Published online 12 November 2011 *Nucleic Acids Research, 2012, Vol. 40, Database issue* ***D321–D324***
doi:10.1093/nar/gkr915

BYKdb: the Bacterial protein tYrosine Kinase database

Fanny Jadeau[1], Christophe Grangeasse[1,*], Lei Shi[2], Ivan Mijakovic[2], Gilbert Deléage[1] and Christophe Combet[1,*]

[1]Unité Bases Moléculaires et Structurales des Systèmes Infectieux; UMR 5086 CNRS - Université Claude Bernard Lyon 1; IBCP FR 3302 - 7, passage du Vercors, 69367 Lyon CEDEX 07, France and [2]Micalis, AgroParisTech/INRA, 78352 Jouy-en-Josas CEDEX, France

Received August 15, 2011; Revised October 5, 2011; Accepted October 8, 2011

ABSTRACT

Bacterial tyrosine-kinases share no resemblance with their eukaryotic counterparts and they have been unified in a new protein family named BY-kinases. These enzymes have been shown to control several biological functions in the bacterial cells. In recent years biochemical studies, sequence analyses and structure resolutions allowed the deciphering of a common signature. However, BY-kinase sequence annotations in primary databases remain incomplete. This prompted us to develop a specialized database of computer-annotated BY-kinase sequences: the Bacterial protein tyrosine-kinase database (BYKdb). BY-kinase sequences are first identified, thanks to a workflow developed in a previous work. A second workflow annotates the UniProtKB entries in order to provide the BYKdb entries. The database can be accessed through a web interface that allows static and dynamic queries and offers integrated sequence analysis tools. BYKdb can be found at http://bykdb.ibcp.fr.

INTRODUCTION

Reversible protein phosphorylation is a major mechanism in the regulation of fundamental signaling events in all living organisms. In bacteria, four types of phosphorylation systems were described: (i) the two-component system (1), (ii) the phosphoenolpyruvate (PEP) transferase system (PTS) (2), (iii) the eukaryotic-like system (3) and (iv) the bacterial tyrosine kinase (BY-kinase) system (4). Those four types differ in the nature of the phosphorylated amino acids (S, T, Y, H and D) and the phosphate donor (adenosine tri-phosphate (ATP) or PEP). The most recently discovered system involves the BY-kinase family.

BY-kinases comprise two domains: a two-pass transmembrane activator domain (TAD) that includes a large extra cellular part, and an intracellular catalytic domain (CD). The BY-kinases CD encompasses three Walker-like motifs (5), called A, A′ and B, and a C-terminal tyrosine-rich region named Y cluster (YC).

These two domains can either be linked in a single polypeptide (in proteobacteria and actinobacteria), or split into two different proteins, encoded by two adjacent genes (in firmicutes). In this case, it has been shown that BY-kinases activity was effective only if the CD was interacting with the region following the second transmembrane segment of TAD.

A number of studies have demonstrated the importance of BY-kinases in several facets of the physiology of the bacterial cell. More precisely, their best characterized function concerns the regulation of the export and the biosynthesis of bacterial extracellular polysaccharide that are recognized important virulence factors. Since these enzymes are not homologs of their eukaryotic counterparts, BY-kinases are particularly interesting in the search of new therapeutic targets to combat bacterial pathogens.

We present here the Bacterial tyrosine kinase database (BYKdb), developed in order to collect, store and manage BY-kinase sequences with standardized annotations. The sequences are identified by using an *in silico* workflow described in a previous work (6). Diverse sequences identified by the workflow have been validated experimentally since then (Dr Mijakovic, personal communication). The sequences are automatically annotated and stored in the database by a second automated workflow. The

*To whom correspondence should be addressed. Tel: +33 4 37 65 29 47; Fax: +33 4 72 72 26 04; Email: christophe.combet@ibcp.fr
Correspondence may also be addressed to Christophe Grangeasse. Tel: +33 4 72 72 26 88; Fax: +33 4 72 72 26 04;
Email: christophe.grangeasse@ibcp.fr

database can be queried via a WWW interface and the query results can be further analyzed with the numerous integrated analysis tools.

