A kinase sequence database: sequence alignments and family assignment

Oleksandr Buzko and Kevan M. Shokat*

Department of Cellular and Molecular Pharmacology, University of California, San Francisco, 513 Parnassus, San Francisco, CA 94143-0450, USA

ABSTRACT

Summary: The Kinase Sequence Database (KSD) located at http://kinase.ucsf.edu/ksd contains information on 290 protein kinase families derived by profile-based clustering of the non-redundant list of sequences obtained from a GenBank-wide search. Included in the database are a total of 5041 protein kinases from over 100 organisms. Clustering into families is based on the extent of homology within the kinase catalytic domain (250–300 residues in length). Alignments of the families are viewed by interactive Excel-based sequence spreadsheets. In addition, KSD features evolutionary trees derived for each family and detailed information on each sequence as well as links to the corresponding GenBank entries. Sequence manipulation tools, such as evolutionary tree generation, novel sequence assignment, and statistical analysis, are also provided.

Availability: The kinase sequence database is a web-based service accessible at http://kinase.ucsf.edu/ksd

Contact: buzko@cmp.ucsf.edu; shokat@cmp.ucsf.edu/ksd

Recent large-scale sequencing efforts have greatly increased the availability of protein sequence information. One of the largest groups of proteins to emerge from these data is the protein kinase superfamily. Kinases are key regulators of almost every eukaryotic signal transduction cascade. Despite the wealth of information about individual kinase functions, studies of trends across the entire superfamily have suffered from a lack of sources of conveniently grouped information. One of the ways to organize kinase sequences is to determine the domain structure (SH2, PDZ, WW, etc.) of a protein kinase and make predictions as to its function. We have concentrated only on the catalytic domain and discriminated between protein kinases by sequence homology. One of our primary goals was to provide not only a source of information about a number of individual proteins, but also to enable researchers to analyze families and larger groups of kinases as a whole.

One of the first attempts to organize protein kinases in this manner has resulted in grouping of approximately 450 protein kinases from various organisms into 55 families (Hanks and Quinn, 1991) presented in the protein kinase resource (PKR; http://www.sdsc.edu/kinases). However, the exceptionally fast pace of sequencing projects in recent years has led to a rapid increase of the number of kinase sequences in public databases, which are largely unstructured and unclustered.

In order to provide an easily accessible resource for kinase researchers, we have compiled a non-redundant database containing all publicly available protein kinase sequences presented in GenBank and SwissProt databases (as of March 2001). As a starting point, we used a set of 50 diverse protein kinases, generated a profile matrix of their alignment (HMMER 2.1.1 Eddy, 1998) and scanned the online databases. The hits were selected based on sequence homology in the catalytic domain with the cutoff set at 0.01. This search produced over 9500 hits, many of them being duplicate entries. We removed redundancy with a BLAST-based tool developed in-house. It uses long word lengths to detect identical regions and discards entries that come from the same organism and have sequence identity greater than 99%. Processing resulted in a set of 5040 kinase sequences.

Clustering of the dataset into families was accomplished using a combination of BLAST (Altschul *et al.*, 1990) and HMMER 2.1.1 searches. Protein kinases are highly homologous within the catalytic domain, therefore, a sensitive method of separating them into clusters was required. As the first step, an each-against-each BLAST search is run and pairwise scores are produced. The highest-scoring pair is selected and used to generate a profile, with which the rest of the sequence pool is searched. The resulting putative family is separated from the rest of the entries, and the process is repeated beginning with the next best-scoring pair. Such recursive clustering is run until either the lower homology score threshold is reached or the pool of sequences has been exhausted. In order to increase precision, a profile is generated for each of the putative families, and the entire pool is scanned

---

*To whom correspondence should be addressed.

again with the final assignment. Families were separated on the basis of the cutoff score produced by HMMER 2.1.1 set to 50 (which approximately corresponds to an E-value of 10^{-15}). (For a more detailed description of the clustering algorithm please refer to the KSD website at http://kinase.ucsf.edu/ksd/data.html).

Clustering produced a total of 290 groups. Direct comparison with the data contained in Protein Kinase Resource showed nearly identical composition of those families present in PKR, thus, rendering Kinase Sequence Database (KSD) a superset of the former. Naming of the kinase families follows the format in Hanks *et al.* New families were named based on their names assigned in the literature. We used CLUSTALW 1.81 (Higgins *et al.*, 1997) to produce alignments of each family, as well as the master alignment of the entire dataset. The latter was also manually edited for alignment errors and is now used for statistical analysis of the sequences. Comparison of the available structural data for Src, CDK anMAP kinases with the alignments have indicated exceptionally close mapping in the catalytic domain.

KSD is implemented as a relational database residing on a MySQL server. Interaction with the database is accomplished via a set of CGI scripts and Java servlets operating through a Web browser, thus, eliminating the need for standalone client applications and removing any platform dependency issues.

GenBank IDs are used as the primary keys of the sequences in the database. KSD can be queried for a name or GenBank ID of a sequence of interest, or, alternatively, for a description with optional filtering of the results by organism or by organism type (animal, plant, fungus, virus, bacteria or archaea). Since many sequences found in GenBank have multiple ID numbers, we included this information in brief sequence description fields, along with the URL links to the corresponding GenBank entries.

Each family is presented as an alignment in a continuous string format (one sequence per line) and can be viewed in HTML or in Microsoft Excel format available in PC and Macintosh versions. The PC version features interactive residue numbering, which allows the user to determine the position of a residue of interest in the corresponding complete sequence. Both versions highlight the residues that directly contact ATP in each kinase. A brief description of each sequence and the corresponding organism are given in the spreadsheet alignments.

We have also provided real-time assignment of novel user-supplied sequences to kinase families contained in the database by virtue of a profile search against the database. The results of the search are given in the form of a ranked list of the most homologous families. In addition, we have included the capability to assess distribution of amino acid side chains at any position in the kinase domain. The ability to assess residue distribution is extremely important when using binding site mutations to single out a kinase of interest (Bishop *et al.*, 1999). Thus, potential interference by the wild type kinases can be evaluated. Other statistical tools include: search for protein kinases whose sequences match a user-defined pattern, display of relative conservation of amino acid side chains and others.

The planned developments include: (a) mapping of protein kinases sequences onto available three-dimensional structures; (b) capability of viewing alignments, which include only sequences from the user-specified organisms; (c) additional statistical analysis of protein kinase sequences, such as tools for evaluation of variability of residues at positions of interest.

We believe that KSD will provide the signal transduction community with a tool for convenient searching and comparing sequences, studying amino acid patterns and evolutionary relationships, and making structural and functional predictions.

Kinase Sequence Database can be freely accessed at http://kinase.ucsf.edu/ksd.

ACKNOWLEDGEMENTS

REFERENCES

Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.

Bishop,A.C., Kung,C.-Y., Shah,K., Witucki,L., Shokat,K.M. and Liu,Y. (1999) Generation of monospecific nanomolar tyrosine kinase inhibitors via a chemicalgenetic approach. *J. Am. Chem. Soc.*, **121**, 627–631.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Hanks,S. 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**, 38–62.

Higgins,D.G., Thompson,J.D. and Gibson,T.J. (1997) Using CLUSTAL for multiple sequence alignments. *Methods Enzymol.*, **266**, 383–402.