Nucleic Acids Research, 2004, Vol. 32, Database issue ***D153–D155***
DOI: 10.1093/nar/gkh019

KinG: a database of protein kinases in genomes

A. Krupa, K. R. Abhinandan and N. Srinivasan*

Molecular Biophysics Unit, Indian Institute of Science, Bangalore 560012, India

Received August 11, 2003; Revised September 3, 2003; Accepted September 11, 2003

ABSTRACT

The KinG database is a comprehensive collection of serine/threonine/tyrosine-specific kinases and their homologues identified in various completed genomes using sequence and profile search methods. The database hosted at http://hodgkin.mbu.iisc.ernet.in/~king provides the amino acid sequences, functional domain assignments and classification of gene products containing protein kinase domains. A search tool enabling the retrieval of protein kinases with specified subfamily and domain combinations is one of the key features of the resource. Identification of a kinase catalytic domain in the user's query sequence is possible using another search tool. The occurrence and location of critical catalytic residues if the query has a catalytic kinase domain, recognition of non-kinase domains in the sequence and subfamily classification of the kinase in the query will help in deciphering the biological role of the kinase. This online compilation can also be used to compare the protein kinases of a given subfamily and domain combinations across various genomes. Another exclusive feature of the database is the collection of the Ser/Thr/Tyr protein kinases and similar sequences encoded in the genomes of archaea and bacteria.

INTRODUCTION

Protein kinases comprise one of the largest families of soluble proteins in eukaryotic genomes. They catalyse the phosphorylation of several cellular proteins on Ser/Thr/Tyr residues altering their functional properties. They are hence central to cellular signalling networks that co-ordinate various activities like metabolism, stress response, transcription, translation, DNA replication and cell cycle control, development of organs, neuronal signalling and apoptosis (1–4). Improper functioning of these enzymes is often manifested in various human diseases and has been implicated in several malignancies (5,6).

Association of the kinase domain with various proteins or non-kinase domains within the same gene product tightly regulates the activity of protein kinases. Hence the non-catalytic domains of protein kinases are critical to our understanding of their biological roles.

The assessment of the diversity and distribution of this important class of enzymes is possible with the availability of large numbers of completely sequenced genomes. This database therefore aims to serve as an online resource for protein kinases identified in the completed genomes. A detailed listing of the domain combinations of protein kinases and their classification is provided for genomes of individual organisms. The database is currently restricted to Ser/Thr/Tyr-specific protein kinases and their homologues and does not include histidine kinases and other classes of protein kinase.

The database therefore adds to the list of existing online resources for protein kinases like Protein Kinase Resource (7: http://pkr.sdsc.edu/html/index.shtml) and http://www.kinase.com. The key features of our database not easily extractable in other databases include the listing by subfamily of protein kinases of an organism with their functional domain assignments. A tool for extraction of protein kinases with specified subfamily and domain combinations is also available. Further unique tools available in KinG include automatic identification of a given query sequence as a kinase (or not) and provision of the location of catalytic residues, associated domains and subfamily classification. This database is therefore expected to be useful to a large community of researchers working on various aspects of the molecular basis of signal transduction by protein kinases.

GENOME-WIDE ASSIGNMENTS

The database currently contains the collection of protein kinases and associated information for five completed eukaryotic genomes including those of *Saccharomyces cerevisiae*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens* and *Arabidopsis thaliana*. In addition protein kinases and similar sequences of eight archaeal genomes and 27 bacterial genomes are provided. The translated ORFs of various genomes have been obtained from publicly available resources including the NCBI (http://ncbi.nlm.nih.gov/), Ensembl (http://www.ensembl.org/), MIPS (ftpmips.gsf.de/cress/arabiprot/), WORMPEP (http://www.sanger.ac.uk/projects/C_elegans/wormpep) and FLYBASE (http://flybase.bio.indiana.edu/). The protein kinases in each of these genomes have been identified using various sensitive sequence and profile search methods like PSI-BLAST (8), HMMER-2 (9), and reverse position specific BLAST (RPS-BLAST) (8) with e-value cut-offs of 0.0001, 0.1 and 0.0001, respectively. The strategies followed to identify a kinase, its functional nature or otherwise and the subfamily assignment are the same as described in our paper on kinases encoded in the human genome (10). Hits identified as kinases

