BactPepDB: a database of predicted peptides from an exhaustive survey of complete prokaryote genomes

Julien Rey1,2,3, Patrick Deschavanne1,2 and Pierre Tuffery1,2,3,*

1INSERM, U973, MTi, F-75205 Paris, France, 2Université Paris Diderot, Sorbonne Paris Cité, F-75205 Paris, France and 3RPBS, F-75205 Paris, France

*Corresponding author: Tel: +(331) 57 27 83 74; Fax: +(331) 57 27 83 72; Email: pierre.tuffery@univ-paris-diderot.fr


Received 16 May 2014; Revised 7 October 2014; Accepted 8 October 2014

Abstract

With the recent progress in complete genome sequencing, mining the increasing amount of genomic information available should in theory provide the means to discover new classes of peptides. However, annotation pipelines often do not consider small reading frames likely to be expressed. BactPepDB, available online at http://bactpepdb.rpbs.univ-paris-diderot.fr, is a database that aims at providing an exhaustive re-annotation of all complete prokaryotic genomes—chromosomal and plasmid DNA—available in RefSeq for coding sequences ranging between 10 and 80 amino acids. The identified peptides are classified as (i) previously identified in RefSeq, (ii) entity-overlapping (intragenic) or intergenic, and (iii) potential pseudogenes—intergenic sequences corresponding to a portion of a previously annotated larger gene. Additional information is related to homologs within order, predicted signal sequence, transmembrane segments, disulfide bonds, secondary structure, and the existence of a related 3D structure in the Protein Databank. As a result, BactPepDB provides insights about candidate peptides, and provides information about their conservation, together with some of their expected biological/structural features. The BactPepDB interface allows to search for candidate peptides in the database, or to search for peptides similar to a query, according to the multiple properties predicted or related to genomic localization.

Database URL: http://www.yeastgenome.org/

Introduction

Peptides and mini proteins have recently met a regain of interest for therapeutic applications (1, 2). For one part, several breakthroughs have allowed significant progress on some traditional weaknesses of peptides as candidate therapeutics. Chemical modifications among which, to cite some, pegylation (3), stapling (4), glycosilation (5), or the construction of chimeric molecules combining cargos and
peptides (6) have shown effective to significantly increase the in vivo stability of peptides or their delivery in targeted cells. In addition, the characterization of new classes of peptides offers promising perspectives for development. Antimicrobial peptides are expected to address the urge for the discovery of new antibiotics (7). Bacterial quorum sensing peptides that participate in cell-to-cell communication and bacterial adaptation to specific conditions could also lead to new ways to control bacterial proliferation (8). Peptides extracted from venoms have shown to target very specifically various receptors (9). Cell penetrating peptides raise new promises for the controlled cell-specific penetration of peptides (10).

In general however, the rate of discovery of new peptides remains low. Most frequently, the characterization of natural peptides still relies on cycles of purification and sequencing that prevent large scale exploration. Alternative routes rely on the use of phage display techniques [see for instance (11, 12)] or combinatorial chemical methods (13, 14) that have led to success in the identification of peptides targeting specific protein–protein interactions [e.g. (15)]. Such approaches are however labor-intensive and costly.

