Published online 8 May 2017 *Nucleic Acids Research, 2017, Vol. 45, Web Server issue* ***W265–W269***
doi: 10.1093/nar/gkx389

P4P: a peptidome-based strain-level genome comparison web tool

Aitor Blanco-Míguez[1,2,3], Florentino Fdez-Riverola[1,2], Anália Lourenço[1,2,4,*] and Borja Sánchez[3]

[1]ESEI-Department of Computer Science, University of Vigo, Edificio Politécnico, Campus Universitario As Lagoas S/N 32004, Ourense, Spain, [2]CINBIO-Centro de Investigaciones Biomédicas, University of Vigo, Campus Universitario Lagoas-Marcosende, 36310 Vigo, Spain, [3]Department of Microbiology and Biochemistry of Dairy Products, Instituto de Productos Lácteos de Asturias (IPLA), Consejo Superior de Investigaciones Científicas (CSIC), Paseo Río Linares S/N 33300, Villaviciosa, Asturias, Spain and [4]CEB-Centre of Biological Engineering, University of Minho, Campus de Gualtar, 4710-057 Braga, Portugal

Received January 31, 2017; Revised April 10, 2017; Editorial Decision April 18, 2017; Accepted May 05, 2017

ABSTRACT

Peptidome similarity analysis enables researchers to gain insights into differential peptide profiles, providing a robust tool to discriminate strain-specific peptides, true intra-species differences among biological replicates or even microorganism-phenotype variations. However, no *in silico* peptide fingerprinting software existed to facilitate such phylogeny inference. Hence, we developed the Peptidomes for Phylogenies (P4P) web tool, which enables the survey of similarities between microbial proteomes and simplifies the process of obtaining new biological insights into their phylogeny. P4P can be used to analyze different peptide datasets, i.e. bacteria, viruses, eukaryotic species or even metaproteomes. Also, it is able to work with whole proteome datasets and experimental mass-to-charge lists originated from mass spectrometers. The ultimate aim is to generate a valid and manageable list of peptides that have phylogenetic signal and are potentially sample-specific. Sample-to-sample comparison is based on a consensus peak set matrix, which can be further submitted to phylogenetic analysis. P4P holds great potential for improving phylogenetic analyses in challenging taxonomic groups, biomarker identification or epidemiologic studies. Notably, P4P can be of interest for applications handling large proteomic datasets, which it is able to reduce to small matrices while maintaining high phylogenetic resolution. The web server is available at http://sing-group.org/p4p.

INTRODUCTION

Molecular-based methods for providing identification and species-level differentiation have proven to be very useful in phylogenetic studies, diagnostics and epidemiological surveillance, particularly where unusual phenotype makes the classical phenotypic identification difficult.

Typical differentiation methods are often challenged when high genetic similarity is shared among species/strains. DNA–DNA hybridization (DDH) is still considered the gold standard in bacterial taxonomy, but the labor-intensive and error-prone nature of DDH experiments and the limited information provided (only DDH values) prevents the establishment of a comparative database and incremental data use (1,2). Given that next-generation sequencing has delivered a rapid and cost-effective approach to obtaining whole-genome sequences of microbial strains, the analysis of genome sequence similarities has emerged as a natural replacement for DDH. Most notably, the existence of standard operating procedures for calculating genome-to-genome distances allows the re-use of genome sequence information in any subsequent comparisons and multiple ways of analysis in assessing taxonomic relationships, discovering new taxa and sharing data between researchers (3,4). Currently, the Genome-to-Genome Distance Calculator web service, implementing the Genome-BLAST Distance Phylogeny (GBDP) method, provides the highest correlation to conventional DDH (5).

We have shown in previous works that whole peptide fingerprinting can be used to complement the outputs of GBDP, i.e. experimental mass spectra may be used to cluster the bacteria, and more specifically it has been found useful for bacterial classification at the species and subspecies levels (6–8). However, till date, no *in silico* software facilitates phylogeny inference by peptide fingerprinting.

