Sma3s: a universal tool for easy functional annotation of proteomes and transcriptomes

Carlos S. Casimiro-Soriguer[1], Antonio Muñoz-Mérida[2],

Antonio J. Pérez-Pulido[1]

[1]Centro Andaluz de Biología del Desarrollo (CABD-CSIC), Universidad Pablo de Olavide, Ctra. Utrera, Km. 1, 41013 Sevilla, Spain; [2]CIBIO-InBIO, Research Network in Biodiversity and Evolutionary Biology, Universidade do Porto, Campus Agrário de Vairão, 4485-661 Vairão, Portugal

Abstract

The current cheapening of next-generation sequencing has led to an enormous growth in the number of sequenced genomes and transcriptomes, allowing wet labs to get the sequences from their organisms of study. To make the most of these data, one of the first things that should be done is the functional annotation of the protein-coding genes. But it used to be a slow and tedious step that can involve the characterization of thousands of sequences.

Sma3s is an accurate computational tool for annotating proteins in an unattended way. Now, we have developed a completely new version, which includes functionalities that will be of utility for fundamental and applied science. Currently, the results provide functional categories such as biological processes, which become useful for both characterizing particular sequence datasets and comparing different projects. But one of the most important implemented innovations is that it has now low computational requirements, and the complete annotation of a simple proteome or transcriptome usually takes less than 24 hours in a personal computer.

Finally, it should be noted that Sma3s has been tested with a large amount of complete proteomes, and it has demonstrated its potential in health science and other specific projects.

Keywords: Functional annotation; Bioinformatic tool; Proteome; Transcriptome

Significance of the study

We present a completely new release of our functional annotation tool Sma3s, together with analyses involving whole proteomes. Sma3s is able to annotate a complete proteome or transcriptome in a plausible time by any researcher using a personal computer. It could allow the extraction of new knowledge in many of the current genomics and proteomics projects. We show the use of this computational tool, annotating a lot of bacterial proteomes and showing how it could lead to the discovery of hidden functional information.

1. Introduction

The current omics era is generating an enormous amount of available proteomes that we need to functionally annotate if we want to make the best of them. In fact, the annotation of all the protein-coding genes from an organism accelerates its knowledge at the molecular level [1,2].

Whereas the transcriptome allows both assessing the gene expression in a particular condition and knowing the specific protein-coding genes expressed by an organism, the genome allows knowing all the genes of this organism. It includes genes encoding for proteins predicted by a gene finder utility or proteogenomics strategy [3]. But, both of them need the functional annotation of protein-coding sequences prior to further analysis.

Functional annotation of a dataset coming from a whole genome or transcriptome can become a slow and tedious job, mainly when researchers have no knowledge on bioinformatics. In this context is usual the manual or semi-manual annotation by assigning functional terms extracted from the best hits in a default similarity search [4,5]. Furthermore, the automatic annotation of large datasets, usually based on sequence similarity, is something that cannot be easily managed by the current computational tools. Most of available automatic annotators only allow the use of web applications with limitations in the number of query sequences to analyze. This is the case of the best scored tools arisen from the *critical assessment of protein function annotation experiment* (CAFA), such as FFPred, Argot, PANNZER, ESG/PFP, or BAR-PLUS [6]. Only two of them allow the annotation of large datasets using specific programming scripts (FFPred [7], PANNZER [8]), but their standalone versions have a lot of requirements, including large databases and specialized software, all of which are elusive for experimental researchers. Blast2GO also belongs to this category [9]. It is a widely used annotation tool that can be difficult to use, especially with large datasets of sequences and without a commercial subscription. All of these tools annotate mainly amino acid sequences, but there are other methods for specifically annotating protein-coding sequences, useful for analyzing transcriptomic data, which have the same weaknesses [10–12].

To overcome the challenges of the current functional annotators we developed Sma3s, which is a standalone tool that has already shown a high accuracy with large sequences datasets, both of proteins and protein-coding nucleotide sequences [13]. From its first version it has been used in

projects with organisms coming from heterogeneous taxonomic divisions such as bacteria [14], fungus [15–17], invertebrates [18], plants [19,20], and animals [21–23], and it provided useful results in all cases. Despite being easy to use, Sma3s had some computational requirements that we have now overcome. Currently, it only depends on the installation of the standard BLAST software, and it allows using Sma3s on any operating system while hardly using computational resources. Additionally, it has now a low requirement of time thanks to the use of a shorter non-redundant database providing the reference annotations. As mentioned above, Sma3s is being used to annotate large datasets, but detailed analysis of the annotations is sometimes performed by other software [15,21], which is not able to exploit the full potential of the results. Thus, Sma3s now provides elaborated results, including functional categories which allow the user the easy creation of charts and the further analysis of the obtained annotations. Finally, Sma3s accuracy has been improved, thanks to the integration of its three previously independent modules, and by the use of quality tags from the annotations, such as the evidence codes coming from Gene Ontology [24].

