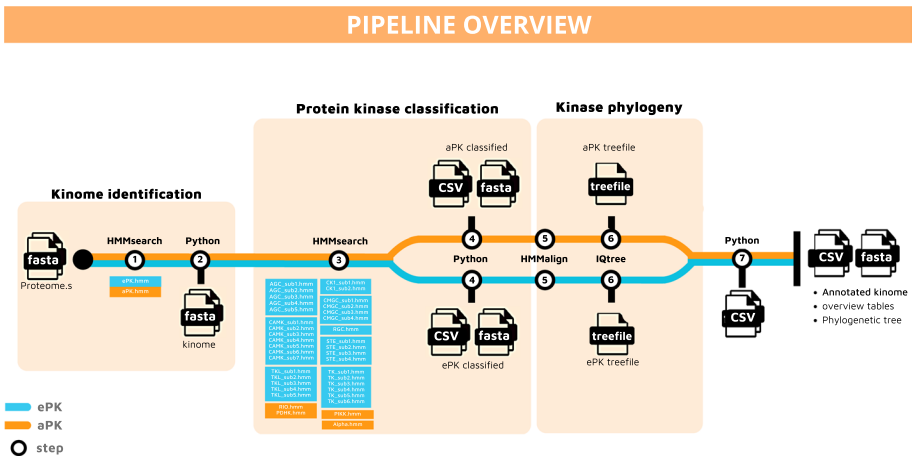

Fig. 1 Overview of the 7 processes from the *KiNext* pipeline

file from this step is a table of identifiers from the proteome homologous to the HMM model of the ePKs and aPKs.

The **second process** is a python script which retrieves the table of sequence identifiers in order to control the assignment to class (ePKs/aPKs) according to the associated e-value (Fig. 1, step 2). The script associates the identifiers with their corresponding fasta sequences extracted from the proteome of the given species. At the end of the second process, 3 fasta files are generated: the complete kinome, and ePKs and aPKs in separate files.

Protein kinase classification

The next objective is to assign protein kinases according to a specific kinase group. This is completed during the **third process**, which uses the previously identified kinome as input to search for sequences homologous to the HMM profiles specific to the kinase groups identified by the *kinomer* project [19] (Fig. 1, step 3). This library used by *KiNext* contains the HMMs profiles of protein kinase families from 4 model organisms: *H. sapiens, D. melanogaster, C. elegans, and S. cerevisiae* [20]. The distinctive feature of the *kinomer* libraries is that each protein kinase group is divided into families. Indeed, Altman et al. [21] found that subdividing large protein groups increased the recognition accuracy of HMMs by allowing the unique characteristics of each family to be captured more accurately. Furthermore, the results of Miranda-Saavedra and Barton [20] indicated that this library of HMM profiles was superior not only to BLAST [22] but also to a general HMM from the catalytic domain of the kinase. This search was carried out using HMMSEARCH v3.3.252 with the following options: "–E 0.05" to set the E-value to its standard value, "–tblout" to summarize the results in a table and "–noali" (Fig. 1, step 3). The output file from this Nextflow process is a table of identifiers homologous to the kinase family profiles.

In the **fourth process**, a python script is used to assign each kinase to the family with the lowest e-value (Fig. 1, step 4). This process ensure that each kinase is attributed to only one kinase group. Kinases that are identified by multiple HMM profiles are listed in the family with the lowest e-value. Once the classification has been checked and validated automatically, this script extracts the fasta sequences corresponding to the identifiers. The end result of this fourth process is the creation of 5 fasta files containing: the complete annotated kinome, the annotated ePKs, the annotated aPKs and another fasta file containing the sequences identified as protein kinases not assigned to a family during the first process of the pipeline. The last fasta file generated corresponds to the full kinome with each kinase annotated to a group (or assigned to the unknown category).

Kinase phylogeny

Once the kinome of the given species is identified, two phylogenetic trees are constructed, one for ePKs and another for aPKs, in order to group kinases according to the similarity of the amino acid sequences. The **fifth process** thus performs sequence alignment and phylogenetic reconstruction of ePKs and aPKs.

Firstly, the fasta file output from the fourth process containing the annotated ePKs is aligned with the HMMALIGN v3.3.2 tool (Fig. 1, step 5). This tool aligns the full kinase sequences with the ePK HMM profile (catalytic domains) and produces a

multiple sequence alignment in ClustalW [23] format. The following options were used in HMMALIGN: the '–trim' option to remove non-homologous residues (assigned to the N and C states in the optimal alignments) from the result of the multiple alignment and the '–outformat' option to write the output alignment in 'clustal' format. The same process is conducted on the annotated aPKs using the aPK HMM profile.

