Amino Acids (2015) 47:2659–2663
DOI 10.1007/s00726-015-2058-2

SHORT COMMUNICATION

PECAS: prokaryotic and eukaryotic classical analysis of secretome

Ana R. Cortazar[1] · José A. Oguiza[2] · Ana M. Aransay[1] · José L. Lavín[1]

Received: 23 December 2014 / Accepted: 16 July 2015 / Published online: 2 August 2015

Abstract Full sets of proteins that are transported to the extracellular space, called secretomes, have been studied for a variety of organisms to understand their potential role in crucial metabolic pathways and complex health conditions. However, there is a lack of tools for integrative classical analysis of secretomes that consider all the data sources available nowadays. Thus, PECAS (Prokaryotic and Eukaryotic Classical Analysis of Secretome) has been developed to provide a well-established prediction pipeline on secreted proteins for prokaryote and eukaryote eukaryote species.

Keywords Analysis pipeline · Bioinformatics · Protein secretion · Secretome · Prediction · Signal peptide

Handling Editor: P. R. Jungblut.

PECAS is freely available at: http://web.bioinformatics.cicbiogune.es/PECAS/index.php.

Electronic supplementary material The online version of this article (doi:10.1007/s00726-015-2058-2) contains supplementary material, which is available to authorized users.

✉ Ana M. Aransay
amaransay@cicbiogune.es

✉ José L. Lavín
jllavin@cicbiogune.es

[1] Genome Analysis Platform, CIC bioGUNE and CIBERehd, Bizkaia Technology Park, 48160 Derio, Spain

[2] Genetics and Microbiology Research Group, Department of Agrarian Production, Public University of Navarre, 31006 Pamplona, Spain

Introduction

Secretome analysis is a trendy topic in different fields such as cancer biomarkers identification, understanding intercellular signaling in neurobiology, stem cells differentiation, discovery of exosomes roles, thyroid hormone regulation, pathogens infection, as well as in numerous aspects of plant and microbial studies (Brown et al. 2012; Greening and Simpson 2013). A major problem is that most existing tools for studying secreted proteins require basic knowledge of command line programming. As part of a solution, SECRETOOL was specifically designed for the analysis of secretomes in fungi (Cortazar et al. 2014). On the other hand, Next-Generation Sequencing (NGS) is currently the main source of raw data at genomics research, but tools for secretome analyses out of such data are lacking (Caccia et al. 2013). Accordingly, we have developed PECAS (Prokaryotic and Eukaryotic Classical Analysis of Secretome), a freely available web tool that deals with proteins that can be secreted through the so-called classical mechanisms for a wide range of organisms, both prokaryotes and eukaryotes. PECAS admits a variety of input options, allowing users to retain most sensible raw sequence data; being to our knowledge, the first tool designed to perform secretome analysis on data derived from NGS technologies. The classification process is carried out through a well-established pipeline on secreted proteins prediction (Brown et al. 2012; Wymelenberg et al. 2005) at a simple web interface through a single submission step.

Materials and methods

Implementation

PECAS presents an interface based on PHP/CSS that relies on a Perl/CGI control module to enable communication between the user and the tools within the server. The web

