RESEARCH NOTE

Open Access

KINtaro: protein kinase-like database

Bartosz Baranowski[1], Marianna Krysińska[2] and Marcin Gradowski[2*]

Abstract

Objective The superfamily of protein kinases features a common Protein Kinase-like (PKL) three-dimensional fold. Proteins with PKL structure can also possess enzymatic activities other than protein phosphorylation, such as AMPylation or glutamylation. PKL proteins play a vital role in the world of living organisms, contributing to the survival of pathogenic bacteria inside host cells, as well as being involved in carcinogenesis and neurological diseases in humans. The superfamily of PKL proteins is constantly growing. Therefore, it is crucial to gather new information about PKL families.

Results To this end, the KINtaro database (http://bioinfo.sggw.edu.pl/kintaro/) has been created as a resource for collecting and sharing such information. KINtaro combines protein sequence information and additional annotations for more than 70 PKL families, including 32 families not associated with PKL superfamily in established protein domain databases. KINtaro is searchable by keywords and by protein sequence and provides family descriptions, sequences, sequence alignments, HMM models, 3D structure models, experimental structures with PKL domain annotations and sequence logos with catalytic residue annotations.

Keywords Protein kinase, Pseudokinase, Phosphotransferase, Structure prediction, HMM, Phosphorylation

Introduction

Kinases are among the most crucial enzymes found in all living organisms. They facilitate phosphorylation reactions, transferring phosphate groups from high-energy compounds like ATP to specific target molecules. Within the PKL superfamily, best known are protein kinases responsible for phosphorylating proteins [1]. Additionally, in the PKL superfamily there are small molecule kinases whose substrates include antibiotics and sugars [2], as well as lipid kinases that target membrane lipids like phospholipids and sphingolipids [3–5].

PKL proteins play critical roles in various biological processes, including cell growth, differentiation, and apoptosis. Dysregulation of these proteins can contribute to the development of numerous diseases, including tumorigenesis [6]. Moreover, PKL proteins can act as promoters of antibiotic resistance [2], aid pathogen survival within host cells [5, 7], and serve as effectors influencing cellular processes in affected cells [8]. Consequently, blocking their activity through various types of inhibitors can be crucial in preventing diseases, infections, and treating cancer [9] providing alternative treatments.

Pseudokinases were initially considered to be non-functional relatives of protein kinases that lost their enzymatic activity due to mutations [10, 11]. However, recent studies have revealed that pseudokinases can exhibit alternative enzymatic activities. For example, the coronavirus NiRAN pseudokinase domain transfers nascent RNA to GDP, using an RNA–protein intermediate, and ultimately forming the core RNA cap structure: GpppA-RNA [12]. The SelO pseudokinase performs AMPylation of proteins involved in redox homeostasis [13]. The bacterial pseudokinase effector SidJ polyglutamylates SidE effectors, blocking their activity which consists of phosphoribosyl

*Correspondence:
Marcin Gradowski
marcin_gradowski@sggw.edu.pl
[1] Laboratory of Plant Pathogenesis, Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warsaw, Poland
[2] Department of Biochemistry and Microbiology, Warsaw University of Life Sciences (SGGW), Warsaw, Poland

Baranowski *et al. BMC Research Notes*   (2024) 17:50

Page 2 of 5

ubiquitination of host Rab GTPases to evade phagocytosis [14], thus modulating the effect on the host cell. Pseudokinases can also serve as allosteric regulators of protein kinases, influencing their activity [15] or stasis for other proteins (for example as part of the secretion system of bacteria) [16].

A number of databases related to protein kinases are known, e.g., the best known database of human kinases according to Manning's classification [17] or the database of protein kinases in genomes—KinG [18], which is based on Pfam [19] domains. The Pfam domains are not always well defined in terms of domain boundaries, e.g., the PIP49_C family does not cover the entire PKL fold [20]. The Pan3_PK pseudokinase family lacks the kinase N-lobe [21]. Moreover, the Pfam clan (superfamily) Pkinase does not include all known PKL families e.g., SelO pseudokinase family—involved in redox homeostasis [13] or FAM198 family which has been recently identified as a potential cancer-associated gene [22]. Other examples are Pox_E2-like—a pseudokinase found in Poxviridae [23] or the CLU [24] pseudokinase present in eukaryotes. In addition, a lot of PKL families are not recognized as domains in the Pfam base, for example, the pseudokinase SidJ [14] or the viral pseudokinase NiRAN [12].