The BYKdb is updated on a monthly basis from the UniProt Knowledgebase (UniProtKB) (7), and stored in the PostgreSQL relational database management system (RDBMS). The programs for the sequence identification and annotation, as well as for the querying and the management of the database are implemented in Java and SQL programming languages.

The first step of the database building procedure is the identification of BY-kinase sequences, thanks to the 'FindBYK' process. It relies on the HMMER software package (version 3.0 of March 2010) (8). The BY-kinases TAD is matched by the Pfam profile PF02706 (9), in a region including the first transmembrane segment. Since this profile identifies proteins that are not BY-kinases, and does not identify CD of two-parts BY-kinases, it is not sufficient to unambiguously identify BY-kinases. Thus, a second profile (6) is used to identify

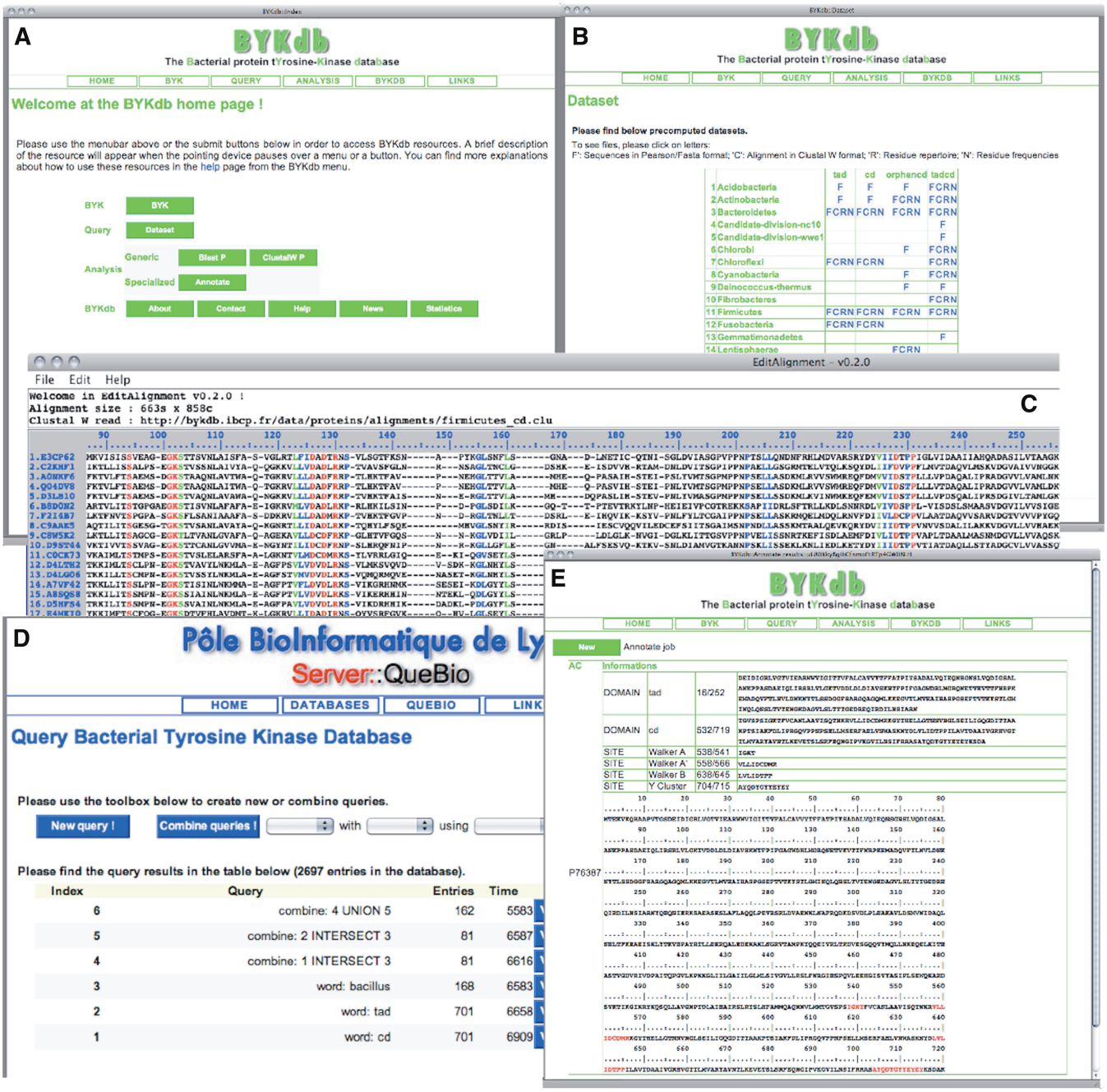

Figure 1. Snapshot of the BYKdb web interface. (**A**) The BYKdb home page, with the resources accessible as clickable buttons (**B**) A pre-computed dataset of BY-kinase and TAD sequences and alignments, ordered by phylum and by cluster. (**C**) The multiple sequence alignment of firmicutes CD viewed in the 'EditAlignment' applet. (**D**) Dynamic queries against the BYKdb. (**E**) The 'Annotate' tool detailed results page for one input sequence.

the CD. The two 'hmmsearch' result sets are analyzed to lead to three clusters of sequences. The first cluster is constituted by sequences matching both HMM profiles ('TADCD' cluster). The second cluster contains sequences with the TAD and CD encoded by two proteins ('TAD' and 'CD' cluster). The sequences are grouped according to their successive accession numbers (AC). The sequences of the third cluster only match the CD profile and have no identified TAD ('orphanCD' cluster). The sequences are further divided into two categories ('confirmed' and 'unsure') according to a filter (isBYK). This filter checks if a sequence contains all the necessary CD motifs (Walker-like motifs and tyrosine cluster) with the right spacing between them (6). The 'unsure' sequences are not integrated in the database.