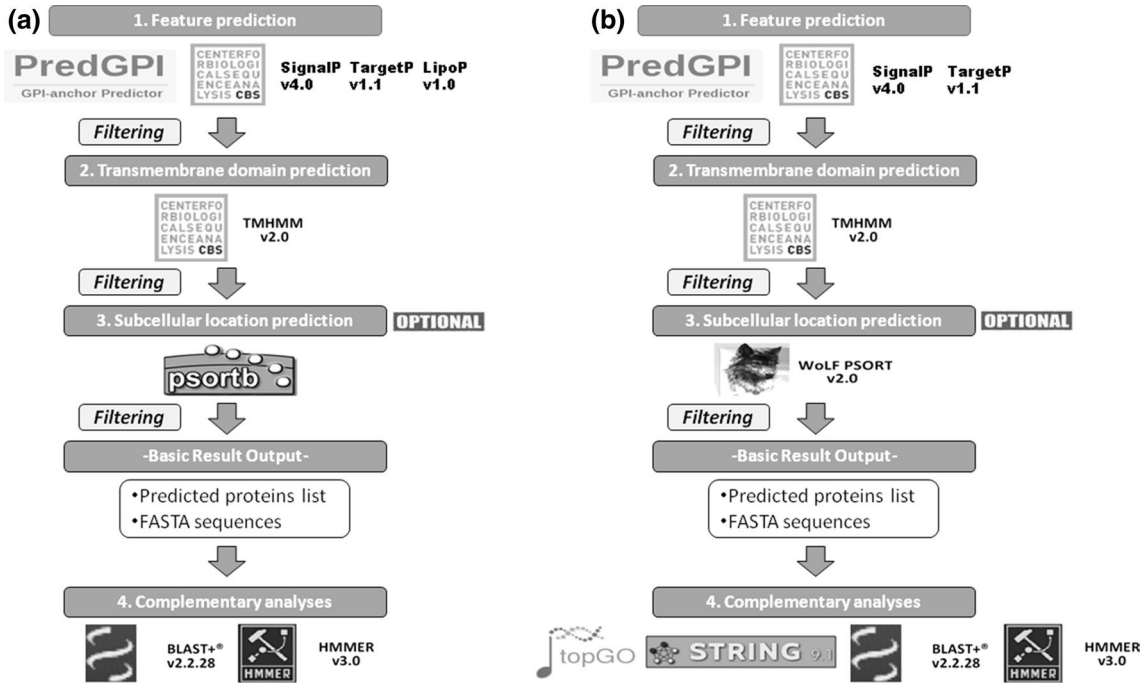

Fig. 1 **a** PECAS's prokaryotic workflow representation and **b** PECAS's eukaryotic workflow representation, showing the main steps of the analysis. Pre-existent bioinformatic tools are represented by their logos and custom developed scripts are displayed as "Filtering" boxes

tool is placed under a Linux environment in an Apache server, to enhance its stability and also to provide a secure background. The prediction pipeline scaffold consists of a set of Perl and R scripts to allow different input formats, to connect individual steps from the analysis pipeline, to filter and process the individual tools' results, and to obtain the corresponding tables and graphical outputs (Fig. 1). Furthermore, a MySQL indexing structure has been implemented to improve predictions retrieval speed for the eligible organisms (i.e., full ENSEMBL list of released vertebrate assemblies) (Supplementary Table S1). All the tools included in PECAS (Supplementary Table S2) are based only upon publicly available software.

Results and discussion

PECAS allows secretome analyses on both prokaryotic and eukaryotic protein sequences considering only the classical protein secretion mechanisms. Accordingly, a different pipeline has been set up for each sort of organisms.

Prokaryotic analysis pipeline

This whole workflow is presented under a one-step query interface, where a FASTA input file of protein sequences can be uploaded. Before submitting data, users can select the cut-off scores for the processing tools and the optional tests. The prokaryotic analysis is executed through a four (two mandatory plus two optional)-stage pipeline (Fig. 1a). The first stage has been designed to carry out secreted protein features prediction, precisely, TargetP identifies the location of the *N*-terminal Signal Peptide (SP), SignalP also recognizes the presence and location of SP cleavage sites (Emanuelsson et al. 2007), PredGPI (Pierleoni et al. 2008) checks the existence of glycosylphosphatidylinositol (GPI) membrane anchoring site and LipoP (Juncker et al. 2003; Rahman et al. 2008) searches for lipoprotein signal peptides for Gram-negative and Gram-positive bacteria. Sequences that enclose any of those features are taken to the second step to be inspected for the presence of transmembrane domains (TMD) by TMHMM (Krogh et al. 2001), where proteins with a number of TMD higher than 1 are discarded. As optional, the third step classifies selected candidates based on their subcellular location by PSORTb (Yu et al. 2010). This sorting is aimed to avoid false positive results from previous steps, but since this is a very stringent filter, user has to be aware that if the selected cut-off value is high, false negative results may arise. At this point, a list of sequence identifiers of the putatively secreted proteins is obtained in FASTA format and PECAS displays a series of graphical results and a table with the features detected for each protein ID considered as secreted. Still complimentary analyses can be chosen in the final stage. There, protein sequences are queried into BLAST+ (Camacho et al. 2009) looking for orthology relations and into HMMER3

scan versus *Pfam* HMM profiles database (Finn et al. 2011, 2014) to determine the domain structure. The results table and graphics along with the orthologs and domain structure determination, if selected, are available for downloading together with the basic results compressed in a unique zip file. Alternatively, user can select to receive a link to the results via e-mail upon job completion.

Eukaryotic analysis pipeline

PECAS's eukaryotic pipeline (Fig. 1b) encompasses different tools for secretome analysis through a one-step query interface, where users can upload input files in one of the following formats: (1) gene identifier lists (IDL) from different repositories (depending on the species), (2) tables of read counts (ToC) out of NGS experiments, (3) differential expression tables (DEt), and (4) protein FASTA files (See PECAS's "Help & FAQS" section for example formats). Cut-off scores for the processing tools can be adjusted and optional and complementary tests might be selected before data submission. Identifiers (IDs) extracted from IDL, ToC, and DEt for all the organisms included within ENSEMBL releases are queried into a MySQL indexing structure to considerably shorten processing times (see comparisons at Supplementary Table S2).