*To whom correspondence should be addressed. Tel: +34 988 387 013; Fax: +34 988 387 000; Email: analia@uvigo.es

Hereby, we present the Peptidomes for Phylogenies (P4P) server, which is the first web service to enable the *in silico* inference of bacterial taxonomy through the analysis of peptidomes. While following the same general principle of existing mass spectrometry approaches, our *in silico* peptide fingerprinting methodology uses whole genome data and *in silico* protein digestion to infer bacterial taxonomy, namely at the species and subspecies levels (9). The primary aim is to be able to generate a valid and manageable list of peptides that are potentially specific to each strain. Most notably, our methodology has been proven to support accurate phylogenetic reconstruction for conventionally challenging groups of organisms, such as the *Bacillus cereus* group of organisms (in combination to GBDP) (9) and the group of organisms in *Bifidobacterium* species (10). These differential peptide profiles could then be further investigated using *in vitro* approaches, such as LC-MS/MS, laying a foundation for the development of biomarker detection and application-specific methods. Notably, P4P can be of interest for applications handling large proteomic datasets, as it is able to reduce larger amounts of proteomic data to small matrices while maintaining high phylogenetic resolution.

MATERIALS AND METHODS

Processing method

P4P web service integrates well-established software tools, such as PSortB (11), mzJava (12), some algorithms from SPECLUST (13) and MrBayes (14). The subcellular locations of the proteins are predicted using the PSortB v3.0 tool. P4P resorts the mzJava tool to digest the proteins, using the major intestinal endoproteases, i.e. trypsin, chymotrypsin and pepsin (low specificity model, ph > 2).

The list of peptides for each strain is sampled based on peptide size, isoelectric point, subcellular location and digestion enzymes. SPECLUST tool is used to identify representative and reproducible peak masses that are present in all spectral profiles of replicates. The consensus spectra matrix is translated to a binary matrix (representing absence or presence of a given peptide mass) in NEXUS file format (15), which is then used to feed MrBayes for phylogenetic analysis purposes.

Since the analysis may be time consuming depending on the number and size of the uploaded data, P4P analysis runs in background and the user is provided with the link of the project's page so that the project status can be consulted at all times. In addition, if the user provided an email of contact, when a long process (such as uploading a project or generating a NEXUS file) finishes, the user will be notified by email.

Inputs and outputs

P4P can be used to analyze different peptide datasets, from bacteria to viruses, eukaryotic species or even metaproteomes, with the inclusion of few modifications regarding the prediction of the protein subcellular location (i.e. a virus cannot be classified as Gram positive or Gram negative). This could be of interest for developing more efficient applications, aimed at managing very large bacterial datasets, such as those required for epidemiologic studies. P4P has two main applications:

i) The tool can accept as input whole proteome data ('.faa' files) obtained from *in silico* methods. This data can be used to generate more or less large strain-specific peptide lists or peptidomes. These peptidomes enable the construction of phylogenetic trees, by running Speclust and MrBayes processes, which are the main outputs. In addition, user can plot protein distribution by subcellular location, isoelectric point and peptide length. A second output is the identification of strain-specific peptides, which can be used to trace back the source protein. This list of peptides may facilitate the development of biomarker detection and application-specific methods (e.g. a dairy starter or a probiotic that has to be traced through the human gut during clinical intervention studies).
ii) P4P can also accept as input experimental mass-to-charge lists originated from mass spectrometers, preferably [M+H]+ monoisotopic lists of masses in Da. This application can be used for handling large datasets, as happens in epidemiological studies, given the ability of the pipeline to handle peptide subsets in binary format whilst keeping the phylogenetic signal. If traced back, our application allows the detection of differential peptide profiles, providing a robust tool to discriminate not only strain-specific peptides, but also true intra-species differences among a set of biological replicates or even microorganism-phenotype variations (e.g. those occurring between biofilm and planktonic populations, or between very close bacteria strains).

Advanced options and customization

When running peptidome similarity analysis for whole proteome data obtained from *in silico* methods, the server allows the specification of the minimum and maximum thresholds for the peptide length and the isoelectric point, as well as the set of digestion enzymes and the protein subcellular locations to be considered in order to diminish the amount of data handled by the service.