The annotator has been compared with one of the last published algorithm to annotate proteins, Argot, and Sma3s obtains the best results with a benchmark of well-annotated sequences where self-annotation was avoided. In here, we also show new uses for Sma3s, such as the annotation of a great dataset of proteomes and the analysis of results to get new knowledge. All of them, as far as we know, will leave Sma3s as virtually one of the easier and faster, keeping its accuracy, functional annotator of proteins and protein-coding sequences currently available.

2. Materials and methods

2.1. Improvements in the algorithm

The three independent modules of Sma3s have been now integrated to give a more complete annotation. The two first modules, which find the most significant homologs, are initially used to assign a gene name and description to each query sequence. But only informative names are used, avoiding those with rare symbols or longer than 6 characters. Then, annotation terms are assigned from either the module 1 or 2, and this preliminary annotation is complemented with the more productive module 3.

One of the principal algorithm improvements concerns to the results. The annotation sources have been extended, and the final annotation report now includes EC numbers for enzymes, together with UniProt keywords and pathways [25], and GO terms [24]. To add functional categories, which are especially useful to undertake great annotation or comparative genomic projects, Sma3s gives GO Slim terms that report more general annotations. The GO Slim terms have been extracted from each GO term in the reference database, using both the Map2Slim script and the Generic GO Slim file from the Gene Ontology web. The results also include four categories used to classify the keywords in UniProt: Biological process, Cellular component, Developmental stage, and Disease.

Due to the exponential growth of the sequence databases, the quality of the annotations is something to keep in mind [26]. Thus, Sma3s allows reporting quality annotations, where only experimental assigned GO terms and keywords will be used. To do this, annotations with the "Inferred Electronic Annotation" evidence code coming from both GO terms (IEA) and UniProt keywords (ECO:0000501, and codes related to this one) can be discarded in the analysis. In addition, non-informative annotations are avoided, as well as database sequences without any GO term and keyword, and predicted sequences from the reference database can be also discarded.

The remaining parameters from Sma3s are now fixed in an automatic way for improving both proteome and transcriptome annotations, and make easier the use of the annotator.

Finally, Sma3s offers the annotations in a text file that can be opened with any spreadsheet program (tab-separated values; TSV), along with a file containing the summary of the results. This latter includes the functional categories, with biological processes and pathways, which allow the easy creation of figures to the end user.

2.2. Requirements to use Sma3s

The installation of the Blast+ package is the only mandatory requirement for Sma3s (ftp://ftp.ncbi.nlm.nih.gov/blast/executables/LATEST), together with the Perl programming language interpreter. Sma3s takes around 24 hours to annotate a simple proteome (around 5000 sequences), and a similar time to annotate a short transcriptome (adding -nucl parameter, which will use blastx instead of blastp). Moreover, Sma3s allows parallelization of the initial Blast step, using high performance computing (HPC) to accelerate the entire process.

The Sma3s script can be found at:

http://www.bioinfocabd.upo.es/sma3s

There, you can see video tutorials about how to use Sma3s in different operating systems.

Finally, you can use UniProt files with .dat extension from its website to perform the annotation (ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/). But we offer a non-redundant database, coming from UniRef90, which allows shorter times to annotate bigger sequence datasets. This database was used to produce the results presented in this work, and it can be downloaded from our website.

2.3. Test dataset of new proteins

To annotate proteins with a current minimum quality, but using a previous database version, where the proteins were not annotated, we proceeded with the following steps. We used the manually curated Swiss-Prot section in UniProt database release 2016_07. Then, we collected entries created from 2015, with only one Accession number, thus avoiding any new entry coming from a previous one. But these proteins could come from TrEMBL, the not curated section in UniProt database. To avoid the latter, we checked the history from each entry. Thus, we selected entries created in UniProt from January 2015 to July 2016 and currently stored in Swiss-Prot. Finally, we remove 41 proteins that lacked GO terms, since they did not allow measuring the accuracy.

