Hindawi Publishing Corporation
ISRN Computational Biology
Volume 2013, Article ID 417634, 5 pages
http://dx.doi.org/10.1155/2013/417634

*Research Article*

# HomoKinase: A Curated Database of Human Protein Kinases

**Suresh Subramani, Saranya Jayapalan, Raja Kalpana, and Jeyakumar Natarajan**

*Data Mining and Text Mining Laboratory, Department of Bioinformatics, Bharathiar University, Coimbatore, Tamil Nadu 641 046, India*

Correspondence should be addressed to Jeyakumar Natarajan; n.jeyakumar@yahoo.co.in

Received 4 March 2013; Accepted 27 March 2013

Academic Editors: G. Colonna and F. Fanelli



HomoKinase database is a comprehensive collection of curated human protein kinases and their relevant biological information. The entries in the database are curated by three criteria: HGNC approval, gene ontology-based biological process (protein phosphorylation), and molecular function (ATP binding and kinase activity). For a given query protein kinase name, the database provides its official symbol, full name, other known aliases, amino acid sequences, functional domain, gene ontology, pathways assignments, and drug compounds. In addition, as a search tool, it enables the retrieval of similar protein kinases with specific family, subfamily, group, and domain combinations and tabulates the information. The present version contains 498 curated human protein kinases and links to other popular databases.

## 1. Introduction

In human genome, the protein kinase is one of the largest recognized protein families which regulate multiple biological processes by posttranslational phosphorylation of serine, threonine, and tyrosine residues [1]. Human genome contains 500 protein kinase genes that constitute about 2% of all genes [2]. Approximately 2000 protein kinases are encoded by human genome. Protein kinases and phosphatases play an important role in regulating and coordinating aspects of metabolism, cell growth, cell motility, cell differentiation and cell division, and signaling pathways involved in normal development and disease [3]. In human genome, 30% to 50% of proteins may undergo phosphorylation; therefore, improper functioning of kinase may lead to various human diseases [4]. Turning on and off of protein kinases and phosphatases maintains the functions of the cellular life in a systematic manner. Further, protein kinases are involved in regulation of many processes, so they are linked to many diseases and act as target for drug design. Protein kinases are the group of enzymes that share conserved catalytic domains involved in stimulating catalytic activity of enzymes and act as ATP binding sites. This result the need and availability of databases specific to protein kinases.

There are many databases for protein kinases present, which include human protein kinases information as well [2, 5, 6]. For example, KinBase [2] contains manually curated kinomes based on Hanks and Hunter classification for nine genomes including humans. KinG [5] contains protein kinases entries for 40 genomes that have been classified by kinome-based sequence search methods. KinWeb [6] is a specific collection of protein kinases encoded in the human genome, and the classification is based on the same orthologous groups present in human and other similar lineages. However, none of the above databases offers high accuracy in classification of human protein kinases due to their underlying classification algorithm. Further, they do not have the options for the retrieval of protein kinases with specific family, subfamily, group, and domain combinations with easy-to-use interface. In this present work, we developed curated human protein kinases database known as "HomoKinase." First, each entry in the database was checked with HGNC to confirm whether it is approved or not. The HGNC approved entry was further confirmed by gene ontology (GO) information based on the presence of three GO terms: (i) ATP binding, (ii) kinase activity, and (iii) protein phosphorylation. The easy-to-use web interface of HomoKinase is shown in Figure 1.


Figure 1: Web interface of HomoKinase.

## 2. Materials and Methods

The HomoKinase database creation involves several steps. First, human genes with their known aliases were downloaded from Entrez Gene (http://www.ncbi.nlm.nih.gov/gene) using the query term "(*Homo sapiens* [Organism]) AND HGNC." Next, the retrieved gene list was crosschecked with the HUGO Gene Nomenclature Committee (HGNC) (http://www.genenames.org/) database to include only the genes with HGNC approved gene name for building the database [7]. The other genes in the list such as pseudogenes, noncoding RNAs, and phenotype which have no HGNC approved name were eliminated.