The 'AnnotateBYK' process ensures the second step of the procedure, i.e. the sequence annotation. The annotation starts with the UniProtKB entries extracted according to the AC of the relevant sequences identified in the first step. The entries are cleared from some data and new data are added. Indeed, the accession numbers, the creation date, the organism and the sequence of each entry are conserved. In the description line (DE), the recommended name field is set to 'Bacterial tyrosine kinase' or to 'Adaptator of bacterial tyrosine kinase' depending on the type of sequences. In the keywords list (KW) are added the 'byk' abbreviation and the cluster of sequences: 'cd', 'tad' or 'tadcd'. The domains and motifs with their start and end positions in the sequence are also added to the feature (FT) lines, respectively as domain and site fields. Besides, for each sequences of the second group, a cross-reference to the related TAD entry is added and reciprocally for the TAD entry. The BY-kinase entries are then loaded in the relational database.

WEB INTERFACE

The BYKdb is accessible through a website (http://bykdb.ibcp.fr) (Figure 1). The user can navigate through the site, thanks to the menubar or the home page buttons. The site is divided into static and dynamic parts.

In the static part, the user can find general information about BY-kinases with links to other external resources ('BYK' link). The user can also access pre-computed datasets ('Dataset' link) sorted according to the phylum and BY-kinase cluster. The datasets include BY-kinase sequences in Pearson/Fasta format and the corresponding multiple sequence alignment. The user can download the corresponding files for further analysis. Furthermore, the alignments can be viewed and edited with the 'EditAlignment' applet developed by our team. In the dynamic part ('Dynamic' link), the user can extract his own datasets by combining multiple criteria (e.g. species and sequence length and sequence cluster). The datasets can be exported as Fasta/Pearson sequences, accession number lists and entry flat files for further analysis with the integrated analysis tools.