P4P enables the use of three proteases representing the major intestinal endoproteases. In turn, subcellular location defines the putative location of the protein in the cell. This information is relevant because, for instance, extracellular proteins are used by the bacterium to communicate with its environment and thereby could help in bacteria differentiation. Finally, the establishment of peptide length and isoelectric point value ranges and the analysis of value distribution may help the user to narrow down the investigation, i.e. looking into a peptidome subset or, in turn, to identify the need to consider different subsets in order to able to achieve the desired differentiation.

RESULTS

Web service

From the user perspective, the analysis implemented in P4P consists of the following steps: (i) input of completely sequenced genomes or experimental mass-to-charge lists, (ii)

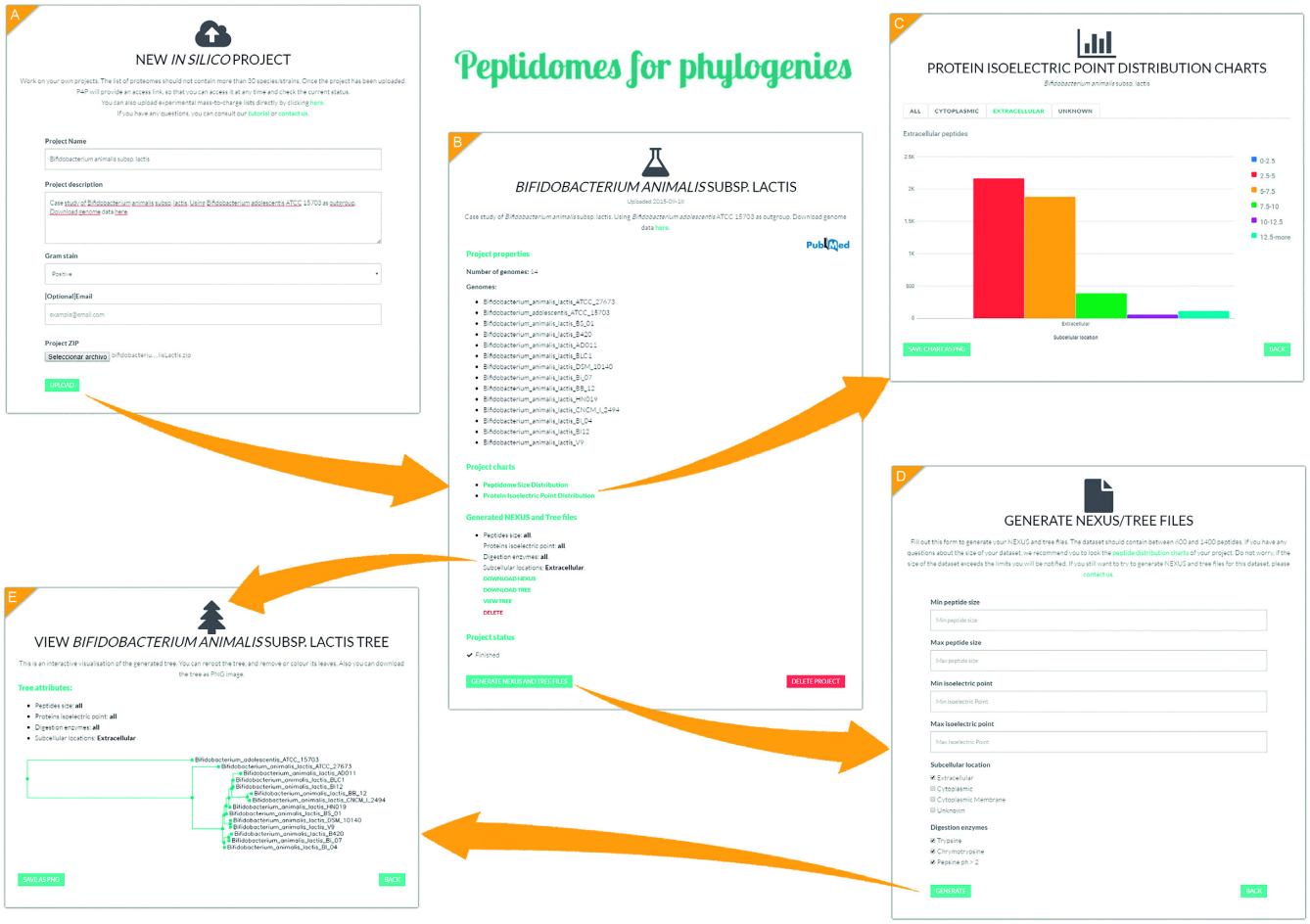

Figure 1. Screenshots of the web service running the analysis of the extracellular peptidomes of *Bifidobacterium animalis* subsp. *lactis* strains (case study number 2) during data upload (**A**), after data upload (**B**), while exploring distribution charts (**C**), during generation of Bayesian phylogenetic tree (**D**) and after generation of the Bayesian phylogenetic tree (**E**).

in silico digestion of proteins using human gut endopeptidases and (iii) comparison of the peptides according to their theoretical mass (peaks) and subsequent computation of consensus peak sets. After initial data processing (Figure 1A), the project is ready to generate the phylogenetic tree (Figure 1B). The user may select all peptides or filter them by peptide length, isoelectric point, subcellular location and/or digestion enzymes (Figure 1D). Peptidome analysis aims to identify the peak masses that are representative of the spectral profiles (Figure 1C). The resulting consensus spectra matrix is translated to a binary matrix that is used to generate a Bayesian phylogenetic tree (Figure 1E).

Case study 1: *Bacillus cereus* species complex

The methodology supporting P4P service was applied to the reconstruction of the *B. cereus* species complex, namely the differentiation of *Bacillus thuringiensis*, *Bacillus anthracis* and *B. cereus* strains (9). Results show that our method, as opposed to genome-sequence homology, is complementary to the proteome-based GBDP analysis and confirmed previous reports of this technology about the misclassification of many strains within the *B. cereus* group (14,15). Another important aspect of this evaluation refers to the computational complexity simplification generated by the P4P method, which