*To whom correspondence should be addressed. Tel: +91 80 293 2837; Fax: +91 80 360 0535; Email: ns@mbu.iisc.ernet.in

have been manually checked to ensure that there are no false positives, as significant e-values can be obtained when the length of the alignment is short. The protein kinase containing gene products encoded in each of the eukaryotic genomes have been classified into subfamilies based on sequence similarity as proposed by Hanks and Hunter (1,11,12). Using HMMER-2 Pfam (13) functional domains have been assigned to these gene products. For every organism, the KinG database provides links to the complete sequence, subfamily and domain composition of each gene product.

Protein kinases and similar sequences encoded in bacterial genomes include lipopolysaccharide kinases (14) and other kinases with distinct domain composition compared with their eukaryotic homologues. This genome-wide distribution information can therefore be used to get an insight into the representation of each kind of protein kinase in various genomes.

SEARCH TOOLS

Domain composition and subfamily-based search

This interface enables the user to search within KinG for protein kinases with a specified combination of domains and/or associated with a particular subfamily. The user can choose the domains from the Pfam (13) functional domains assigned to the protein kinases in the database. A subfamily could be selected from the list of various protein kinases derived from the Protein Kinase Resource (7: http://pkr.sdsc.edu/html/index.shtml). Furthermore, the user can download the individual amino acid sequences of the various protein kinases that meet the user search criteria. Links to domains and subfamily associated with each gene product returned as a result of the search, provide the broad functionality of the domain and subfamily. The putative biological roles of these various protein kinases and the likely signalling pathway in which a given kinase may be involved in can be inferred from the information retrieved on the kinase subfamily and associated domains.

Identification of protein kinase and associated domains

A second component of the database could be used to identify a catalytic kinase domain in the users' query sequence. RPS-BLAST is used to scan the query sequence across a library of various kinase subfamily profiles created using PSI-BLAST. The search returns results indicating the presence or absence of the kinase catalytic domain in the query sequence. The most critical features of the functional kinase catalytic domain are the glycine-rich loop involved in binding to ATP and the catalytic base (an aspartate residue) required for the phosphotransfer reaction. Details regarding the location, extent of conservation of these key functional residues and the associated subfamily are specified in the results if the sequence shares similarity with any known kinase catalytic domain. The query sequence containing the catalytic base is further categorized as an 'RD' or 'non-RD' kinase, based on the presence or absence of arginine preceding the catalytic base (aspartate). The requirement for phosphorylation in the activation segment of most of the 'RD' kinases for regulation has been suggested previously by Johnson *et al.* (15). The occurrence of other functional domains for a kinase-catalytic-domain-containing query could be investigated by a RPS-BLAST-based search, in a library of protein domains derived from Pfam (13) or SMART (16). The multiple sequence alignments of protein domains compiled in the SMART database has been obtained from the Conserved Domain Database (CDD) (17) for the generation of profiles searchable by RPS-BLAST. The user can further check for the functional domains present in the query sequence to get more complete information about the roles of the given kinase. These various features of the query sequence as suggested from the results of the search could be used to understand the gross biological role of the kinase.

