BIOINFORMATICS DISCOVERY NOTE Vol. 24 no. 21 2008, pages 2427–2430
doi:10.1093/bioinformatics/btn462

Sequence analysis

Identification of the idiosyncratic bacterial protein tyrosine kinase (BY-kinase) family signature

Fanny Jadeau, Emmanuelle Bechet, Alain J. Cozzone,
Gilbert Deléage,Christophe Grangeasse and Christophe Combet*

Institut de Biologie et Chimie des Protéines; UMR5086, CNRS, Université Lyon 1, IFR128 BioSciences Lyon-Gerland, 7, passage du Vercors, 69367 Lyon CEDEX 07, France

Received on June 6, 2008; revised on August 1, 2008; accepted on August 25, 2008

Advance Access publication September 3, 2008

Associate Editor: Limsoon Wong

ABSTRACT

Motivation: Most of the protein tyrosine kinases found in bacteria have been recently classified in a new family, termed BY-kinase. Indeed, they share no sequence homology with their eukaryotic counterparts and have no known eukaryotic homologues. They are involved in several biological functions (e.g. capsule biosynthesis, antibiotic resistance, virulence mechanism). Thus, they can be considered interesting therapeutic targets to develop new drugs to treat infectious diseases. However, their identification is rendered difficult due to slow progress in their structural characterization and comes most often from biochemical experiments. Moreover BY-kinase sequences are related to many other bacterial proteins involved in several biological functions (e.g. ParA family proteins). Accordingly, their annotations in generalist databases, sequence analysis and classification remain partial and inhomogeneous and there is no bioinformatics resource dedicated to these proteins.

Results: The combination of similarity search with sequence-profile alignment, pattern matching and sliding window computation to detect the tyrosine cluster was used to identify BY-kinase sequences in UniProt Knowledgebase. Cross-validations with keywords searches, pattern matching with several patterns and checking of motifs conservation in multiple sequence alignments were performed. Our pipeline identified 640 sequences as BY-kinases and allowed the definition of a PROSITE pattern that is the signature of the BY-kinases. The sequences identified by our pipeline as BY-kinases share a good sequence similarity with BY-kinases that have already been biochemically characterized, and they all bear the characteristic motifs of the catalytic domain, including the three Walker-like motifs followed by a tyrosine cluster.

Availability: http://bykdb.ibcp.fr

Contact: c.combet@ibcp.fr

1 INTRODUCTION

In bacteria, protein phosphorylation was unknown until the 1970s. Since then, four types of bacterial phosphorylation systems were described: (i) the two-component system (Bourret *et al*., 1989), (ii) the phosphoenolpyruvate transferase system (PTS) (Reizer *et al*., 1988), (iii) the eukaryotic-like system (Bakal and Davies, 2000) and (iv) the bacterial tyrosine kinase (BY-kinase) system (Grangeasse *et al*., 2007). The phosphorylated amino acids and the phosphate donor differ according to each system. The first two systems represent a hallmark of bacterial signalling. The PTS seems to be restricted to bacteria whereas two-component systems are more widespread and they have been found in a variety of eukaryotic species including plants, fungi and yeasts (Mizuno, 2005). The latter system that involves BY-kinases has also quickly turned out to be quite different from that of eukaryotes and it represents a promising regulatory tool of bacterial physiology.

BY-kinases comprise two domains: a two-pass transmembrane activator domain (TAD) with a large extracellular part and an intracellular catalytic domain (CD). In actinobacteria and proteobacteria, both domains belong to a single protein whereas in firmicutes, the two domains are found in the form of two distinct proteins encoded by two adjacent genes. In addition, it has been shown that BY-kinases of firmicutes were only active when CD was interacting with TAD and more precisely with the region following the second transmembrane segment of TAD. In contrast, BY-kinases from proteobacteria are constitutively active. Another difference between proteobacteria and firmicutes is the length of the external loop, which is longer in proteobacteria.

The TAD is partially matched by the Pfam profile PF02706 (Finn *et al*., 2006). The profile encompasses a region of ∼130 (firmicutes) or ∼240 (proteobacteria) residues comprising the first transmembrane segment of the TAD. Thus, this profile could help to identify BY-kinases but is not sufficient. Indeed, it matches proteins that are not BY-kinases as PCP1 family proteins that are involved in the length regulation of the oligosaccharidic chain of lipolysaccharide. Also, BY-kinases CD of firmicutes cannot be found with this profile. The BY-kinases CD encompasses three Walker-like motifs (called A, A′ and B; Walker *et al*., 1982) and a tyrosine cluster (YC). It seems that the level of phosphorylation of the YC regulates the activity of the BY-kinases.