was proven to reduce larger amounts of proteomic data to small matrices without losing phylogenetic signal. Therefore, P4P is considered of interest for developing more efficient applications aimed at managing very large bacterial peptide datasets, such as those generated in epidemiologic studies. Input data are provided in Supplementary Material S1 and as an example dataset in the web service. Also, some complementary benchmarking data is available in Supplementary Material S4.

Case study 2: *Bifidobacterium animalis* subsp. *lactis* strains

P4P pipeline was applied to the analysis of the peptidomes of publicly available *Bifidobacterium animalis* subsp. *lactis* strains, which have a genome identity of 99.975% (16), with the purpose of facilitating the identification of biomarkers and the development of application-specific detection methods (10). *B. animalis* subsp. *lactis* is by far the bifidobacteria more used in functional food products (17), and it is usually the sole viable bifidobacteria species in fermented milks (18). Proper strain identification is a very valuable trait for both producers and consumers, as close probiotic strains may have different effects on host health (i.e. at the

immunomodulation level) (19). P4P peptidome-based trees were fairly similar to those generated by single nucleotide polymorphisms (SNP)/insertion-deletion polymorphism-based allelic typing (20). Yet, our method enabled the identification of specific peptides in each of the strains, specifically more than 50 specific peptides per strain, which could be used to construct antibody-based tests and thus, may efficiently detect a defined strain in fermented milks or within the gut microbiota during clinical trials. Input data are provided in Supplementary Material S2 and as an example dataset in the web service.

Case study 3: *Ralstonia solanacearum* species complex

The peptide mass fingerprints of 27 strains of the plant pathogen *Ralstonia solanacearum*, produced using MALDI-TOF-MS, are provided as an example of an experimental mass-to-charge dataset. Note that usually this data consists of [M+H]+monoisotopic lists of masses. These data originated from an experimental study that based on genomic and proteomic evidence supported the division of the *R. solanacearum* species complex into three species (21). Classification of *R. solanacearum* species complex has been matter of controversy during the last 50 years, and a taxonomic review was mandatory in order to better cluster different groups of this microorganism for better optimize applications such as identification of resistance to bacterial wilt, or identification of new pathogenic strains.