Following this strategy, we found 349 proteins from Swiss-Prot release 2016_07, with manually assigned annotations, which did not exist in the release from 2014 (Suppl. file 1). Thus, annotating these proteins using Swiss-Prot release 2014_11, and checking accuracy with Swiss-Prot release 2016_07, constitutes a good procedure to test annotation tools avoiding self-annotation.

To check the prediction, we used GO Slim terms and UniProt keywords from Sma3s, and convert GO annotations found by Argot to GO Slim, using the Map2Slim script by the Gene Ontology Consortium.

We used Sma3s with the following reference databases: Swiss-Prot 2014_11, UniProt 2014_11, UniRef90 2014_11, and UniRef90 2016_07 (taking away the 349 query sequences). And we used Argot release 2.5 from its website (July, 2016), which used databases from 2014.

2.4. Annotation of *Bacillus* proteomes

To annotate different proteomes from *Bacillus* genus we downloaded the protein dataset files from Ensembl bacteria database [27]. We selected only reference species whose names were composed by two words. Thus, we finally downloaded 52 different proteomes. All of these were used to run Sma3s, using default parameters and a HPC cluster to accelerate the process.

3. Results and discussion

3.1. Sma3 algorithm overview

The Sma3s algorithm has been completely rewritten to simplify its use and offer more useful information in the results, in addition to reduce the entire run time (Fig. 1). Originally, Sma3s was composed of three independent but complementary modules. The first of them searches for very similar sequences and its annotations, the second searches for orthologs, and the third one uses all significant alignments from a Blast search to retrieve shared annotations with significant expected values. In the current version, all the three modules have been merged to both improve the ease to use and complete the results with more accurate assignments.

Since both of the initial modules are based on more similar sequences, gene names and descriptions are mainly assigned from them rather than the third module. But annotations overall are assigned from all the 3 modules to improve the final sensitivity.

Nevertheless, users can run Sma3s with a single module to check if their sequences are already annotated in the database (using only the module 1), or if there are orthologs in the database that enable the annotation (using only the module 2).

3.2. Proving the accuracy of Sma3s for annotating protein sequences

To test the accuracy of Sma3s, a dataset of 349 manually curated proteins was used. These proteins entered in Swiss-Prot in 2015 or later. So, to avoid self-annotation Sma3s was used with Swiss-Prot 2014 as the reference database for the annotation. The result shows that despite of they are new sequences in the public database of proteins, the obtained accuracy is high, with recall around 60% to both keywords and GO terms (Fig. 2).

Sma3s was initially based on Swiss-Prot as the reference database, but we wanted to increase the accuracy using a greater database. Thus, UniProt database was selected to annotate the same dataset. This is a database with a number of sequences 100 times higher (Table 1). In this case, the results show that though the precision increases slightly, the run time makes it unfeasible, especially if we want to annotate huge sequence datasets.

In order to decrease the run time, we wanted to use a shorter database but maintaining a more complete collection of protein sequences than with Swiss-Prot. To do that we used UniRef90, where sequence redundancy is reduced using 90% identity as a threshold. In this case, the number of sequences is 5 times lower than with UniProt (Table 1). So, when we run Sma3s with UniRef90 (2014) as the reference database, the time run is similar to that obtained when we used Swiss-Prot. But also more importantly, the accuracy is almost as high as when using UniProt. In fact, the number of significant Blast alignments is similar to that obtained when we used the complete UniProt database.

The query dataset is composed of new sequences that are currently better known. So, when we use Sma3s with the current release of UniRef90 (2016), the accuracy is now the highest, with a recall close to 80% and a precision of 63%. All of this proposes the last version of UniRef90 as the reference database to annotate proteins and protein-coding sequences with a high accuracy and a short run time.

To compare Sma3s with other protein functional annotator we chose Argot, since it has recently showed a high accuracy [28]. We annotate the query dataset with this computational tool, which assigns GO terms, and used databases from 2014 in its web version. The obtained accuracy by this method shows a recall higher than Sma3s using UniRef90 2014 (but not when using the 2016 version), though its precision is lower (31%) versus Sma3s (37%) (Fig. 2b).

It is important to note that, even though the recall is higher with Argot, this value changes depending on the type of GO term. The higher recall of Argot seems related to the most generic terms coming from the cellular component ontology (Fig. 3). However the molecular function and biological process ontologies have better results with Sma3s versus Argot, especially in more specific terms. Although cellular localization is important to know the place where a protein exerts

its function, molecular functions and even more biological processes are more demanded in projects analyzing gene expression or comparing the annotation of phylogenetically related organisms.