BY-kinases are interesting new therapeutic targets as tyrosine phosphorylation plays a key role in many biological functions (Ilan *et al*., 1999; Vincent *et al*., 2000), especially the virulence mechanisms of some pathogens, and as they have no eukaryotic homologues.

We introduce below an efficient *in silico* pipeline to identify BY-kinases using a sequence-profile alignment similarity search, pattern matching of the three Walker-like motifs and a sliding window computation to detect the YC.

*To whom correspondence should be addressed.

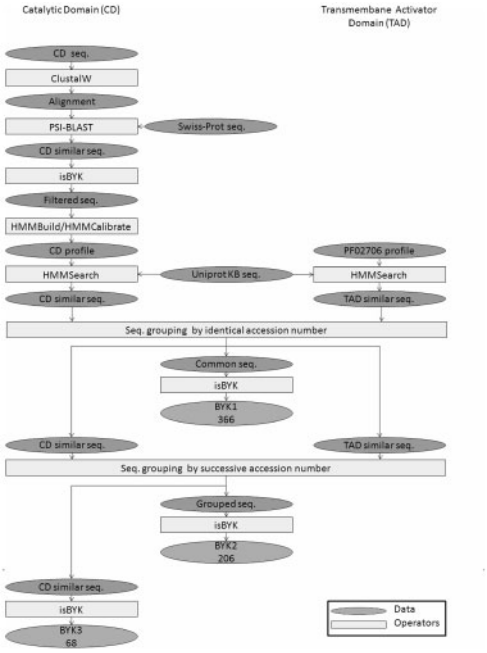

Fig. 1. Flowchart of the BY-kinase sequences identification pipeline. Seq.: sequences. *isBYK* is the filtering algorithm looking for the motifs and the YC cluster. *BYK1*, *BYK2* and *BYK3* are the three groups of putative BY-kinases identified by our pipeline.

Table 1. Motifs and distances (Δ) used for sequence filtering (*isBYK*)

A	Δ1	A′	Δ2	B	Δ3	YC
GK[ST]	7,27	[ILVFM](3)DXDXR	60,80	[ILVFM](3)DX(2)P	≥30	2/7

2 METHODS

A keyword search on Swiss-Prot database (UniProt Consortium, 2008) with protein and organisms names allows the sequence retrieving of the 11 biochemically characterized BY-kinase sequences (UniProtKB accession numbers: P0C0T9, P38134, P58764, P76387, Q04663, Q3K0T0, Q54520, Q8X7L9, Q8XC28, Q9AFI1, Q9AHD2).

The Clustal W (Thompson *et al*., 1994) multiple sequence alignment (version 1.8, slow mode with default parameters) of the CD (residues 531–720 of UniProtKB:P76387) of the 11 BY-kinase sequences was used as an input for a Position Specific Iterated - Basic Local Alignment Search Tool (PSI-BLAST) (version 2.2.13, *E*-value 1e-6, one iteration) (Schäffer *et al*., 2001) search against Swiss-Prot (release 55.3; 29-APR-2008; 366 226 sequences) in order to enrich the initial profile (Fig. 1). The resulting sequences were filtered (*isBYK* filter described below) for the presence of the three motifs and the YC (Table 1). The remaining sequences were aligned and used to compute a Hidden Markov Models (HMM) profile thanks to the HMMER (version 2.3.2) package. This CD profile was used to perform a global profile-local sequence HMM search (different *E*-value tested) against UniProtKB (release 13.3; 29-APR-2008; 6 074 524 sequences). A second HMM search was performed with the Pfam PF02706 profile (E-value ≤ 10).

Results of the two HMM searches were cross-correlated using the accession numbers of sequences and filtered with the *isBYK* method. After this treatment, the sequences are divided in three groups: *BYK1*, *BYK2* and *BYK3*. *BYK1* sequences (366) are found by the two HMM searches and they contain both domains (TAD + CD). *BYK2* sequences have successive accession numbers (successive genes belonging to the same operon), one contains the TAD (found by the Pfam profile) and the other contains the CD (found by our CD profile). The *BYK2* sequences (206) correspond to firmicutes type BY-kinases. The remaining sequences (68), matching only the CD profile, have been classified the *BYK3*. For *BYK3* sequences, the TAD domains are possible novel, having neither accession numbers successive to the CD sequences nor sequence details. A final Clustal W alignment for each group (*BYK1–3*) of sequences and all sequences has been computed from scratch.