The **sixth process** consists of phylogenetic reconstruction in IQtree [24] v2.1.258 based on the alignments obtained during the fifth process (Fig. 1, step 6). IQTree is set by default for use with *-st AA* (amino acid sequences). The *-m MFP* option allows IQtree to automatically select the best evolutionary substitution model for tree construction. And to estimate the robustness of the tree topology, the "ultrafast bootstrap" method *-bb 1000* and the SH-aLRT *-alrt 1000*, an approximate Shimodaira-Hasegawa likelihood ratio test, were used. The phylogenetic reconstruction of the aPKs can help the *KiNext* user to confirm and refine the classification of atypical protein kinase families. It is important to use these trees with caution and to be aware of their limitations, particularly pertaining to aPKs. The phylogenetic trees generated by IQtree were viewed manually using the ETE3 framework [25] to automatically colored them by kinase group. ETE3 simplifies the reconstruction, analysis, and visualization of phylogenetic trees. An ETE3 python script is available in the *KiNext* Gitlab repository (https://gitlab.ifremer.fr/bioinfo/workflows/kinext).

The **seventh and final process** of the *KiNext* pipeline is a python script that builds two summary tables (.csv) from all the files produced during the *KiNext* run (Fig. 1, step 7). The first table is composed of 6 columns including the identifiers of the protein kinases found by the pipeline, a second column indicating whether they are ePKs or aPKs, the next column indicating the "e-value" obtained for this result, the fourth and fifth columns correspond respectively to the family and the score obtained for this classification, and finally the last column corresponds to the description of the sequence (predicted annotation). The second table summarizes the number of sequences identified for each of the protein kinase families, the count of sequences that were identified as being a protein kinase but for which the pipeline was unable to assign a group, and finally, the total number of sequences making up the kinome. These two final tables provide users an overview of results obtained that can be used as a starting point for more in-depth analyses [1].

Finally, the user can conduct a validation process by checking manually the e-values of the newly predicted kinase domains against those of annotated kinase proteins to ensure consistency and reliability. AlphaFold, a state-of-the-art protein structure prediction tool [26], can be employed by the *KiNext* user to verify the structural integrity of the predicted proteins by modeling the 3D conformation of the protein and then, making a structural superposition against a reference kinase to and verifying the presence of the catalytic domain. Another option is to use Foldseek [27] with the Alphafold results and search for similarities and structural overlap against references databases, like Alphafold/Swiss-Prot. This validation approach was used in trial runs of *KiNext* on the genomes of *C. gigas* and *O. tauri* to verify that the protein kinases identified by the pipeline based on structural similarity.

Results and discussion

KiNext improves kinome identification: validation with *C. gigas* and *O. tauri*

The aim of *KiNext* is to search for, identify, classify and determine the phylogeny of protein kinases in any organism, on the basis of its proteome (annotated genome), by improving and integrating the kinomer strategy and using an automated workflow. The already published, complete, detailed and sufficiently precise kinome of the Pacific oyster *C. gigas* [28] or green alga *O. tauri* [29] were compared to the kinome obtained by *KiNext* (Additional file 1: Table S1–S2). *KiNext* recovered all 371 protein kinases identified for *C. gigas* in previous work [28], plus 21 additional identified kinases (Fig. 2A). In addition, the classification of a number of kinases were updated: 13 additional kinases in the AGC group, 37 in the CAMK group, 5 in the CMGC group, 1 in the RGC group, 16 in the STE group, 7 in the TK group, and 5 in the TKL group (Fig. 2C). Of the 21 new kinases identified by *KiNext*, 12 were classified as ePK and the remaining 9 as aPK. Nineteen could be modeled by AlphaFold and among them, 15 exhibit a kinase catalytic domain, identified with Foldseek using the Alphafold/Swiss-Prot database (Additional file 1: Table S3).