Thanks to the progress of high throughput sequencing techniques, an increasing amount of complete genome sequences is becoming available which could constitute an important source for the identification of candidate peptides. Unfortunately, mining this amount of information for the discovery of new active peptides still faces some challenges. First, the identification of candidate genes from genomic sequences cannot be directly related to their effective expression. For eukaryotes, even for expressed genes, the knowledge of genomic sequences does not provide sure information about the expressed sequences, owing to the presence of exons and the fact that peptides such as hormones, for instance, often result from the maturation of preproteins [e.g. (16)]. Prokaryotes do not possess such features to the same extent, and gene identification and chromosomal information should in theory be more straightforward, although not exhaustive, because peptides effectively expressed can be encoded in non-genomic nucleic acid, such as that of plasmids. However, the identification of short coding sequences (SCSs) can be challenging (17), and as a consequence, standard annotation pipelines have shown to identify only few such SCSs (18). Indeed, very few candidate genes of size <50 residues are annotated in the RefSeq database (curated non-redundant sequence database of genomes) (19) in FASTA format, together with the organism taxonomy and the corresponding annotations about gene location in the GenBank format. The complete genome and plasmids sequences are then processed using BactgeneSHOW (39), a program specifically designed for the prokaryote genome-wide identification of SCSs. BactgeneSHOW relies on Hidden Markov Models that account for the presence of ribosome-binding sites (RBS) and four types of nucleotide composition of the coding sequences. It analyses both direct and reverse strands. It has been successfully used to identify new SCSs whose expression has been further validated by transcriptomic analyses (17). Predicted SCSs are then translated into amino acid sequences, and only sequences of size between 10 and 80 amino acids are considered for further analysis.

**Construction and Content**

**Genome wide identification of the candidate peptides**

An overview of the processing of BactPepDB is presented in Figure 1. Full genome sequences are collected from RefSeq database (curated non-redundant sequence database of genomes) (19) in FASTA format, together with the organism taxonomy and the corresponding annotations about gene location in the GenBank format. The complete genome and plasmids sequences are then processed using BactgeneSHOW (39), a program specifically designed for the prokaryote genome-wide identification of SCSs. BactgeneSHOW relies on Hidden Markov Models that account for the presence of ribosome-binding sites (RBS) and four types of nucleotide composition of the coding sequences. It analyses both direct and reverse strands. It has been successfully used to identify new SCSs whose expression has been further validated by transcriptomic analyses (17). Predicted SCSs are then translated into amino acid sequences, and only sequences of size between 10 and 80 amino acids are considered for further analysis.
Compared to the classical upper limit of peptides of 50 amino acids, the limit of 80 amino acids comes since in some cases, a leader sequence usually on the order of 20–30 amino acids can exist. All such candidate SCSs corresponding to genes already annotated in RefSeq (19) are first identified and inserted in BactPepDB along with their annotations. Then, all the SCSs and RBS genomic coordinates identified by BactgeneSHOW are integrated into the database and the newly detected SCSs are labeled as belonging to intergenic regions or coding regions (entity-overlapping sequences).

Potential pseudogenes detection

In order to facilitate the identification of false positives, we also perform a search for potential pseudogenes. Pseudogenes are dysfunctional relatives of genes that have lost their protein-coding ability or are otherwise no longer expressed in the cell. Pseudogenes can arise from a partial duplication of active genes and can be found in the intergenic regions. We thus perform a similarity-based approach to detect partially duplicated genes. This approach takes the set of genes detected in intergenic regions of a genome and compares it to the set of RefSeq annotated gene sequences of the same genome. This comparison is accomplished by using the blastp algorithm (40). Cutoff values of 50% identity and 50% coverage of the shortest sequence are set to filter those sequence hits. We recall however that as BactGeneSHOW accounts for the presence of the RBS, all BactPepDB SCSs labeled as potential pseudogenes are possibly expressed.

Similarity searching

Similarity searching is performed using blastp (40) on the protein sequences. Inter-species sequence conservation inference is achieved by performing an all-against-all search over all the peptide sequences of all the species of an order (as given by organismal taxonomy of RefSeq), considering a sequence identity and coverage of >50% each. Intergenic SCSs are then classified into three different categories according to Warren et al. (18). Intergenic SCSs that align to annotated genes from other species are classified as ‘similar to RefSeq entry’ in the genome to which they belong. Intergenic SCSs that align to entity-overlapping sequences from other species are classified as ‘genomic artifacts’. Finally, intergenic SCSs that align to intergenic sequences from other species are labeled as ‘potentially missing’. For these last ones, we require that hit sequence(s) belong to a different taxonomic family. Indeed, a requirement based on different species would not be satisfactory, as there are prokaryotes classified as different species with very similar intergenic sequences due to lack of divergence (for example, Brucella species (41)). Furthermore, the species and genera levels of classification have been shown to be highly variable in prokaryotes (42–45). As such, the next highest taxonomic level (the family) is considered. This requirement is the main evidence used to distinguish sequences that are likely to be real genes from sequences that represent some other conserved elements. In order to give a qualitative estimate of the conservation, two levels of conservation are also provided: a weak conservation stands for the identification of a candidate ortholog in another species of a genus (191 genera), whereas strong conservation stands for the existence of a candidate ortholog in more than two species of a genus, whatever the number of species (105 genera).