Finally, gene ontology based refinement was performed to classify the protein kinase genes from the HGNC approved list of human protein-coding genes. In general, GO is mainly focused on three significant ontology terms such as molecular function, cellular component, and biological process. A single gene product may be annotated to multiple GO terms, detailing a range of functional attributes, using both manual and electronic annotation methods [8, 9]. The conserved protein kinase core consists of two lobes: a smaller N-terminal lobe (N-lobe) with ATP binding site and a larger C-terminal lobe (C-lobe) with catalytic site responsible for kinase activity [3, 10]. In addition, the biological processes correspond to protein phosphorylation. These three unique terms of gene ontology (GO) provide precise information about the annotated gene, gene products, and other terms which in turn provide a deep insight about kinases to the researchers. So, we classify the HGNC approved human genes which confirms these three GO terms: (i) ATP binding, (ii) kinase activity, and (iii) protein phosphorylation as true protein kinases. Gene ontology search was performed using two web tools, namely, Quick Go [11] and Amigo Go [12] with automated PHP scripts. The HGNC approved human genes, which satisfy all these three GO criteria, were classified as human protein kinases and used to build the database.

The predicted list of protein kinases were further divided into groups, families, subfamilies, and domains. The group classifications were done using the PhosphoSite database [13], whereas the superfamily, family, subfamily, and domain level classifications were retrieved from UniProt [14]. In addition, various biological information such as official symbol, full name, biological IDs, other known aliases, amino acid sequences, functional domain, gene ontology, pathway assignments, and drug compounds were extracted from various biological databases such as (i) NCBI, (ii) UniProt, (iii) Amigo Go, (iv) KEGG, and (v) DrugBank. Figure 2 depicts a schematic summary of the HomoKinase data warehouse creation process.

The curated human protein kinase names and their related information retrieved from other databases were used to develop the HomoKinase database. The HomoKinase database is implemented as client/server architecture with easy-to-use web interface. The server is made of MySQL database, and the web client and programs for the human protein kinase retrieval, annotation, and query interface were designed using PHP programming language.

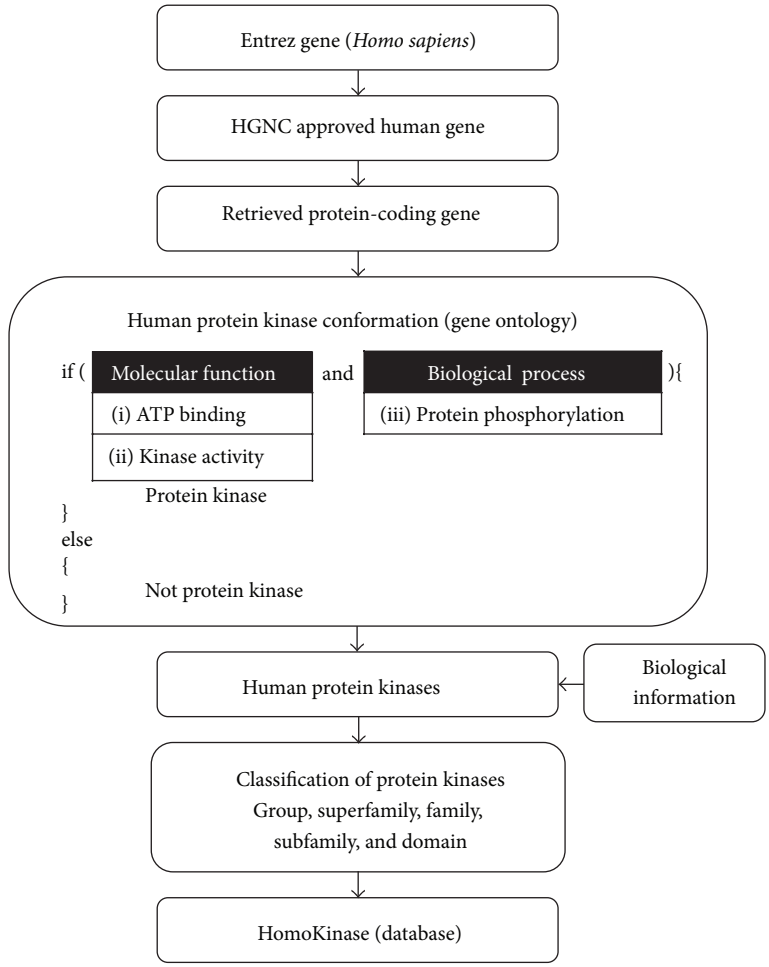


Figure 2: Architecture of HomoKinase database.

Entrez Gene stores information on 1,93,709 genes specific to *Homo sapiens* (as on October 2012). We retrieved 33,489 human genes/proteins specific to our query term "(*Homo sapiens* [Organism] AND HGNC)." On further comparison with HGNC database, only the 19,026 genes have official HGNC gene symbol, and the remaining were 8399 pseudogenes, 4230 noncoding RNAs, 707 phenotype, and 1127 other genes.