The InterPro database, which absorbed Pfam is still missing many known PKL families [25].

Other databases dedicated to protein kinases are specialized, e.g., KLIFS—a database based on structural knowledge allowing to navigate in the space of kinase-ligand interactions [26], KinaseMD—a database collecting most updated information on mutations, unique annotations of drug response, especially drug resistance and functional sites of kinases [27], BYKdb—Bacterial tYrosine-Kinase database [28]. There is no specialized database collecting information on all the proteins that share the common PKL structure.

Earlier, we studied the pan-proteome of the *Legionella* genus bioinformatically. Some of the *Legionella* PKL families seem to be unique to this bacterium [29].

Together with information from our own research, databases and literature our database contains 72 updated and carefully prepared PKL families (Additional file 1: Table S1) and basic information about each family from all domains of life. The available 3D structure models and domain structures can help in search strategies for further PKL homologs [30].
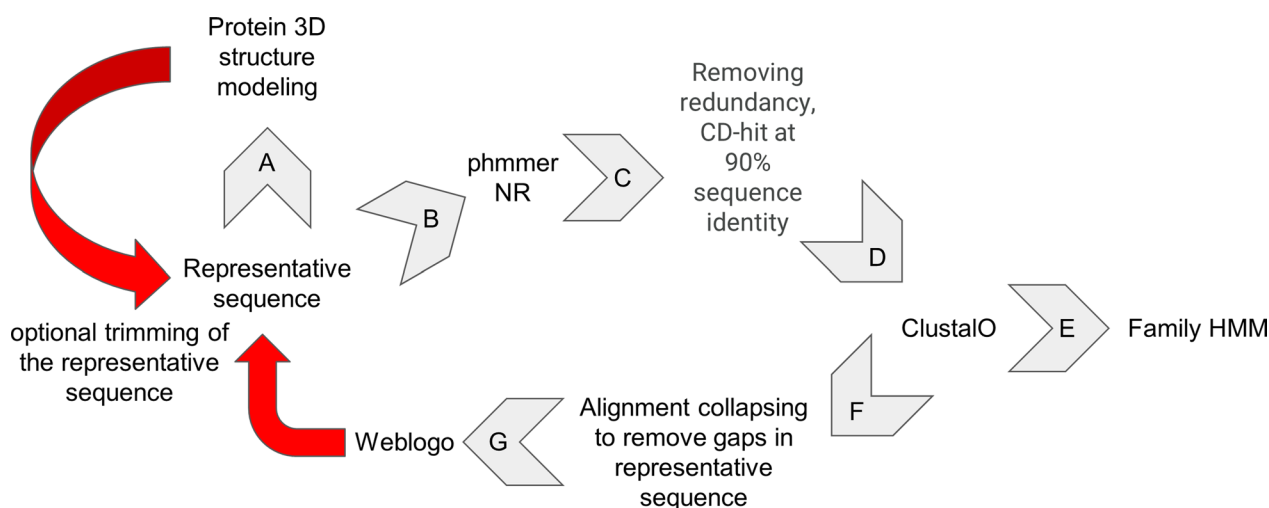
We believe that our semi-automatic approach of constructing the PKL domain family sequence models based on the protein structure model is better than automatic approaches used in other protein domain databases.

The main value of the database lies in its searchable presentation of 32 novel annotated families, previously unrecognized as PKL, along with the assignment of active sites to each family.
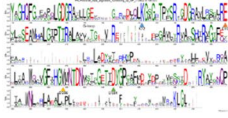
Methods and materials
KINtaro protein family model
For defining protein kinase families, we adopted an approach similar to the protein database Pfam [19], now part of InterPro [25]. However, Pfam's "PKinase (CL0016)" protein clan as mentioned before was not adequately updated, and their family models were not always accurate [19]. In our pipeline, we initiated the process of defining a new family with a representative sequence. These sequences were obtained from existing PKL families in Pfam, and also, for families missing in Pfam, from known 3D structures possessing the PKL

Fig. 1 Family model construction scheme

Baranowski *et al. BMC Research Notes*   (2024) 17:50

Page 3 of 5

FAMILY: SelO (PKLF000038)

Description

SelO pseudokinase displays a protein kinase-like fold, with ATP flipped in the active site, thus providing a structural basis for catalysis. SelO are localized to the mitochondria and AMPylate proteins involved in redox homeostasis. SelO activity plays a crucial role in the cellular response to oxidative stress [PMID: 30270044].