Integration of predicted features

For each candidate peptide, several supplementary analyses are performed and include: (i) the prediction of secondary structure using PsiPred (46), (ii) the prediction of the local conformation prediction as a Structural Alphabet profile (47), (iii) the search for the existence of candidate structure by performing a blastp against the protein data bank (PDB) (48), (iv) the prediction of putative disulfide bonds using DIPro (49), (v) the prediction of transmembrane segments using TMHMM (50), and (vi) the prediction of signal peptides using SignalP (51). Finally, cross references with external databases have also been considered but are presently limited to Bactibase (23).
Update strategy

Periodical updates of the database are scheduled. Due to the computational cost however, we intend to keep the database up to date on a 6-month basis only.

Database architecture

BactPepDB was implemented using the combination of the Perl CGI programming language (5.10.1) and a MySQL relational database (5.1.66). The site is running on Apache server (2.2.16-6) installed on a Kernel-based Virtual Machine with Debian 6.0.6 as operating system.

Database access


Database interface

The search page proposes two possibilities to query BactPepDB. The first is to search for homologs of a peptide sequence among the BactPepDB collection, the second is to search for peptides of BactPepDB matching various criteria. Different parameters can be combined to focus the search: (i) source organism, which represents the sequenced genome the peptides are predicted from, (ii) peptide status, which makes possible to restrict the search to predicted peptides or to peptides already annotated in RefSeq (19), (iii) predicted peptide features, which makes possible to specify information such as the location of the predicted gene (coding or intergenic region), and for peptides in the intergenic regions, if it is a potential pseudo-gene or not, (iv) sequence features like sequence length or the ability to use regular expressions, and (v) other predicted features such as the presence of a transmembrane segment or a disulfide bond, and peptide conservation across different species of an order (see Similarity Searching section). After running a search, a brief summary of the number of hits matching the different criteria is returned, followed by a list of the hits with the details of the values. This list can be sorted interactively using each criterion as sort key. Finally, links to external resources are provided. For the similarity search results, pair-wise alignments of the queries with the corresponding hits are shown, but the user also has the possibility to request for a multiple alignment of all the hits using Clustal Omega (54) along with a representation of the residue frequencies at each position produced by WebLogo (55). Each peptide in the database has a unique accession number (e.g. BPDB: 0845925). Subsequently clicking on a BPDB id will lead to the peptide entry content which contains all gathered information concerning this peptide. At any time, users can move back to the search page and narrow their search as search parameters are preserved.

Results and Discussion

On date of 8 August 2014, the database contains 1 747 413 Peptides from 1226 species (2240 strains including 1598 plasmids) belonging to 557 genera, 218 families and 97 orders. Looking at the agreement between the candidates identified by BactGeneSHOW and the genes previously annotated in RefSeq, we observe that 74% of the SCSs annotated in RefSeq are detected by BactGeneSHOW, in which 66% of the SCSs match exactly, and only 8% differ by either their start or stop positions. The remaining 26% could correspond to SCSs missed by BactGeneSHOW, to SCSs of RefSeq not identified using an automated annotation pipeline and thus possibly having non-typical start codons but observed as expressed by biologists or to genes discarded by the truncature to 80 amino acids. This highlights the limits of a fully automated procedure, but still, a significant amount of information can be retrieved.