The *isBYK* filter starts by searching all the occurrences of each motif (Table 1) on each sequence identified by the CD profile thanks to a regular expression matching. Then, a combinatorial computation is performed with all these occurrences. The resulting combinations of the three motifs are filtered thanks to the distance (Table 1) between motifs. This approach saves CPU time for highly varying distances. The YC is defined by two tyrosine counted in a sliding window of seven residues with a step of one residue and searched within sequences that passed the previous motif and distance filtering. The overlapping windows are merged in a single window. Sequences that passed this *isBYK* filter are considered as BY-kinase sequences.

BY-kinases found by our pipeline were compared with the results of a keyword search (description: '*tyrosin* *kinase*'; Taxon: 'achaebacteria | bacteria') with Sequence Retrieval System (SRS) (Etzold *et al*., 1996) and four pattern searches (PROSITE syntax: pattern1 = G-K-[ST]-X(7,27)-[ILVFM](3)-D-X-D-X-R-X(60,80)-[ILVFM](3)-D-X(2)-P-X(30,150)-Y-X(0,5)-Y, pattern2 = G-K-[ST]-X(7,27)-[ILVFM](3)-D-X-D-X(62,82)-[ILVFM](3)-D-X(33,153)-Y-X(0,5)-Y, pattern3 = G-K-[ST]-X(7,27)-[ILVFM](3)-D-X-D-X(62,82)-[ILVFM](3)-D and pattern4 = X(6)-G-K-[ST]-X(10,30)-D-X-D-X(62,82)-D-X-[ILVFM](3)D-X(52,72)-Y-X(0,5)-Y) with the PattInProt algorithm available on the NPS@ server (http://npsa-pbil.ibcp.fr) (Combet *et al*., 2000).

3 RESULTS

3.1 CD profile building

A PSI-BLAST search against Swiss-Prot with the CD alignment of 11 biologically characterized BY-kinases matched 45 sequences. The *isBYK* filter selected 21 sequences among the 45. According to a keyword search against Swiss-Prot, 22 sequences are annotated as tyrosine kinases. The two approaches share in common 21 sequences. In Swiss-Prot, entry Q1DB00 is annotated as a tyrosine kinase. However, it is not found by a HMM search with the CD profile as the sequence is very different from known BY-kinases with only one [ILVFM](3) motif (residues 135–138) and two YC (residues 170–173 and 203–205). According to a BLAST search against Swiss-Prot, this sequence is similar to that of serine/threonine kinases (e.g. P54738). All these data lead to the elimination of Q1DB00 from our shortlist. The alignment of the 21 sequences (Fig. 2) was used for the following HMM search.

3.2 Pipeline results

The HMM search with the PF02706 profile found 1123 sequences. The total number of sequences and the number of sequences filtered as BY-kinases found by the HMM search with the CD profile depending on the selected *E*-value are shown in Table 2 for each group (*BYK1*, *BYK2* and *BYK3*).

According to the results reported in Table 2, the CD profile alone cannot correctly identify and distinguish at the same time BY-kinases. For comparison with SRS and pattern search the results