Origin

phmmer: 0.0001; database: nr; sequence_cutoff: 100aa; clustering: cdhit: 90%; catalytic residues: UNUSUAL (SelO pseudokinase lacks the catalytic Asp (PKA D166)), based on - collapsed family logo, 3D structure model, HHpred, PDB(6EAC), PMID: 30270044

Model

Download model pLDDT 96.8331700860357

Structure

Download structure

PKL domain

Download PKL domain

HMM Model

Download HMM model

WebLogo

Sequences

Aligned domain sequences
Download sequence

Unaligned domain sequences
Download sequence

Full sequences
Download sequence

External databases

- Pfam: PF02696
- InterPro: IPR003846

Fig. 2 Family card interface

fold, from novel PKL families described in the literature or from our own sequence/structure searches. Such representative sequences served as a query for 3D structure modeling. Model was created based on the representative sequence (Fig. 1, arrow A) using ColabFold (AlphaFold2

using MMseqs2) or ESMfold, the final model was chosen based on the pLDDT score [31, 32].

To find all members of a PKL family, a representative sequence also served as a query for phmmer [33] against the NR database [34] with an E-value threshold of 0.0001 (Fig. 1, arrow B). Next, we filtered out homologous sequences shorter than 100 amino acids and clustered them at 90% sequence identity [35] (Fig. 1, arrow C). The clustered sequences were then aligned using the ClustalO program [36] to build the family's hidden Markov model (HMM) [33] (Fig. 1, arrow D and E). The alignment was collapsed, where gaps were removed from the representative sequence (Fig. 1, arrow F). A sequence logo was generated from the collapsed alignment using Weblogo [37] (Fig. 1, arrow G). In the final optional step, an iterative approach was used to enhance the family model by adjusting the domain boundaries, where we evaluated the collapsed logo and structure model (Fig. 1, red arrows). For convenience, in the database in the "Family" tab (Fig. 2), the "origin" of the family is recorded, which includes the parameters used and information about any customized steps used in family model construction.

Two large and highly similar Pfam families PF00069 (Pkinase) and PF07714 (PK_Tyr_Ser-Thr) were combined into one family of classical kinases PKLF000033 (Pkinase). Instead of using phmmer, here we employed HMMsearch (with an E-value threshold of 0.0001) and HMM [33] derived from seed alignments (PF00069 and PF07714) from the Pfam database [19]. This HMM was employed to gather homologs, which were then clustered at a 30% sequence identity level.

Each family is assigned a unique identifier (Additional file 1: Table S1; Fig. 2), beginning sequentially with the abbreviation "Protein Kinase-Like—PKL+F" followed by the family's ordinal number. Additionally, each family possesses its own distinctive name.

Results

Database implementation

All PKL families and their relevant information were deposited into a local postgreSQL database. The KINtaro database website (http://bioinfo.sggw.edu.pl/kintaro/) was developed with the Django framework on a Linux machine. All KINtaro data is accessible for all users without registration or login. One can register to maintain sequence search history.

What KINtaro provides

KINtaro offers concise descriptions in family cards (Fig. 2) along with sequence logos collapsed to representative sequences [36] with annotated catalytic residues (when possible) corresponding to canonical kinase

Baranowski *et al. BMC Research Notes*   (2024) 17:50

Page 4 of 5

catalytic residues. The active site assignments (as originally described by Hanks) is based on literature [1], family sequence logos, 3D structure models, known structures and homology. Family structure models are provided, generated using either AlphaFold2 [31] or ESMfold [32]. Additionally, curated representative protein structures from PDB and individual PKL domain structures are provided [38]. The database also includes, for every family, a HMM sequence model, sets of full and clustered sequences of family members, accompanied by their alignments, full sequences containing the PKL domain and links to external databases. Family HMMs can be used to enrich, for example, genomic annotations. The provided sets of PKL sequences can be used for example, for finding new families (e.g. by cluster analysis through quasi-distances between sequences [39]). Structures and models, as mentioned earlier, can be used to search for distant kinase homologs [30]. Such a well-curated dataset can support research into novel (pseudo) enzymatic PKL families.

PKL family search in KINtaro
KINtaro enables users to conduct PKL domain searches with their own sequences using HMMscan (HMMER [33]). KINtaro also is searchable by keywords.