Among the candidates identified in BactPepDB that are not annotated in RefSeq, one denotes, as illustrated in Figure 2, a clear increase of the number of new entries for smaller SCSs. As a consequence of our flexible definition of potential pseudogenes, the fraction of non-potential pseudogenes new intergenic SCSs for sizes of > 50 amino acids is actually marginal. This suggests that the approach does not identify false positives with high frequencies. Conversely, very few of the candidate sequences identified correspond to already annotated sequences for sizes between 10 and 30 amino acids. Furthermore, one observes some kind of compensation between the increasing number of new intergenic candidates and the decreasing number of RefSeq entries that maintains the number of such peptides rather constant from 80 down to 20 amino acids. Some increase of new intergenic candidates is observed for sizes from 10 to 15 amino acids, which is difficult to interpret at the moment, and remains the subject for further investigation. The fraction of the RefSeq SCSs, the potential pseudogenes peptides, the intra and intergenic SCSs are reported in Table 1, using a three class separation depending on peptide size, to distinguish regions in which new intergenic SCSs are, roughly, predominant, equal and marginal compared to previous RefSeq entries. Overall, SCSs not previously annotated in RefSeq
correspond to 70% of BactPepDB. Among these, over 263 000 (15%) are predicted to have disulfide bonds, over 173 000 (10%) to have transmembrane segments and over 30 000 (1.7%) to have signal peptides. Structural homologs of the PDB could be detected for over 100 000 new SCSs.

Considering the intra-genus and intra-order conservation, we observe that close to 184 000 new SCSs detected in the intergenic regions are conserved to some extent across different species of a genus, whereas 112 000 of them are conserved across different taxonomic families of an order. Table 2 presents a comparison of the fraction of the conserved SCSs for these new intergenic SCSs and those preexisting in RefSeq. Interestingly, the fraction of intergenic SCSs that are conserved is similar to that of RefSeq, which suggests that the information of these newcomers is consistent with preexisting one. The fraction of the conserved peptides identified in the intergenic regions appears stable depending of peptide size when considering the different species of a genus (intra-genus) or the different families of an order (intra-order). Overall, depending on the taxonomic level chosen as a requirement for conservation significance, at least 18% of the new intergenic SCSs are conserved.

Table 1. BactPepDB entries by categories

<table>
<thead>
<tr>
<th></th>
<th>Small</th>
<th>Medium</th>
<th>Large</th>
</tr>
</thead>
<tbody>
<tr>
<td>RefSeq</td>
<td>3946 (0.6)</td>
<td>143 157 (30.4)</td>
<td>362 395 (63.2)</td>
</tr>
<tr>
<td>Potential pseudogenes</td>
<td>201 228 (28.6)</td>
<td>51 752 (11.0)</td>
<td>11 743 (2.1)</td>
</tr>
<tr>
<td>Intergenic</td>
<td>324 533 (46.2)</td>
<td>189 573 (40.2)</td>
<td>121 897 (21.3)</td>
</tr>
<tr>
<td>Entity-overlapping</td>
<td>173 270 (24.6)</td>
<td>86 907 (18.4)</td>
<td>77 012 (13.4)</td>
</tr>
</tbody>
</table>

The three categories correspond to small (<30 amino acids), medium (30–50 amino acids) and large (>50 amino acids) peptide sizes. Peptides already annotated in RefSeq are distinguished from newcomers of BactPepDB categorized as potential pseudogenes, intergenic and entity-overlapping. Fractions in % within brackets.