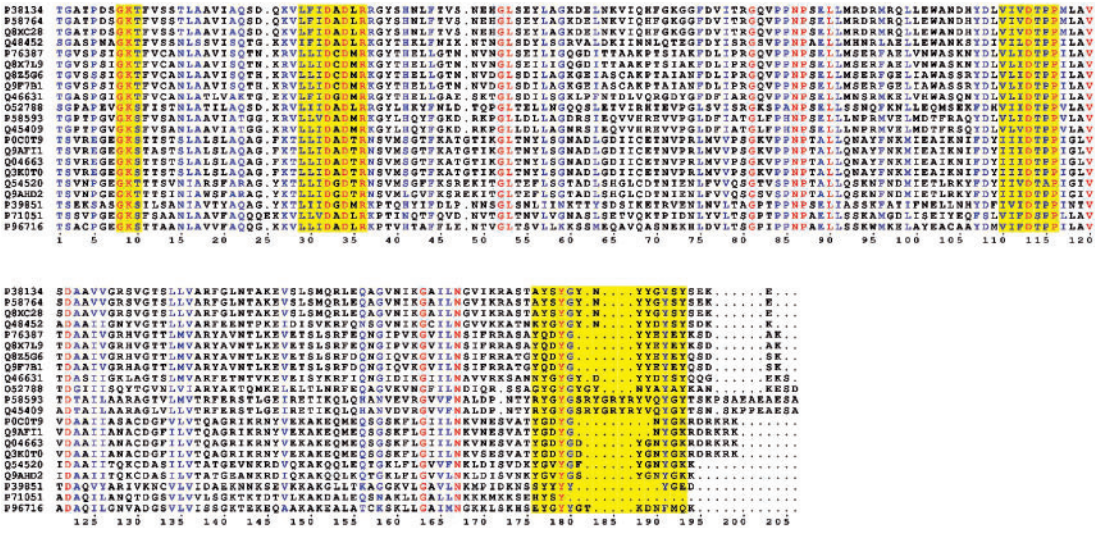

Fig. 2. BY-kinase CD alignment used for the HMM search. Similar residues are in blue and identical in red. The three Walker-like motifs and the YC cluster are highlighted in yellow: motif A (columns 8–10), motif A′ (columns 29–36), motif B (columns 110–116) and YC (columns 176–193). Figure prepared with ESPript (Gouet *et al*., 1999).

Table 2. HMM search with the CD profile, filtering and grouping results

E-value	1e-100	1e-50	1e-30	1e-12	1e-6	10
BYK1	85	241	312	349	365	366
BYK2	38	190	198	203	204	206
BYK3	13	49	59	62	67	68
BYK	136	480	569	614	636	640
All	139	549	727	874	993	4116
BYK/All	0.98	0.87	0.78	0.70	0.64	0.16

of the run with *E*-value threshold of 10 were used. The remaining sequences identified by the HMM CD profile mainly belong to a few number of protein families: ParA, cobyrinic acid ac-diamine synthase, Soj, MRP or MinD.

For the three *BYK* groups, a Clustal W alignment was computed and shows the conservation of the three Walker-like motifs, as well as the alignment of the 640 sequences.

3.3 Comparison with keyword search

The keyword search on UniProtKB with SRS retrieved 422 sequences. Among these sequences, 244 are found in common with the ones identified by our pipeline as BY-kinases. The CD profile matched 115 more sequences in common with the SRS set. These sequences are excluded by the *isBYK* filter because they do not match all the criteria defined by the motifs and distances between them. Thus, 73/115 have amino acid changes (non-synonymous mutation due to one base substitution in the codon) in at least one motif, two do not have the YC, four have bad distance between the three motifs and/or YC and 33 for several of the previous reasons. The last 3/115 are ParA family proteins (A0Y4Q0, Q3IK41 and A4B860).

The remaining 63/422 sequences specifically found in SRS were checked by reading the entry text and by running a BLAST against Swiss-Prot. Thirty-nine sequences are similar to serine/threonine kinases. They contain in their description field both tyrosine kinase and serine and/or threonine kinase (e.g. Q47KC4). Seven proteins are aminoglycoside phospotransferases (e.g. A3SDE7). A5ELQ8 is similar to the Acyl-CoA dehydrogenase family member 10 (Q6JQN1). The Q7UV17 sequence (74 residues) has similarity in ankyrin repeat regions of several sequences (e.g. Q9D738). A9I9E4 has similarity with IPT/TIG domain of hepatocyte growth factor receptor precursor (e.g. Q769I5). Q8A2E1 has similarity with the cell division protein kinase five homolog. A6M3Q6 is a protein phosphatase (e.g. P08538). B1IT65 is a S-adenosylmethionine synthetase (e.g. Q3YXS9). Q8EXB4, described as a receptor tyrosine kinase, has no significant similarity and remains of unknown function. Five sequences are TAD of BY-kinases (e.g. Q65E44). The five remaining sequences could be BY-kinase. A4LFX9 is a sequence of 42 residues. Q3EM42, Q4MT00, Q6HKX2 and Q81FJ7 are divergent BY-kinases with insertions–deletions and/or mutations of at least one motif. For these divergent/truncated sequences, a new sequencing and analysis of the genome region is needed.