Conclusions
The family of proteins with the PKL fold is continuously expanding. In 2020, we counted over 50 families [40], and in 2022, nearly 70 [29]. Primarily composed of kinases, this group also includes proteins with diverse enzymatic functions and proteins with non-enzymatic roles [15, 16]. To summarize, our database represents a meticulously curated compilation of PKL proteins, serving as a comprehensive and up-to-date resource for information on this rapidly expanding protein superfamily.

Limitations
For some novel families, the PKL assignment is not experimentally confirmed but only predicted by sequence and structure similarities.

Abbreviations
PDB   Protein Data Bank
PKL   Protein Kinase-like
pLDDT   Per-residue confidence metric of structure models

Supplementary Information
The online version contains supplementary material available at https://doi.org/10.1186/s13104-024-06713-y.

Additional file 1: Table S1. KINtaro families. Columns: ID_family—KINtaro family id, name_family—KINtaro family name, pfam_id—pfam family id, interpro_id—interpro family id.

Declarations

Ethics approval and consent to participate
Not applicable.

Consent for publication
Not applicable.

Competing Interests
The authors declare there are no competing interests.

References
1. Hanks SK, Hunter T. The eukaryotic protein kinase superfamily: kinase (catalytic) domain structure and classification [1]. FASEB J. 1995;9(8):576–96.
2. Kannan N, Taylor SS, Zhai Y, Venter JC, Manning G. Structural and functional diversity of the microbial kinome. PLOS Biol. 2007;5(3): e17.
3. Itoh T, Ishihara H, Shibasaki Y, Oka Y, Takenawa T. Autophosphorylation of type I phosphatidylinositol phosphate kinase regulates its lipid kinase activity. J Biol Chem. 2000;275(25):19389–94.
4. Heath CM, et al. Lipid kinases play crucial and multiple roles in membrane trafficking and signaling. Histol Histopathol. 2003;18:989–98.
5. Li G, Liu H, Luo ZQ, Qiu J. Modulation of phagosome phosphoinositide dynamics by a Legionella phosphoinositide 3-kinase. EMBO Rep. 2021;22(3): e51163.
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. Dong N, Niu M, Hu L, Yao Q, Zhou R, Shao F. Modulation of membrane phosphoinositide dynamics by the phosphatidylinositide 4-kinase activity of the Legionella LepB effector. Nat Microbiol. 2016;2:16236–16236.
8. St. Louis BM, Quagliato SM, Lee PC. Bacterial effector kinases and strategies to identify their target host substrates. Front Microbiol. 2023. https://doi.org/10.3389/fmicb.2023.1113021.
9. Castelo-Soccio L, Kim H, Gadina M, Schwartzberg PL, Laurence A, O'Shea JJ. Protein kinases: drug targets for immunological disorders. Nat Rev Immunol. 2023;15:1–20.
10. James MM, Peter DM, Patrick AE. Live and let die: insights into pseudoenzyme mechanisms from structure. Curr Opin Struct Biol. 2017;5(47):95–104.
11. Murphy JM, Farhan H, Eyers PA. Bio-Zombie: the rise of pseudoenzymes in biology. Biochem Soc Trans. 2017;45(2):537–44.
12. Park GJ, Osinski A, Hernandez G, Eitson JL, Majumdar A, Tonelli M, et al. The mechanism of RNA capping by SARS-CoV-2. Nature. 2022;609(7928):793–800.

Baranowski *et al. BMC Research Notes*   (2024) 17:50

Page 5 of 5

13. Sreelatha A, Yee SS, Lopez VA, Park BC, Kinch L, Pilch S, et al. Protein AMPylation by an evolutionarily conserved pseudokinase. Cell. 2018;175(3):809-821.e19.

14. Black MH, Osinski A, Gradowski M, Servage KA, Pawłowski K, Tomchick DR, et al. Bacterial pseudokinase catalyzes protein polyglutamylation to inhibit the SidE-family ubiquitin ligases. Science. 2019;364(6442):787–92.

15. Zhang H, Zhu Q, Cui J, Wang Y, Chen MJ, Guo X, et al. Structure and evolution of the Fam20 kinases. Nat Commun. 2018;23(9):1218.

16. Tassinari M, Doan T, Bellinzoni M, Chabalier M, Ben-Assaya M, Martinez M, et al. The antibacterial type VII secretion system of *Bacillus subtilis*: structure and interactions of the Pseudokinase YukC/EssB. MBio. 2022;13(5): e0013422.