The 19,032 HGNC approved human genes were further classified into protein kinases by checking the presence of three GO annotation terms (i) ATP binding property, (ii) kinase activity, and (iii) protein phosphorylation property. The HGNC approved genes fulfilling the above three GO properties (e.g., CDK1, MARK1) were classified as protein kinases and included in the database. Protein kinases missing any one of the above GO properties were filtered and eliminated as nonprotein kinase. The examples of proteins with missing kinase information were (i) absence of ATP binding (e.g., PRKAG2, ADCK4), (ii) absence of kinase activity (e.g., ACTR2, EPHA8), and (iii) absence of protein phosphorylation (e.g., RIOK1, TRIB2). In addition, few genes with lipid kinase activity (e.g., PIK3C2B) and nonprotein kinase (e.g., CKM) were also filtered out. The GO curation and filtration resulted in 498 human genes marked as validated human protein kinases which were included in the final HomoKinase database.

## 3. Results

The HomoKinase database was compared with KinBase [2] and KinWeb [6], the two currently available databases which include human protein kinases. KinBase consists of 506 entries, whereas KinWeb contains 508 entries. HomoKinase excludes the genes which were not approved by HGNC (e.g., NIM1, MST4 in KinBase; ZAK, SgK223 in KinWeb) and genes without proper GO kinase annotation (e.g., ADCK4, TRRAP in KinBase; BRDT, SRM in KinWeb). As a result the number of entries in HomoKinase is reduced into 498. In addition, some of the other common mistakes identified in both databases include (i) gene ID replaced with another gene ID (e.g., SPEG in KinBase; TAO2, Trad in KinWeb), (ii) genes without information on Entrez Gene ID (e.g., sgk424 in KinBase and KinWeb), and (iii) pseudogenes (e.g., PRKY in KinBase and KinWeb). In total, we identified 31 genes with incomplete information (such as error in gene ID and gene

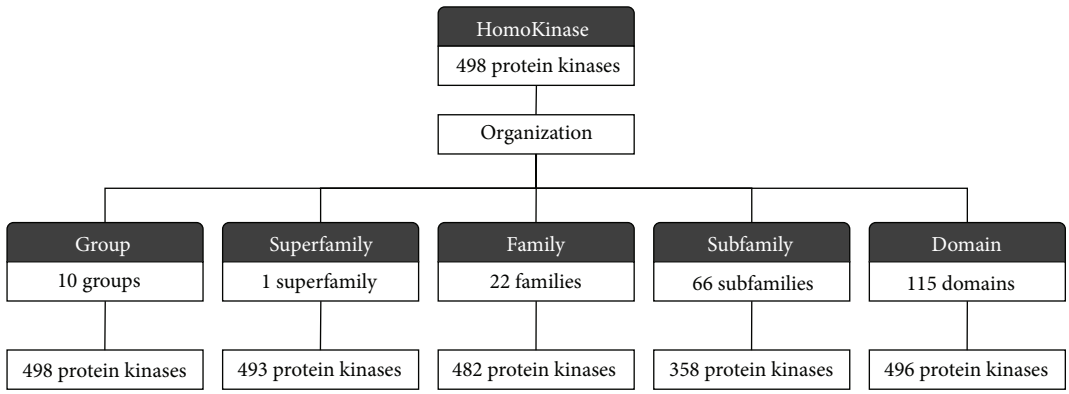


Figure 3: HomoKinase database organization.

name) in KinBase and 8 genes in KinWeb. Table 1 shows the overall comparison of the three databases.

## 4. Discussion

We have developed a curated database of human protein kinases. The salient feature of HomoKinase database is that it provides individual protein name search as well as group search (e.g., family, subfamily, domain, etc.). Individual search can be carried out by giving official symbol (provided by HGNC), Entrez Gene ID, HGNC ID, Ensembl ID, and UniProt ID) and other aliases/designations. The group search can be carried out by classification of protein kinases into different kinase groups, families, subfamilies, and domains. The different group classification of protein kinases in HomoKinase is discussed below.

The 498 human protein kinases entries in the database were classified into 10 groups, 1 superfamily, 22 families, 66 subfamilies, and 115 domains. All 498 protein kinases fall in any one of the 10 groups. However, only 482 proteins were classified into 22 families, and 14 proteins do not belong to any family. Further, 358 proteins belong to 66 subfamilies, whereas for 140 proteins, the subfamily information is missing. In addition, each protein has one-to-many domains, and in total, 115 domains were found among 496 kinases. The database group search can be performed using any one of the above classes. The group search lists out all protein kinases that belong to that search category in a tabular form from which individual protein search can be carried out. The HomoKinase database classification and organization is shown in Figure 3.