3.4 Comparison with pattern searches

The pattern search against UniProtKB with the pattern1 (the BY-kinase signature identified in this work) matched 640 sequences. They are the same sequences that our pipeline identified as BY-kinases. The pattern2 (pattern1 without R residue in A′ motif and P residue in B motif) found 1026 sequences. Among the 386 other proteins, there are mainly ParA family proteins, cobyrinic acid ac-diamine synthase proteins, sporulation initiation proteins (Soj) or MinD proteins. The pattern3 (pattern2 without the YC) identified 1879 sequences. The 856 supplementary sequences are mainly of the four families cited above. The pattern4 was defined according to conserved aligned blocks previously identified in the literature (Grangeasse *et al*., 2007). It matched 823 sequences among which 559 are common with the 640 found by our pipeline. The 264 remaining sequences belong to the families described above, the MRP family and the elongation factor 3 of eukaryotes. Altogether,

these results confirmed that the signature described by pattern1 is characteristic of the BY-kinase protein family.

4 CONCLUSION

BY-kinases are a recently discovered family of proteins involved in several biological functions including virulence mechanisms of some pathogens. Their sequence is related to other protein families but is distinct by the presence of three specific Walker-like motifs (one Walker-A-like followed by two Walker-B-like) and a YC. This work allowed the identification of the BY-kinase family signature summarized as a PROSITE pattern: G-K-[ST]-X(7,27)-[ILVFM](3)-D-X-D-X-R-X(60,80)-[ILVFM](3)-D-X(2)-P-X(30,150)-Y-X(0,5)-Y. The HMM profile will be useful to catch sequences with errors that are BY-kinases. The sequences identified by our pipeline will be annotated and collected in a database. This database will be made available through Internet as part of a dedicated bioinformatics resource for BY-kinases.

ACKNOWLEDGEMENTS

The authors acknowledge Pr. P. Gouet for his help with the ESPript software.

Funding: French Agence Nationale de la Recherche (ANR Jeune Chercheur BACTYRKIN, ANR-07-JCJC0125-01).

Conflict of Interest: none declared.

REFERENCES

Bakal,C.J. and Davies,J.E. (2000) No longer an exclusive club: eukaryotic signalling domains in bacteria. *Trends Cell Biol*., **10**, 32–38.

Bourret,R.B. *et al*. (1989) Protein phosphorylation in chemotaxis and two-component regulatory systems of bacteria.*J. Biol. Chem.*, **264**, 7085–7088.

Combet,C. *et al*. (2000) NPS@: network protein sequence analysis. *Trends Biochem. Sci.*, **25**, 147–150.

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

Etzold,T. *et al*. (1996) SRS: information retrieval system for molecular biology data banks. *Methods Enzymol.*, **266**, 114–128.

Finn,R.D. *et al*. (2006) Pfam: clans, web tools and services. *Nucleic Acids Res.*, **34**, D247–D251.

Gouet,P. *et al*. (1999) ESPript: multiple sequence alignments in PostScript. *Bioinformatics*, **15**, 305–308.

Grangeasse,C. *et al*. (2007) Tyrosine phosphorylation: an emerging regulatory device of bacterial physiology. *Trends Biochem. Sci.*, **32**, 86–94.

Ilan,O. *et al*. (1999) Protein tyrosine kinases in bacterial pathogens are associated with virulence and production of exopolysaccharide. *EMBO J.*, **18**, 3241–3248.

Mizuno,T. (2005) Two-component phosphorelay signal transduction systems in plants: from hormone responses to circadian rhythms. *Biosci. Biotechnol. Biochem.*, **69**, 2263–2276.

Reizer,J. *et al*. (1988) The phosphoenolpyruvate: sugar phosphotransferase system in gram-positive bacteria: properties, mechanism, and regulation. *Crit. Rev. Microbiol.*, **15**, 297–338.

Schäffer,A.A. *et al*. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res.*, **29**, 2994–3005.

UniProt Consortium (2008) The universal protein resource (UniProt). *Nucleic Acids Res*., **36**, D190–D195.

Thompson,J.D. *et al*. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.*, **22**, 4673–4680.

Vincent,C. *et al*. (2000) Relationship between exopolysaccharide production and protein-tyrosine phosphorylation in gram-negative bacteria. *J. Mol. Biol.,* **304**, 311–321.

Walker,J.E. *et al*. (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.