For the green algae *O. tauri*, *KiNext* recovered 118 of the 122 proteins kinases identified in previous work [29], plus 19 additional kinases (Fig. 2B). As with *C. gigas*, the

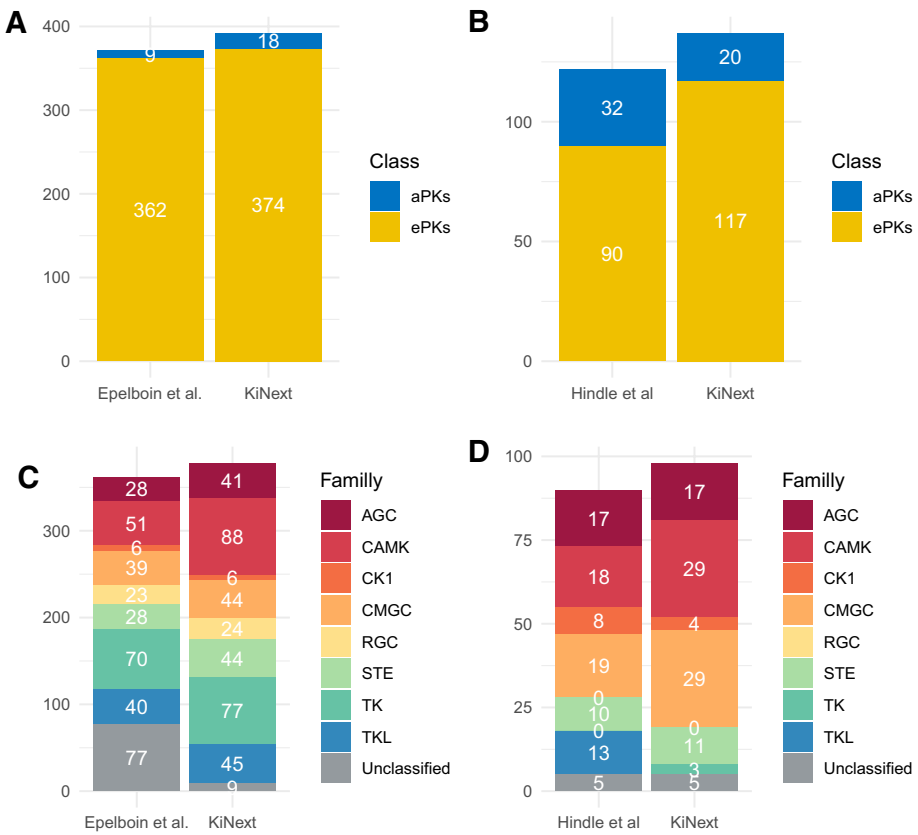

Fig. 2 The kinome of *C. gigas* (**A**, **C**) or *O. tauri* (**B**, **D**) obtained by *Kinext* and compared to previously published kinomes (Epelboin et al. for *C. gigas*; Hindle et al. for *O. tauri*). **A**, **B**: Total number of kinases. **C**, **D**: Number of ePK in each group

classification of a number of kinases were updated. Classification according to *KiNext* placed 11 additional kinases in the CAMK group, 10 in the CMGC group, 1 in the TKL group and 3 in the TK group (Fig. 2D). Ten kinases were re-classified from aPK to ePK. Among the 19 new kinases identified by *KiNext* as ePK, 17 could be modeled by AlphaFold and among them, 16 exhibit a kinase catalytic domain, identified with Foldseek using Alphafold/Swiss-Prot database (Additional file 1: Table S4).

Protein kinase identification and classification for *C. gigas* and *O. tauri* demonstrate that the pipeline functions adequately in two phylogenetically distant species and is likely to be suitable for a broad range of organisms. *KiNext* recovered all previously identified kinases, and further improved upon previous results by returning kinases not previously identified. The pipeline therefore appears to perform as well, if not better, than previous methods, with the added advantage of being reproducible and adhering to FAIR principles. With regard to the classification of kinases, *KiNext* provided group level classifications for all but 9 ePK in the *C. gigas* kinome, while 77 ePK had no group assignment in the previously published kinome. With regard to the classification of kinases in the microalgae *O. tauri*, kinases that had not previously been identified were found and assigned to families. In both test datasets, *KiNext* was able to classify a greater number of kinases compared to previous methods. This improvement in kinase classification is likely the result of generating one HMM per kinase family, as previously reported [20, 21].