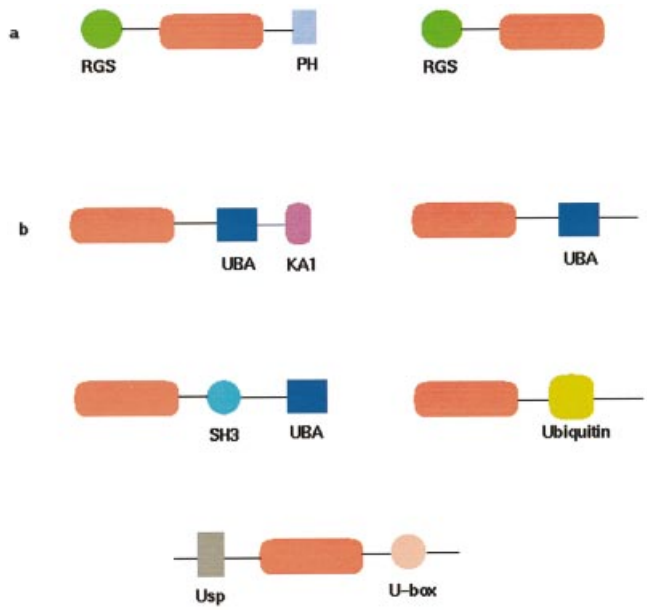

Figure 1. Domain diversity of representative protein kinases belonging to various subfamilies of protein kinases extracted from the KinG database. Protein kinase domains are shown in red. (**a**) Domain combination of protein kinases with the RGS domains characteristic of the AGC_4 subfamily of protein kinases, involved in desensitization of G-protein-coupled receptors. PH, pleckstrin homology domain. Protein kinase representatives from this family are found in *H.sapiens*, *D.melanogaster* and *C.elegans*. Homologues of this subfamily of kinases are not found in *S.cerevisiae*. (**b**) Protein kinases with domains known to be involved in ubiquitination. UBA, ubiquitin-associated domain; Usp, universal stress protein domain; U-box, modified ring finger domain found in proteins involved in ubiquitination. UBA-domain-containing protein kinases are found in the CAMK, AGC and PTK subfamilies of *A.thaliana*, *H.sapiens* and *D.melanogaster*. Protein kinases containing ubiquitin domains encoded in *H.sapiens* and *D.melanogaster* belong to the CAMK subfamily. *Arabidopsis thaliana* in addition encodes another class of ubiquitin-related protein kinases with a U-box, which belong to the plant receptor kinase subfamily.

IMPLICATIONS OF DOMAIN COMPOSITION AND SUBFAMILY ASSIGNMENT TO THE KINASE DOMAIN

The influence of associated functional domains on the activity of protein kinases has been revealed by earlier studies (18,19). Domain compositions are often conserved in a given subgroup of protein kinases indicating that they share common function and modes of regulation. The absence of a particular

functional domain or a given subgroup of kinases in any organism would therefore suggest the lack of the functionality associated with the domain. Such inferences could be drawn by comparison of the domain combination of protein kinases across various genomes. For example, the regulator of G-protein signalling (RGS), a GTPase activating protein domain, is associated with G-protein-coupled receptor kinases (GRKs) (Fig. 1a). In addition to phosphorylation of the G-protein-coupled receptors by GRKs, leading to their desensitization, the RGS domains of GRKs help in restoring the heterotrimeric state of G-proteins, to check downstream signalling events. However, none of the protein kinases identified in *S.cerevisiae* has an RGS domain, suggesting the absence of GRK-mediated deactivation of G-proteins in this organism. Similarly a search for protein kinases with ubiquitin or other ubiquitin-associated domains like UBA and U-box reveal a set of protein kinases (Fig. 1b) of distinct subfamilies in various genomes, suggesting the influence of ubiquitination in signalling pathways mediated by these protein kinases.

The KinG database therefore provides a platform to investigate the functional roles of protein kinases in completed genomes. It would also help the user in understanding the biological role of a protein kinase sequence of specific interest in great detail. The information on protein kinases in the database will be updated with the availability of complete genome information for other model organisms and also with the increase in the knowledge of the protein families derived from databases such as Pfam and SMART.

ACKNOWLEDGEMENTS

The authors thank S. Abhiman and Rana Bhadra for their help in the development of the database. A.K. is supported by a fellowship from the Council of Scientific and Industrial Research, India. K.R.A is supported by the Wellcome Trust, UK. This research is supported by the award of an International Senior Fellowship in biomedical sciences to N.S. by the Wellcome Trust, UK and by the computational genomics project funded by the Department of Biotechnology, India.