3.3. Annotating complete proteomes and analyzing functional categories

Sma3s is useful to annotate complete proteomes in a short time. When it finishes, it offers a report with different annotation types, including the most probable gene name and protein description, EC numbers for enzymes, GO terms, UniProt keywords and pathways. But one of the most demanded usability for massive annotators is the possibility of giving summaries with information about functional categories. This could be useful, for example, to compare annotations of different organisms in the same project.

To enable this functionality, Sma3s now reports a summary with different categories and the number of sequences belonging to each category. To present this new functionality, we selected *Bacillus subtilis* as a complete proteome to annotate. From the 3940 proteins from *B. subtilis*, Sma3s was able to assign annotations to 3583 of them (91%), with GO and keywords as the most abundant annotated terms (Suppl. file 2). From the results, the different biological processes of this bacterium can be studied from GO Slim as well as from UniProt keywords categories (Fig. 4). The coverage from the former is higher (it offers an average of 2 terms by each protein), but keywords offer novel terms which can be very useful, as it can be checked in the present example with the group *Sporulation*. Hence, Sma3s found 255 proteins related to *Sporulation* in this well-known sporulated bacterium [29]. These different functional categories sources offer complementary results. For example, whereas the number of *transport* proteins is much higher in keyword than in GO Slim category, the number of proteins involved in *carbohydrate metabolism* is lower (Fig. 4). All of this allows a more complete collection of the biological processes characteristics of the organism to annotate.

These reports with functional categories can be used to analyze the annotation of a simple proteome or transcriptome, but also to compare different functional annotations coming from different organisms. To show this utility, we annotate 52 proteomes of the *Bacillus* genus using Sma3s, and compare 2 representative annotations from both biological process categories: *Cell*

motility from GO Slim, and *Antibiotic biosynthesis* from UniProt keywords (Fig. 5). All annotated proteomes have a similar number of proteins involved in cell motility, with the exception of *B. gaemokensis* [30], and *B. mycoides*, which is one of the rare *Bacillus* that has been previously reported to lack of motility [31]. On the other hand, several species highlight if we check the other biological process, *antibiotic biosynthesis*. For *B. subtilis*, the representative species of the genus, Sma3s finds 45 proteins related with this annotation, and it is known that this species produces more than two dozen of different antibiotics of a great variety of types [29]. Further, we have been able to support this interesting ability in two more species, *B. megaterium*, and different strains of *B. amyloliquefaciens* [32–34]. All together shows a practical example of a discovering of bacterial strains of interest in a specific field, with for example application in medicine.

3. Concluding remarks

Here we show that Sma3s is a useful computational tool for annotate large datasets of protein sequences. In fact, Sma3s does not only annotate protein sequences but also transcriptomes in a short time and with minimal requirements. Furthermore, the most important characteristic of Sma3s is that it can be used to annotate complete sequence datasets by any user without computational knowledge, which will allow increasing the knowledge about their organisms in study.

Funding

This research was supported by the Ministry of Economy and Competitiveness of the Spanish Government grant BFU2013-46923-P. Authors declare no competing interest.

Acknowledgements

We would like to thank C3UPO and CICA for the HPC support.

References

[1] Rougon-Cardoso, A., Flores-Ponce, M., Ramos-Aboites, H.E., Martínez-Guerrero, C.E., et al., The genome, transcriptome, and proteome of the nematode Steinernema carpocapsae: evolutionary signatures of a pathogenic lifestyle. *Sci. Rep.* 2016, 6, 37536.

[2] Castro, J.C., Maddox, J.D., Cobos, M., Requena, D., et al., De novo assembly and functional annotation of Myrciaria dubia fruit transcriptome reveals multiple metabolic pathways for L-ascorbic acid biosynthesis. *BMC Genomics* 2015, 16, 997.

[3] Renuse, S., Chaerkady, R., Pandey, A., Proteogenomics. *Proteomics* 2011, 11, 620–630.

[4] Prochnik, S.E., Umen, J., Nedelcu, A.M., Hallmann, A., et al., Genomic analysis of organismal complexity in the multicellular green alga Volvox carteri. *Science* 2010, 329, 223–226.

[5] Venturini, L., Ferrarini, A., Zenoni, S., Tornielli, G.B., et al., De novo transcriptome characterization of Vitis vinifera cv. Corvina unveils varietal diversity. *BMC Genomics* 2013, 14, 41.

[6] Radivojac, P., Clark, W.T., Oron, T.R., Schnoes, A.M., et al., A large-scale evaluation of computational protein function prediction. *Nat. Methods* 2013, 10, 221–227.