KiNext as a new FAIR tool for genome analysis

There is currently a global drive to sequence biodiversity, with numerous genomes sequencing programs currently underway. The "Darwin Tree of Life" (DtOL) which aims to sequence the genomes of 70,000 eukaryotic genomes from organisms in Britain and Ireland [30], the “Vertebrate Genomes Project” (VGP) aims to produce and additional 70,000 genomes from extant vertebrate species [31], while the “Earth Biogenome Project” aims to sequence, catalog and characterize the genomes of all of Earth’s eukaryotic biodiversity over a period of ten years [32], In marine biology, the ATLASea research program (PEPR Atlas of marine genomes) will sequence the genomes of 4,500 species from the French Exclusive Economic Zone (EEZ), including 1,000 from overseas territories, and will be used, among other things, to identify biological functions of industrial or medical interest and to study invasive marine species. Each of these programs will produce thousands of new genomes, which will require structural and functional annotation of important families of genes. Future work that integrates *KiNext*, as well as other tools for gene annotation, into larger annotation workflows will require scalibility. *KiNext* has been designed and implemented in NextFlow with future applications in mind. As science moves forward in the genomics era pipelines such as *KiNext* will facilitate the understanding of protein kinase function and evolution by harnessing the information contained in the multitude of genomes (and kinomes) currently being produced.

Conclusion

The *KiNext* pipeline is a new tool for the analysis of genomes. *KiNext* identifies protein kinases and provides a classification to group and family for any eukaryotic species whose genome is annotated. The pipeline can be readily extended for use with

other organisms such as Archaea and Bacteria, provided a suitable protein kinase HMM model is given as input. Identifying and studying complex networks of kinases open the way to new discoveries in the field of cell signaling and regulatory pathways. Studying kinomes will improve our knowledge on sensing and signaling pathways involved in adaptation and resistance to environmental stress across species, such as pollution or climate change.

Availability and requirements

Project name: *KiNext.*

Project gitlab page: https://gitlab.ifremer.fr/bioinfo/workflows/kinext

Operating system(s): Platform independent.

Programming language: Java (>=8), nextflow v22.10.4, Docker/Apptainer.

Other requirements: Java 8 or higher, Nextflow v22 or higher, Python 3.11 or higher, Apptainer 3.4 or higher.

License: Affero GPL.

Any restrictions to use by non-academics: None.

Abbreviations

AGC	CAMP-dependent kinase/protein kinase G/protein kinase C
ATP	Adenosine tri-phosphate
aPK	Atypical protein
CaMK	Ca2+/calmodulin-dependent protein kinase
CK1	Casein Kinase I
CMGC	CDK, MAPK, GSK3 and CLK containing group
CSV	Comma-separated values
ePK	Eukaryotic protein kinase
FAIR	Findable Accessible Interoperable Reusable
HMM	Hidden Markov Model
PEPR	Priority Research Programmes and Equipment
PTK	Protein tyrosine kinase
RGC	Receptor Guanylate Cyclase
STE	For "sterile", containing homologs of sterile yeast kinases
TK	Protein Tyrosine Kinase
TKL	Similar to Tyrosine Kinases

Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-024-05953-w.

Additional file 1: TablesThe supplementary tables include kinase identification results and functional annotation results with Alphafold and Foldseek

Acknowledgements

The authors would like to Diego Miranda-Saavedra and Geoffrey J. Barton, for providing HMM models from the Kinomer library and for discussions and recommendations. We also thank two anonymous reviewers for their constructive comments.

Author contributions

EH developed the methodology and implemented the pipeline. AC and EH led the writing of the manuscript. CC, FN and AC provided guidance and oversight, helped to refine the design and suggested improvements to the manuscript. AC, CC and FN provided financial support. All authors read and approved the final manuscript.

Funding

Not applicable.

Availability of data and materials

No datasets were generated or analysed during the current study.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing interests
The authors declare that they have no competing interests.

Received: 11 February 2024 Accepted: 7 October 2024
Published online: 25 October 2024