REFERENCES

1. Hanks,S.K., Quinn,A.M. and Hunter,T. (1988) The protein kinase family: conserved features and deduced phylogeny of the catalytic domains. *Science*, **241**, 42–52.
2. Pawson,T. (1994) Introduction: protein kinases. *FASEB J.*, **8**, 1112–1113.
3. Zhang,Z., Yu,X., Zhang. Y., Geronimo.,B., Lovlie,A., Fromm,S.H. and Chen,Y. (2000) Targeted misexpression of constitutively active BMP receptor-IB causes bifurcation, duplication and posterior transformation of digit in mouse limb. *Dev. Biol.*, **220**, 154–167.
4. Frost,D.O. (2001) BDNF/trkB signaling in the developmental sculpting of visual connections. *Prog. Brain Res.*, **134**, 35–49.
5. Lee,M.H. and Yang,H.Y. (2001) Negative regulators of cyclin-dependent kinases and their roles in cancers. *Cell. Mol. Life Sci.*, **58**, 907–922.
6. Irby,R.B., Mao,W., Coppola,D., Kang,J., Loubeau,J.M., Trudeau,W., Karl,R., Fujita,D.J., Jove,R. and Yeatman,T.J. (1999) Activating SRC mutation in a subset of advanced human colon cancers. *Nature Genet.*, **21**, 187–190.
7. Smith,C.M., Shindyalov,I.N., Veretnik,S., Gribskov,M., Taylor,S.S., Ten Eyck,L.F. and Bourne,P.E. (1997) The protein kinase resource. *Trends Biochem. Sci.*, **22**, 444–446.
8. Altschul,S.F., Madden,T.L., Schaffer,A.A., Zhang,J., Zhang,Z., Miller,W. and Lipman,D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.*, **25**, 3389–3402.
9. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
10. Krupa,A. and Srinivasan,N. (2002) The repertoire of protein kinases encoded in the draft version of the human genome: atypical variations and uncommon domain combinations. *Genome Biol.*, **3**, RESEARCH0066.1–14.
11. Hunter,T. (1987) A thousand and one protein kinases. *Cell*, **50**, 823–829.
12. Hanks,S.K. and Quinn,A.M. (1991) Protein kinase catalytic domain sequence database: identification of conserved features of primary structure and classification of family members. *Methods Enzymol.*, **200**, 38–42.
13. Bateman,A., Birney,E., Cerruti,L., Durbin R., Etwiller L., Eddy,S.R., Griffiths-Jones,S., Howe K.L., Marshall,M. and Sonnhammer,E.L. (2002) The Pfam protein families database. *Nucleic Acids Res.*, **30**, 276–280.
14. Krupa,A. and Srinivasan,N. (2002) Lipopolysaccharide phosphorylating enzymes encoded in the genomes of Gram-negative bacteria are related to the eukaryotic protein kinases. *Protein Sci.*, **11**, 1580–1584.
15. Johnson,L.N., Noble,M.E. and Owen,D.J. (1996) Active and inactive protein kinases: Structural basis for regulation. *Cell*, **85**, 149–158.
16. Letunic,I., Goodstadt,L., Dickens,N.J., Doerks,T., Schultz,J., Mott,R., Ciccarelli,F., Copley,R.R., Ponting,C.P. and Bork,P. (2002) Recent improvements to the SMART domain-based sequence annotation resource. *Nucleic Acids Res.*, **30**, 242–244.
17. Marchler-Bauer,A., Anderson,J.B., DeWeese-Scott,C., Fedorova,N.D., Geer,L.Y., He,S., Hurwitz,D.I., Jackson,J.D., Jacobs,A.R., Lanczycki,C.J. *et al.* (2003) CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res.*, **31**, 383–387.
18. Lodowski,D.T., Pitcher,J.A., Capel,W.D., Lefkowitz,R.J. and Tesmer,J.J. (2003) Keeping G proteins at bay: a complex between G protein-coupled receptor kinase 2 and Gβγ. *Science*, **300**, 1256–1262.
19. Nagar,B., Hantschel,O., Young,M.A., Scheffzek,K., Veach,D., Bornmann,W., Clarkson,B., Superti-Furga,G. and Kuriyan,J. (2003) Structural basis for the autoinhibition of c-Abl tyrosine kinase. *Cell*, **112**, 859–871.