## 5. Conclusion

In summary, HomoKinase is an easy-to-use interface to a curated database of human protein kinases. We plan for the future expansion of the database which includes high number of eukaryotic species for relative comparison. In addition, there are plans for expansion with inclusion of protein secondary and tertiary structure and pathway information on kinases. Protein structure information is vital in understanding protein function and evolutionary relationships, and pathway information will help to understand the various metabolic and signaling pathways in which the kinases were involved.

Table 1: Comparison of HomoKinase with KinBase and KinWeb.

| Category | Database | | |
|---|---|---|---|
| | HomoKinase | KinBase | KinWeb |
| **Number of protein kinases** | **498** | **506** | **508** |
| HGNC not approved | — | 9 | 10 |
| Gene ID replaced with another Gene ID | — | 1 | 4 |
| Absence of information in Entrez Gene | — | 1 | 1 |
| Pseudogene | — | 1 | 1 |
| Absent of ((i) ATP binding, (ii) kinase activity, and (iii) protein phosphorylation) | — | 12 | 21 |
| Incomplete information—wrong gene ID, wrong gene name | — | 31 | 8 |

## Availability

The database is hosted and available online at http://www.biomining-bu.in/homokinase/.

## Acknowledgments

This work is supported by Grant from the Department of Information Technology (DIT), Government of India (no. DIT/R&D/BIO/15(22)/2008). Suresh Subramani and Raja Kalpana acknowledge the support received from the grant.

## References

[1] G. Manning, G. D. Plowman, T. Hunter, and S. Sudarsanam, "Evolution of protein kinase signaling from yeast to man," *Trends in Biochemical Sciences*, vol. 27, no. 10, pp. 514–520, 2002.

[2] G. Manning, D. B. Whyte, R. Martinez, T. Hunter, and S. Sudarsanam, "The protein kinase complement of the human genome," *Science*, vol. 298, no. 5600, pp. 1912–1934, 2002.

[3] L. N. Johnson, M. E. M. Noble, and D. J. Owen, "Active and inactive protein kinases: structural basis for regulation," *Cell*, vol. 85, no. 2, pp. 149–158, 1996.

[4] C. Y. Yang, C. H. Chang, Y. L. Yu et al., "PhosphoPOINT: a comprehensive human kinase interactome and phospho-protein database," *Bioinformatics*, vol. 24, no. 16, pp. i14–i20, 2008.

[5] A. Krupa, K. R. Abhinandan, and N. Srinivasan, "KinG: a database of protein kinases in genomes," *Nucleic Acids Research*, vol. 32, pp. D513–D515, 2004.

[6] L. Milanesi, M. Petrillo, L. Sepe et al., "Systematic analysis of human kinase genes: a large number of genes and alternative splicing events result in functional and structural diversity," *BMC Bioinformatics*, vol. 6, no. 4, article S20, 2005.

[7] R. L. Seal, S. M. Gordon, M. J. Lush, M. W. Wright, and E. A. Bruford, "Genenames.org: the HGNC resources in 2011," *Nucleic Acids Research*, vol. 39, no. 1, pp. D514–D519, 2011.

[8] D. Binns, E. Dimmer, R. Huntley, D. Barrell, C. O'Donovan, and R. Apweiler, "QuickGO: a web-based tool for Gene Ontology searching," *Bioinformatics*, vol. 25, no. 22, pp. 3045–3046, 2009.

[9] M. Ashburner, C. A. Ball, J. A. Blake et al., "Gene ontology: tool for the unification of biology. The Gene Ontology Consortium," *Nature Genetics*, vol. 25, no. 1, pp. 25–29, 2000.

[10] S. S. Taylor, E. Radzio-Andzelm, Madhusudan, X. Cheng, L. Ten Eyck, and N. Narayana, "Catalytic subunit of cyclic AMP-dependent protein kinasestructure and dynamics of the active site cleft," *Pharmacology and Therapeutics*, vol. 82, no. 2-3, pp. 133–141, 1999.

[11] QuickGO, 2013, http://www.ebi.ac.uk/QuickGO.

[12] AmiGO, 2013, http://amigo.geneontology.org/.

[13] PhophoSitePlus, 2013, http://www.phosphosite.org/.

[14] UniProt, 2013, http://www.uniprot.org/.