References

1. Manning G, Plowman GD, Hunter T, Sudarsanam S. Evolution of protein kinase signaling from yeast to man. Trends Biochem Sci. 2002;27:514–20.
2. Tagliabracci VS, Pinna LA, Dixon JE. Secreted protein kinases. Trends Biochem Sci. 2013;38:121–30.
3. Hanks SK, Quinn AM, Hunter T. The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. Science. 1988;241:42–52.
4. 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. 1999;96:13603–10.
5. Krebs EG, Beavo JA. Phosphorylation-dephosphorylation of enzymes. Annu Rev Biochem. 1979;48:923–59.
6. Hanks SK. Genomic analysis of the eukaryotic protein kinase superfamily: a perspective. Genome Biol. 2003;4:111.
7. Leonard CJ, Aravind L, Koonin EV. Novel families of putative protein kinases in bacteria and archaea: evolution of the "eukaryotic" protein kinase superfamily. Genome Res. 1998;8:1038–47.
8. Scheeff ED, Bourne PE. Structural evolution of the protein kinase-like superfamily. PLoS Comput Biol. 2005;1: e49.
9. Walker EH, Perisic O, Ried C, Stephens L, Williams RL. Structural insights into phosphoinositide 3-kinase catalysis and signalling. Nature. 1999;402:313–20.
10. Yamaguchi H, Matsushita M, Nairn AC, Kuriyan J. Crystal structure of the atypical protein kinase domain of a trp channel with phosphotransferase activity. Mol Cell. 2001;7:1047–57.
11. Goldberg JM, Griggs AD, Smith JL, Haas BJ, Wortman JR, Zeng Q. Kinannote, a computer program to identify and classify members of the eukaryotic protein kinase superfamily. Bioinformatics. 2013;29:2387–94.
12. Stroehlein AJ, Young ND, Gasser RB. Improved strategy for the curation and classification of kinases, with broad applicability to other eukaryotic protein groups. Sci Rep. 2018;8:6808.
13. Adeyelu T, Bordin N, Waman VP, Sadlej M, Sillitoe I, Moya-Garcia AA, et al. KinFams: de-novo classification of protein kinases using CATH functional units. Biomolecules. 2023;13:277.
14. Wilkinson MD, Dumontier M, Aalbersberg IJ, Appleton G, Axton M, Baak A, et al. The FAIR guiding principles for scientific data management and stewardship. Sci Data. 2016;3:160018.
15. Djaffardjy M, Marchment G, Sebe C, Blanchet R, Bellajhame K, Gaignard A, et al. Developing and reusing bioinformatics data analysis pipelines using scientific workflow systems. Comput Struct Biotechnol J. 2023;21:2075–85.
16. Tommaso PD, Chatzou M, Floden EW, Barja PP, Palumbo E, Notredame C. Nextflow enables reproducible computational workflows. Nat Biotechnol. 2017;35:316–9.
17. Kurtzer GM, Sochat V, Bauer MW. Singularity: Scientific containers for mobility of compute. PLoS ONE. 2017;12: e0177459.
18. Eddy SR. Accelerated profile HMM searches. PLoS Comput Biol. 2011;7: e1002195.
19. Martin DMA, Miranda-Saavedra D, Barton GJ. Kinomer v. 1.0: a database of systematically classified eukaryotic protein kinases. Nucleic Acids Res. 2009;37(1):D244–50.
20. Miranda-Saavedra D, Barton GJ. Classification and functional annotation of eukaryotic protein kinases. Proteins Struct Funct Bioinform. 2007;68:893–914.
21. Altman RB, Jung TA, Klein TE, Dunker AK, Hunter L, Brown D, et al. Subfamily HMMS in functional genomics. Biocomput. 2005;2004:322–33.
22. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
23. Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 1994;22:4673–80.
24. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. Mol Biol Evol. 2020;37:1530–4.
25. Huerta-Cepas J, Serra F, Bork P. ETE 3: reconstruction, analysis, and visualization of phylogenomic data. Mol Biol Evol. 2016;33:1635–8.
26. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021;596:583–9.
27. van Kempen M, Kim SS, Tumescheit C, Mirdita M, Lee J, Gilchrist CLM, et al. Fast and accurate protein structure search with Foldseek. Nat Biotechnol. 2024;42:243–6.
28. Epelboin Y, Quintric L, Guévélou E, Boudry P, Pichereau V, Corporeau C. The kinome of pacific oyster crassostrea gigas, its expression during development and in response to environmental factors. PLoS ONE. 2016;11: e0155435.
29. Hindle MM, Martin SF, Noordally ZB, van Ooijen G, Barrios-Llerena ME, Simpson TI, et al. The reduced kinome of Ostreococcus tauri: core eukaryotic signalling components in a tractable model species. BMC Genom. 2014;15:640.

30. Consortium TDT of LP, Blaxter M, Mieszkowska N, Palma FD, Holland P, Durbin R, et al. Sequence locally, think globally: the darwin tree of life project. Proc Natl Acad Sci. 2022;119:1.
31. Rhie A, McCarthy SA, Fedrigo O, Damas J, Formenti G, Koren S, et al. Towards complete and error-free genome assemblies of all vertebrate species. Nature. 2021;592:737–46.
32. Lewin HA, Robinson GE, Kress WJ, Baker WJ, Coddington J, Crandall KA, et al. Earth biogenome project: sequencing life for the future of life. Proc Natl Acad Sci. 2018;115:4325–33.

Publisher's Note