17. Kinase.com. 2023. http://kinase.com/web/current/. Accessed 10 May 2023.

18. Krupa A, Abhinandan KR, Srinivasan N. KinG: a database of protein kinases in genomes. Nucleic Acids Res. 2004;32(1):D153–5.

19. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: the protein families database in 2021. Nucleic Acids Res. 2021;49(D1):D412–9.

20. Tennant-Eyles AJ, Moffitt H, Whitehouse CA, Roberts RG. Characterisation of the FAM69 family of cysteine-rich endoplasmic reticulum proteins. Biochem Biophys Res Commun. 2011;406(3):471–7.

21. Christie M, Boland A, Huntzinger E, Weichenrieder O, Izaurralde E. Structure of the PAN3 pseudokinase reveals the basis for interactions with the PAN2 deadenylase and the GW182 proteins. Mol Cell. 2013;51(3):360–73.

22. Zheng X, Chen J, Nan T, Zheng L, Lan J, Jin X, et al. FAM198B promotes colorectal cancer progression by regulating the polarization of tumor-associated macrophages via the SMAD2 signaling pathway. Bioengineered. 2023;13(5):12435–45.

23. Gao WND, Gao C, Deane JE, Carpentier DCJ, Smith GL, Graham SC. The crystal structure of vaccinia virus protein E2 and perspectives on the prediction of novel viral protein folds. J Gen Virol. 2022;103(1): 001716.

24. Schaeffer RD, Zhang J, Kinch LN, Pei J, Cong Q, Grishin NV. Classification of domains in predicted structures of the human proteome. Proc Natl Acad Sci. 2023;120(12): e2214069120.

25. Paysan-Lafosse T, Blum M, Chuguransky S, Grego T, Pinto BL, Salazar GA, et al. InterPro in 2022. Nucleic Acids Res. 2023;51(D1):D418–27.

26. Kanev GK, de Graaf C, Westerman BA, de Esch IJP, Kooistra AJ. KLIFS: an overhaul after the first 5 years of supporting kinase research. Nucleic Acids Res. 2021;49(D1):D562–9.

27. Hu R, Xu H, Jia P, Zhao Z. KinaseMD: kinase mutations and drug response database. Nucleic Acids Res. 2021;49(D1):D552–61.

28. Jadeau F, Grangeasse C, Shi L, Mijakovic I, Deléage G, Combet C. BYKdb: the Bacterial protein tYrosine Kinase database. Nucleic Acids Res. 2012;40(D1):D321–4.

29. Krysińska M, Baranowski B, Deszcz B, Pawłowski K, Gradowski M. Pan-kinome of Legionella expanded by a bioinformatics survey. Sci Rep. 2022;12(1):21782.

30. Black MH, Gradowski M, Pawłowski K, Tagliabracci VS. Methods for discovering catalytic activities for pseudokinases. Methods Enzymol. 2022;667:575–610.

31. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. Nat Methods. 2022;19(6):679–82.

32. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Language models of protein sequences at the scale of evolution enable accurate structure prediction. bioRxiv. 2022. https://doi.org/10.1101/2022.07.20.500902v1.

33. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. Nucleic Acids Res. 2011;39(2):W29-37.

34. Sayers EW, Bolton EE, Brister JR, Canese K, Chan J, Comeau DC, et al. Database resources of the national center for biotechnology information. Nucleic Acids Res. 2021;50(D1):D20–6.

35. Li W, Jaroszewski L, Godzik A. Tolerating some redundancy significantly speeds up clustering of large protein databases. Bioinformatics. 2002;18(1):77–82.

36. Sievers F, Higgins DG. Clustal Omega for making accurate alignments of many protein sequences. Protein Sci. 2018;27(1):135–45.

37. Crooks GE, Hon G, Chandonia JM, Brenner SE. WebLogo: a sequence logo generator. Genome Res. 2004;14(6):1188–90.

38. Burley SK, Bhikadiya C, Bi C, Bittrich S, Chao H, Chen L, et al. RCSB Protein Data Bank (RCSB.org): delivery of experimentally-determined PDB structures alongside one million computed structure models of proteins from artificial intelligence/machine learning. Nucleic Acids Res. 2023;51(D1):D488-508.

39. Frickey T, Lupas A. CLANS: a Java application for visualizing protein families based on pairwise similarity. Bioinformatics. 2004;20(18):3702–4.

40. Gradowski M, Baranowski B, Pawłowski K. The expanding world of protein kinase-like families in bacteria: forty families and counting. Biochem Soc Trans. 2020;48(4):1337–52.

Publisher's Note