[7] Minneci, F., Piovesan, D., Cozzetto, D., Jones, D.T., FFPred 2.0: improved homology-independent prediction of gene ontology terms for eukaryotic protein sequences. *PloS One* 2013, 8, e63754.

[8] Koskinen, P., Törönen, P., Nokso-Koivisto, J., Holm, L., PANNZER: high-throughput functional annotation of uncharacterized proteins in an error-prone environment. *Bioinformatics* 2015, 31, 1544–1552.

[9] Conesa, A., Gotz, S., Garcia-Gomez, J.M., Terol, J., et al., Blast2GO: a universal tool for annotation, visualization and analysis in functional genomics research. *Bioinformatics* 2005, 21, 3674–6.

[10] Grabherr, M.G., Haas, B.J., Yassour, M., Levin, J.Z., et al., Full-length transcriptome assembly from RNA-Seq data without a reference genome. *Nat. Biotechnol.* 2011, 29, 644–652.

[11] Chen, T.-W., Gan, R.-C.R., Wu, T.H., Huang, P.-J., et al., FastAnnotator--an efficient transcript annotation web tool. *BMC Genomics* 2012, 13 Suppl 7, S9.

[12] Hotz-Wagenblatt, A., Hankeln, T., Ernst, P., Glatting, K.H., et al., ESTAnnotator: A tool for high throughput EST annotation. *Nucleic Acids Res.* 2003, 31, 3716–9.

[13] Muñoz-Mérida, A., Viguera, E., Claros, M.G., Trelles, O., Pérez-Pulido, A.J., Sma3s: a three-step modular annotator for large sequence datasets. *DNA Res.* 2014, 21, 341–353.

[14] García-Romero, I., Pérez-Pulido, A.J., González-Flores, Y.E., Reyes-Ramírez, F., et al., Genomic analysis of the nitrate-respiring Sphingopyxis granuli (formerly Sphingomonas macrogoltabida) strain TFA. *BMC Genomics* 2016, 17, 93.

[15] Mikhailov, K.V., Simdyanov, T.G., Aleoshin, V.V., Genomic survey of a hyperparasitic microsporidian Amphiamblys sp. (Metchnikovellidae). *Genome Biol. Evol.* 2016.

[16] Masuya, H., Manabe, R.-I., Ohkuma, M., Endoh, R., Draft Genome Sequence of Raffaelea quercivora JCM 11526, a Japanese Oak Wilt Pathogen Associated with the Platypodid Beetle, Platypus quercivorus. *Genome Announc.* 2016, 4.

[17] Cho, O., Ichikawa, T., Kurakado, S., Takashima, M., et al., Draft Genome Sequence of the Causative Antigen of Summer-Type Hypersensitivity Pneumonitis, Trichosporon domesticum JCM 9580. *Genome Announc.* 2016, 4.

[18] Perina, A., González-Tizón, A.M., Meilán, I.F., Martínez-Lage, A., De novo transcriptome assembly of shrimp Palaemon serratus. *Genomics Data* 2017, 11, 89–91.

[19] Pascual, J., Alegre, S., Nagler, M., Escandón, M., et al., The variations in the nuclear proteome reveal new transcription factors and mechanisms involved in UV stress response in Pinus radiata. *J. Proteomics* 2016, 143, 390–400.

[20] Carmona, R., Zafra, A., Seoane, P., Castro, A.J., et al., ReprOlive: a database with linked data for the olive tree (Olea europaea L.) reproductive transcriptome. *Front. Plant Sci.* 2015, 6, 625.

[21] Kumar, V., Kutschera, V.E., Nilsson, M.A., Janke, A., Genetic signatures of adaptation revealed from transcriptome sequencing of Arctic and red foxes. *BMC Genomics* 2015, 16, 585.

[22] Benzekri, H., Armesto, P., Cousin, X., Rovira, M., et al., De novo assembly, characterization and functional annotation of Senegalese sole (Solea senegalensis) and common sole (Solea solea) transcriptomes: integration in a database and design of a microarray. *BMC Genomics* 2014, 15, 952.

[23] Nourisson, C., Muñoz-Merida, A., Carneiro, M., Sequeira, F., De novo transcriptome assembly and polymorphism detection in two highly divergent evolutionary units of Bosca's newt (Lissotriton boscai) endemic to the Iberian Peninsula. *Mol. Ecol. Resour.* 2016.

[24] Gene Ontology Consortium, Gene Ontology Consortium: going forward. *Nucleic Acids Res.* 2015, 43, D1049-1056.