Table 2. Conserved SCSs

<table>
<thead>
<tr>
<th></th>
<th>Small</th>
<th>Medium</th>
<th>Large</th>
</tr>
</thead>
<tbody>
<tr>
<td>Intra-genus RefSeq</td>
<td>978 (25)</td>
<td>40 913 (29)</td>
<td>152 429 (42)</td>
</tr>
<tr>
<td>New intergenic SCSs</td>
<td>84 276 (26)</td>
<td>57 584 (30)</td>
<td>41 447 (34)</td>
</tr>
<tr>
<td>Intra-order RefSeq</td>
<td>750 (19)</td>
<td>20 041 (14)</td>
<td>60 194 (17)</td>
</tr>
<tr>
<td>New intergenic SCSs</td>
<td>61 662 (19)</td>
<td>28 047 (15)</td>
<td>21 940 (18)</td>
</tr>
</tbody>
</table>

Numbers and fractions (% within brackets) of peptide entries that are conserved across species of a genus (intra-genus) and across families of an order (intra-order). Fractions are relative to the total number of entries in each category (see Table 1). The three categories correspond to small (<30 amino acids), medium (30–50 amino acids) and large (>50 amino acids) peptide sizes. Peptides already annotated in RefSeq are distinguished from the new intergenic SCSs of BactPepDB.
Evaluation of predictions

As the database contains predicted candidates, it is important to assess how likely it can assist the identification of truly expressed peptides. A way to assess BactPepDB-added value comes from experimental studies focusing on specific genomes. For instance, 14 CDSSs that were missing from the initial annotation of Vibrio splendidus LGP32 were recently uncovered (56, 58). We found that 12 of our predictions overlap these missing CDSSs. To assess this on a larger scale, we have also compared two versions of BactPepDB based on two versions of the RefSeq database (on date of 11 June 2013 and 30 September 2013) and have found that 125 newly annotated peptides of size comprised between 10 and 80 amino acids were added for genomes that are common to both versions, 33 of which are not of the ‘predicted’ kind and were biologically confirmed. BactGeneSHOW had correctly predicted 89 of these newcomers in the previous version of BactPepDB, among which 24 are now biologically confirmed, which means that about 70% of those newly annotated peptides were already present in BactPepDB before making it to the RefSeq database. Among those predicted peptides, 83 are conserved across different species and only six were unique in their respective order, supporting that peptide conservation is a good measure of peptide expression likelihood.

As the core of BactPepDB relies on BactGeneSHOW, we have also run other gene prediction programs over these genomes to assess BactGeneSHOW performance. GeneMarkHMM 2.6 was able to retrieve 94 of these newcomers, whereas Prodigal 2.5 could only find 38 of them. Although GeneMarkHMM 2.6 slightly outperformed BactGeneSHOW, it is important to note that GeneMarkHMM was apparently inefficient for some genomes, for instance in Flavobacterium psychrophilum JIP02/86 where none of the five newly annotated genes was detected, whereas BactGeneSHOW retrieves them. Indeed, GeneMarkHMM relies on precalculated heuristic models which may not be suitable for all species whereas BactGeneSHOW relies on a self-learning algorithm.

Finally, another important point to assess is the expected proportion of false positives present in the database. Although this is a very difficult question to answer, we recently gained some insight through RNA deep sequencing data, which reveals smaller intergenic transcripts and mRNA extensions. Analysis of new transcripts from Escherichia coli str. K-12 substr. MG1655 (56, 58) showed that only 74 predicted sequences of BactPepDB were overlapping the 1094 potentially non-coding transcripts (ncRNA) and long 5’-UTR extensions detected in the intergenic regions of MG1655. This is interesting enough because only a very small fraction of these 1094 transcribed regions is supposed to code for peptides.