P4P analysis based on the proteomic profiles was consistent with the classification obtained in this study using different whole genome based distances, including GBDP, showing against the complementary features that peptidomes can add to genome sequencing. Moreover, that work showed the discriminative potential of peptidome similarity analysis (i.e. identification and comparison of unique peptide profiles) as well as the ability of P4P to work with both *in silico* and experimental proteomic data. Input data are provided in Supplementary Material S3 and as an example dataset in the web service.

DISCUSSION

We have introduced P4P, a novel *in silico* peptidome fingerprinting tool to explore similarities between microbial proteomes and simplify the process of obtaining new biological insights into their phylogeny. P4P is a versatile tool that can help biologists in many ways, for example, improving phylogenetic analyses in challenging taxonomic groups, biomarker identification or epidemiologic studies. Our method is complementary to *in silico* DNA-to-DNA hybridization and, if tuned adequately, has the ability to reduce large peptidome datasets without losing phylogenetic signals (see benchmarking data in Supplementary Materials S4 and 5). Moreover, experimental peptidomes can be obtained from single bacteria cultures to small consortia or even complex populations such as the human gut microbiota.

P4P allows the discrimination of strain-specific peptides, true intra-species differences among biological replicates, or even microorganism-phenotype variations. Indeed, the flexible customization options of P4P can be used to analyze different peptide datasets, i.e. bacteria, viruses, eukaryotic species and metaproteomes. Also, it is able to work with whole proteome datasets as well as experimental mass-to-charge lists originated from mass spectrometers.

Continued efforts are being made to optimize the speed, file size capacity and rendering features of the tool. In future releases of P4P we expect to be adding support for MS/MS data. more functionality with regard to downstream data processing, namely a more powerful tree viewer, support for a larger set of tree data export formats and enable user customization of SPECLUST and MrBayes parameters. We believe that P4P is a valuable time-saving resource that will become an integral part of the day to day work of many research groups worldwide.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

SING group thanks CITI (Centro de Investigación, Transferencia e Innovación) from University of Vigo for hosting its IT infrastructure.

FUNDING

Spanish 'Programa Estatal de Investigación, Desarrollo e Innovación Orientada a los Retos de la Sociedad' [AGL2013-44039R]; Portuguese Foundation for Science and Technology (FCT) under the scope of the strategic funding of UID/BIO/04469/2013 unit and COMPETE 2020 [POCI-01-0145-FEDER-006684]; INOU16-05 project from the University of Vigo; Fundación AECC. Funding for open access charge: Spanish 'Programa Estatal de Investigación, Desarrollo e Innovación Orientada a los Retos de la Sociedad' [AGL2013-44039R].
Conflict of interest statement. None declared.

REFERENCES

1. Rosselló-Móra,R. (2012) Towards a taxonomy of Bacteria and Archaea based on interactive and cumulative data repositories. *Environ. Microbiol.*, **14**, 318–34.
2. Auch,A.F., von Jan,M., Klenk,H.-P. and Göker,M. (2010) Digital DNA-DNA hybridization for microbial species delineation by means of genome-to-genome sequence comparison. *Stand. Genomic Sci.*, **2**, 117–134.
3. Colston,S.M., Fullmer,M.S., Beka,L., Lamy,B., Gogarten,J.P. and Graf,J. (2014) Bioinformatic genome comparisons for taxonomic and phylogenetic assignments using Aeromonas as a test case. *Mbio*, **5**, e02136.
4. Auch,A.F., Klenk,H.-P. and Göker,M. (2010) Standard operating procedure for calculating genome-to-genome distances based on high-scoring segment pairs. *Stand. Genomic Sci.*, **2**, 142–148.
5. Meier-Kolthoff,J.P., Auch,A.F., Klenk,H.-P. and Göker,M. (2013) Genome sequence-based species delimitation with confidence intervals and improved distance functions. *BMC Bioinformatics*, **14**, 60.
6. Alves,G., Wang,G., Ogurtsov,A.Y., Drake,S.K., Gucek,M., Suffredini,A.F., Sacks,D.B. and Yu,Y.-K. (2016) Identification of microorganisms by high resolution tandem mass spectrometry with accurate statistical significance. *J. Am. Soc. Mass Spectrom.*, **27**, 194–210.