The available analysis tools are either generic or specialized. The generic analysis tools (e.g. BLAST or Clustal W) are available through the NPS@ server (10), which is an integrated sequence analysis web server. The specialized analysis tool ('Annotate' link) available on the site allows the annotation of one or several sequences, which can be useful to scan protein sequences from a newly sequenced bacterial genome in order to identify BY-kinases and their adaptators. In the results page, the user can access the start/end positions and the sequences of the different domains and sites. Furthermore, the domains and sites are mapped in color on the BY-kinase sequence.

STATISTICS

BYKdb has been available since July 2011. The release 3.0 (October 2011) comprises 2,746 sequences, including 713 CD sequences with their corresponding TAD sequences, 1,169 TADCD sequences and 151 orphanCD sequences.

CONCLUSION AND PERSPECTIVE

BYKdb collects BY-kinase sequences, identified and annotated, thanks to a computer-automated system. The automatic annotation process used to generate BYKdb entries guarantees harmonized annotations and allows efficient keyword searches. The BYKdb website gives access to pre-computed static datasets and allows dynamic queries. The extracted data can be further analyzed with generic or specialized bioinformatics algorithms. In the future, the annotations will be enriched with new data (e.g. information about protein structure) and new analysis tools (e.g. phylogeny) will be added.

ACKNOWLEDGEMENTS

We acknowledge the UniProt consortium for giving us the permission to use some of the UniProtKB data.

FUNDING

BYKdb is funded by the Agence Nationale de la Recherche (ANR-07-JCJC0125-01 BACTYRKIN) and the Pôle Rhône-Alpes de BioInformatique (PRABI) platform by the Groupement d'Intérêt Scientifique Infrastructures en Biologie, Santé et Agronomie (GIS IBiSA). Funding for open access charge: Agence Nationale de la Recherche (ANR-07-JCJC0125-01 BACTYRKIN).

Conflict of interest statement. None declared.

REFERENCES

1. Bourret,R.B., Hess,J.F., Borkovich,K.A., Pakula,A.A. and Simon,M.I. (1989) Protein phosphorylation in chemotaxis and two-component regulatory systems of bacteria. *J. Biol. Chem.*, **264**, 7085–7088.
2. Reizer,J., Saier,M.H. Jr, Deutscher,J., Grenier,F., Thompson,J. and Hengstenberg,W. (1988) The phosphoenolpyruvate: sugar phosphotransferase system in gram-positive bacteria: properties, mechanism, and regulation. *Crit. Rev. Microbiol.*, **15**, 297–338.

3. Bakal,C.J. and Davies,J.E. (2000) No longer an exclusive club: eukaryotic signalling domains in bacteria. *Trends Cell Biol.*, **10**, 32–38.
4. Grangeasse,C., Cozzone,A.J., Deutscher,J. and Mijakovic,I. (2007) Tyrosine phosphorylation: an emerging regulatory device of bacterial physiology. *Trends Biochem. Sci.*, **32**, 86–94.
5. Walker,J.E., Saraste,M., Runswick,M.J. and Gay,N.J. (1982) Distantly related sequences in the alpha- and beta-subunits of ATP synthase, myosin, kinases and other ATP-requiring enzymes and a common nucleotide binding fold. *EMBO J.*, **1**, 945–951.
6. Jadeau,F., Bechet,E., Cozzone,A.J., Deléage,G., Grangeasse,C. and Combet,C. (2008) Identification of the idiosyncratic bacterial protein-tyrosine kinase (BY-kinase) family signature. *Bioinformatics*, **24**, 2427–2430.
7. The UniProt Consortium. (2011) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res.*, **39**, D214–D219.
8. Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.
9. Finn,R.D., Mistry,J., Schuster-Böckler,B., Griffiths-Jones,S., Hollich,V., Lassmann,T., Moxon,S., Marshall,M., Khanna,A., Durbin,R. *et al.* (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.
10. Combet,C., Blanchet,C., Geourjon,C. and Deléage,G. (2000) NPS@: network protein sequence analysis. *Trends Biochem. Sci.*, **25**, 147–150.