Searching for Bactibase homologs

We illustrate here the use of BactPepDB to the search for homologs of Bactibase (23). Bactibase is, in our experience, the only database devoted to antimicrobial peptides for which the complete sequence collection could be downloaded. Among the 219 entries, 197 peptide sequences match the condition to have sequences of size between 10 and 80 residues without non-standard amino acids in their sequence. The genomic information (chromosomal and plasmidic) corresponding to the genus/species was present in BactPepDB for 146 of them. However, this condition does not imply the information should be present in BactPepDB since for one part some variation between strains of a species can occur, and for another part, some peptides can result from the cleavage of preproteins larger than 80 residues, thus out of the scope of BactPepDB. A careful inspection of the literature reporting peptide identification for each Bactibase entry showed that peptides not found in BactPepDB correspond to 11 cases for which preproteins are larger than 80 amino acids, and 37 cases for which it was not possible to conclude, owing to the fact that the peptide sequence was not elucidated using genomic information or that it was not possible to conclude between chromosomal or plasmidic encoding. As a result, 98 peptides only were clearly in the scope of BactPepDB.

A similarity-based search in BactPepDB—accepting a correct identification for a hit in the same species, and with over 90% identity—led to the identification of 56 of them. RefSeq annotations were only present for 34 cases over 56. Thus BactPepDB was able to infer new knowledge for 22 cases over 98, a gain of 22%. Furthermore, hits at a lesser sequence identity were found for 22 more peptides. BactPepDB was thus able to grab information for 76 peptides over 98. We remind that all strains of a species are not expected to produce all antimicrobial peptides [see for instance (59–62)]. Overall, such results illustrate that the re-annotation of the complete genome using a method specialized for SCSS can have added value, at least as a preliminary step to confront with additional information.

Comparison with BAGEL, a database of predicted bacteriocins

We have also analysed the consistency of BactPepDB with BAGEL (24), a resource predicting bacteriocins from genomic data, over a collection of 15 genomes of different genera: Acaryochloris marina MBIC11017, Achromobacter xylosoxidans A8, Bacillus cereus AH187, Bacillus subtilis BSn5, Enterobacter cloacae SCF1, Escherichia coli W, Geobacillus sp CS6 T3, Lactobacillus casei W56, Methanococcus voltae A3, Mycobacterium tuberculosis
Conclusions and Future Directions

BactPepDB is a database of predicted peptides from a exhaustive survey of complete prokaryote genomes. BactGeneSHOW being a generic approach to the search for SCsS, taking into account the complete spectrum of prokaryotes from archaea to bacteria and the diversity of each category, it is expected that due to the variability in start codon and codon usage, some part of the truly expressed SCsS are not detected. Genome coding specificity, particularly that existing for bacteria and archaea, could be integrated in BactGeneSHOW but this remains the subject for further work. In addition, from our analyses, BactPepDB already shows the ability to retrieve a large part of previously annotated biological peptides when in the scope of the database. BactPepDB could be improved in several other directions. At present, this precludes important sources of prokaryotic information such as those with unusual codons, as well as the incomplete genomes available in RefSeq or other databases from which it should be possible to increase the knowledge of the degree of conservation of candidates. Particularly, it could be of interest to add data from the Ensembl Bacteria database (63) as it contains, on average, more strains per species. Another limit is related to the impossibility to detect peptides resulting from the maturation of large proteins, which is presently beyond the scope of BactPepDB.

Accepting these limitations, it remains that BactPepDB appears to contain new knowledge about SCsS compared to previous RefSeq entries. Although, it is difficult to exactly assess the amount of candidate peptides that may be expressed in some physiological conditions, or that may have a biological activity, BactPepDB provides a rather unique panorama of SCsS over the complete collection of genomes available, at the level of individual sequences but also considering their conservation through genera. The close to 18% of BactPepDB newcomers conserved to some extent could be seeds for further investigations. The detection of small peptides being more difficult using biochemical analyses, BactPepDB is thus expected to assist the experimental discovery of new bioactive peptides.

Acknowledgements

The authors thank Pierre Nicolas and the MIG team for making available their BactGeneSHOW program, the eBio platform (Université Paris-Sud) for RNA-seq data and analysis, and F. Guyon and J. Muzard for useful discussions.

Funding

INSERM-University Paris Diderot UMR-S 973 and IBiSA (for deployment on the RPBS platform).

Conflict of interest: None declared.

References