7. Singhal,N., Kumar,M., Kanaujia,P.K. and Virdi,J.S. (2015) MALDI-TOF mass spectrometry: an emerging technology for microbial identification and diagnosis. *Front. Microbiol.*, **6**, 791.
8. Karlsson,R., Gonzales-Siles,L., Boulund,F., Svensson-Stadler,L., Skovbjerg,S., Karlsson,A., Davidson,M., Hulth,S., Kristiansson,E. and Moore,E.R.B. (2015) Proteotyping: Proteomic characterization, classification and identification of microorganisms–A prospectus. *Syst. Appl. Microbiol.*, **38**, 246–257.
9. Blanco-Míguez,A., Meier-Kolthoff,J.P., Guitiérrez-Jácome,A., Göker,M., Fdez-Riverola,F., Sánchez,B. and Lourenço,A. (2016) Improving phylogeny reconstruction at the strain level using peptidome datasets. *PLoS Comput. Biol.*, **12**, e1005271.
10. Blanco-Míguez,A., Gutiérrez-Jácome,A., Fdez-Riverola,F., Lourenço,A. and Sánchez,B. (2016) A peptidome-based phylogeny pipeline reveals differential peptides at the strain level within Bifidobacterium animalis subsp. lactis. *Food Microbiol.*, **60**, 137–141.
11. Yu,N.Y., Wagner,J.R., Laird,M.R., Melli,G., Rey,S., Lo,R., Dao,P., Cenk Sahinalp,S., Ester,M., Foster,L.J. *et al.* (2010) PSORTb 3.0: Improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, **26**, 1608–1615.
12. Horlacher,O., Nikitin,F., Alocci,D., Mariethoz,J., Müller,M. and Lisacek,F. (2015) MzJava: an open source library for mass spectrometry data processing. *J. Proteomics*, **129**, 63–70.
13. Johansson,P., Alm,R. and Emanuelsson,C. (2006) SPECLUST: a web tool for clustering of mass spectra. *J. Proteome Res.*, 785–792.
14. Huelsenbeck,J.P. and Ronquist,F. (2001) MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, **17**, 754–755.
15. Maddison,D.R., Swofford,D.L. and Maddison,W.P. (1997) NEXUS: an extensible file format for systematic information. *Syst. Biol.*, **46**, 590–621.
16. Lomonaco,S., Furumoto,E.J., Loquasto,J.R., Morra,P., Grassi,A. and Roberts,R.F. (2015) Development of a rapid SNP-typing assay to differentiate Bifidobacterium animalis ssp. lactis strains used in probiotic-supplemented dairy products. *J. Dairy Sci.*, **98**, 804–812.
17. Gueimonde,M., Delgado,S., Mayo,B., Ruas-Madiedo,P., Margolles,A. and De Los Reyes-Gavilán,C.G. (2004) Viability and diversity of probiotic Lactobacillus and Bifidobacterium populations included in commercial fermented milks. *Food Res. Int.*, **37**, 839–850.
18. Jayamanne,V.S. and Adams,M.R. (2006) Determination of survival, identity and stress resistance of probiotic bifidobacteria in bio-yoghurts. *Lett. Appl. Microbiol.*, **42**, 189–194.
19. Hill,C., Guarner,F., Reid,G., Gibson,G.R., Merenstein,D.J., Pot,B., Morelli,L., Canani,R.B., Flint,H.J., Salminen,S. *et al.* (2014) Expert consensus document: The International Scientific Association for Probiotics and Prebiotics consensus statement on the scope and appropriate use of the term probiotic. *Nat. Rev. Gastroenterol. Hepatol.*, **11**, 506–514.
20. Briczinski,E.P., Loquasto,J.R., Barrangou,R., Dudley,E.G., Roberts,A.M. and Roberts,R.F. (2009) Strain-specific genotyping of Bifidobacterium animalis subsp. lactis by using single-nucleotide polymorphisms, insertions, and deletions. *Appl. Environ. Microbiol.*, **75**, 7501–7508.
21. Prior,P., Ailloud,F., Dalsing,B., Remenant,B., Sanchez,B. and Allen,C. (2016) Genomic and proteomic evidence supporting the division of the plant pathogen Ralstonia solanacearum into three species. *BMC Genomics*, **17**, 90.