[25] The UniProt Consortium, UniProt: the universal protein knowledgebase. *Nucleic Acids Res.* 2017, 45, D158–D169.

[26] Holliday, G.L., Davidson, R., Akiva, E., Babbitt, P.C., Evaluating Functional Annotations of Enzymes Using the Gene Ontology. *Methods Mol. Biol.* 2017, 1446, 111–132.

[27] Kersey, P.J., Allen, J.E., Armean, I., Boddu, S., et al., Ensembl Genomes 2016: more genomes, more complexity. *Nucleic Acids Res.* 2016, 44, D574-580.

[28] Lavezzo, E., Falda, M., Fontana, P., Bianco, L., Toppo, S., Enhancing protein function prediction with taxonomic constraints--The Argot2.5 web server. *Methods* 2016, 93, 15–23.

[29] Stein, T., Bacillus subtilis antibiotics: structures, syntheses and specific functions. *Mol. Microbiol.* 2005, 56, 845–857.

[30] Nakamura, L.K., Jackson, M.A., Clarification of the Taxonomy of Bacillus mycoides. *Int. J. Syst. Evol. Microbiol.* 1995, 45, 46–49.

[31] Jung, M.Y., Jung, M.-Y., Paek, W.K., Park, I.-S., et al., Bacillus gaemokensis sp. nov., isolated from foreshore tidal flat sediment from the Yellow Sea. *J. Microbiol.* 2010, 48, 867–871.

[32] Malanicheva, I.A., Kozlov, D.G., Sumarukova, I.G., Efremenkova, O.V., et al., Antimicrobial activity of Bacillus megaterium strains. *Mikrobiologiia* 2012, 81, 196–204.

[33] Jeong, H., Park, S.-H., Choi, S.-K., Genome Sequence of Antibiotic-Producing Bacillus amyloliquefaciens Strain KCTC 13012. *Genome Announc.* 2015, 3.

[34] Arguelles-Arias, A., Ongena, M., Halimi, B., Lara, Y., et al., Bacillus amyloliquefaciens GA1 as a source of potent antibiotics and other secondary metabolites for biocontrol of plant pathogens. *Microb. Cell Factories* 2009, 8, 63.

Figure legends

Figure 1. Organigram of the Sma3s pipeline. Every sequence from the query dataset is compared to the reference database using Blast (blastp for proteomes and blastx for transcriptomes). This database can be filtered to taking in account only sequences of expected high quality. Sma3s prioritizes annotations (especially gene name and description) coming from very similar or orthologous proteins. The last module enriches the annotations with proteins sharing short similarity regions but with common annotations. Finally, Sma3s gives the results in two files, one of them with all the annotations for each query sequence, and the other with functional categories that can be used to both create figures and compare different annotations.

Figure 2. Annotation accuracy when using different reference databases. Recall and precision is represented by percent values, and run times (dashed line) is represented in minutes. The results are shown both a) when UniProt keywords, and b) GO terms are evaluated. In all cases the 2014 versions of the databases were used, except for UniRef90 version 2016. Argot, which only predicts GO terms, use a 2014 version.

Figure 3. Number of predicted GO terms by different databases and methods. The background in grey color represents the number of specific GO terms in the protein entries from UniProt database. GO terms coming from different ontologies are highlighted in red (Cellular Component), blue (Molecular Function), and green (Biological Process) color.

Figure 4. Number of proteins predicted to be involved in different biological processes. The figure represents the percentage of proteins belonging to different biological processes from a) GO Slim and b) UniProt keyword categories. The number of proteins predicted to each process is shown in brackets.

Figure 5. Number of proteins predicted to be involved in antibiotic biosynthesis and motility from different species of *Bacillus*. Proteins annotated in the antibiotic biosynthesis were extracted from UniProt keyword category, and those annotated in cell motility were extracted from GO Slim category.

Tables

Table 1. Number of annotated proteins within different databases and number of Blast hits obtained for all the sequences in the test dataset. Two different versions are used for the different databases.

	2014			2016	
	Swiss-Prot	UniProt	UniRef90	UniProt	UniRef90
Number of annotated proteins	524,643	52,924,113	10,809,500	47,715,549	21,330,801
Number of Blast hits	36,000	65,636	63,421	71,995	68,811

Supporting information

Supplementary file 1. Test dataset of 349 proteins in FASTA format.

Supplementary file 2. Summary offered by Sma3s for the annotation of *B. subtilis* proteome.

Every different category has been highlighted using different colors.