The eukaryotic pipeline is also organized in four stages that are slightly different from those described for the prokaryotic pipeline. Feature prediction is just achieved through SignalP and TargetP in addition to the recognition of glycosylphosphatidylinositol (GPI) membrane anchoring

Table 1 PECAS prediction power by calculation of Sensitivity (Sn) and Specificity (Sp) parameters (numbers in bold)

Sensitivity (Sn)						
Eukarya	#Classically secreted set	SignalP	TargetP	Phobius	WoLFPSORT	PECAS
H. sapiens	1217	**0.33** (396)	**0.61** (741)	**0.63** (767)	**0.56** (676)	**0.59** (717)
R. norvegicus	869	**0.53** (457)	**0.78** (678)	**0.81** (704)	**0.72** (622)	**0.78** (676)
M. musculus	1724	**0.42** (727)	**0.76** (1304)	**0.79** (1356)	**0.68** (1171)	**0.75** (1295)
D. rerio	319	**0.39** (126)	**0.80** (256)	**0.81** (257)	**0.73** (234)	**0.80** (255)
D. melanogaster	324	**0.64** (207)	**0.88** (284)	**0.87** (283)	**0.76** (245)	**0.87** (281)
C. elegans	260	**0.48** (126)	**0.85** (221)	**0.89** (232)	**0.78** (202)	**0.85** (221)
C. albicans	299	**0.38** (113)	**0.85** (254)	**0.87** (259)	**0.64** (192)	**0.85** (253)
Prokarya	#Classically secreted set	SignalP	TargetP	Phobius	PSORTb	PECAS
L. pneumophila	175	**0.08** (14)	**0.35** (61)	**0.43** (76)	**0.19** (33)	**0.35** (61)
E. faecalis	279	**0.04** (11)	**0.80** (222)	**0.74** (206)	**0.08** (22)	**0.75** (208)
L. acidophilus	163	**0.09** (14)	**0.66** (108)	**0.71** (116)	**0.10** (17)	**0.68** (111)
Specificity (Sp)						
Eukarya	#OXPHOS proteins	SignalP	TargetP	Phobius	WoLFPSORT	PECAS
H. sapiens	251	**1** (0)	**0.95** (13)	**0.88** (29)	**0.87** (33)	**0.99** (3)
R. norvegicus	106	**1** (0)	**0.90** (11)	**0.86** (15)	**0.89** (12)	**0.98** (2)
M. musculus	145	**1** (0)	**0.93** (10)	**0.87** (19)	**0.88** (18)	**0.99** (3)
D. rerio	54	**1** (0)	**0.80** (11)	**0.91** (5)	**0.78** (12)	**0.96** (2)
D. melanogaster	66	**1** (0)	**0.83** (11)	**0.91** (6)	**0.88** (8)	**0.98** (1)
C. elegans	71	**1** (0)	**0.86** (10)	**0.96** (3)	**0.90** (7)	**1** (0)
C. albicans	107	**1** (0)	**0.96** (4)	**0.96** (4)	**1** (0)	**0.98** (2)
Prokarya	NS Set	SignalP	TargetP	Phobius	PSORTb	PECAS
L. pneumophila	100	**0.99** (1)	**0.99** (1)	**0.95** (5)	**0.95** (5)	**0.98** (2)
E. faecalis	100	**1** (0)	**1** (0)	**0.98** (2)	**1** (0)	**0.99** (1)
L. acidophilus	100	**1** (0)	**0.99** (1)	**0.96** (4)	**1** (0)	**0.99** (1)

True positive predictions are shown for each tool in the Sn table and false positives in the Sp table (numbers in parenthesis). Test sets for Sn calculation were downloaded from secretome curated databases to obtain a classically secreted set (first column at Sn table), as explained in Supplementary methods. Sp estimation required sets of true non-secreted proteins (first column of Sp table): eukaryotes sequences from the oxidative phosphorylation pathway (OXPHOS) of each organism were downloaded from UniProt (http://www.uniprot.org/) and, in the case of prokaryotes, a random set of 100 protein sequences not predicted as secreted (NS set) were extracted from each prokaryotic proteome for this assessment

sites, while step two is identical to the prokaryotic one. The optional third filtering phase is carried out by WoLFPSORT (Horton et al. 2007), for eukaryotic subcellular location sorting. Again as a final stage, complementary analyses can be selected in order to get the resulting protein sequences queried into BLAST+ for orthology information and scanned versus *Pfam* database, to determine the domain structure of the output proteins. Furthermore, if the query was either IDL, ToC, or DEt, corresponding to one of the so far 66 species considered in our input file menu, the resulting IDL is tested for protein–protein interactions through STRING database (Franceschini et al. 2013) and/or Gene Ontology (GO) enrichment analysis by topGO (Alexa and Rahnenfuhrer 2010). The enrichment examination is accomplished by topGO, which relies on two types of statistics test: (i) Fisher's exact test, which is based on gene counts, that applies when the input is a list of interesting genes (as in the case of IDLs) if a sufficient number of them are available to carry out the required statistical tests and (ii) a Kolmogorov–Smirnov like test that computes enrichment based on gene scores such as those available for DEt or ToC input type.

The basic results are a list with the IDs of the putatively secreted proteins and a file where the sequences corresponding to those IDs are stored in FASTA format. In addition, the results page provides a comprehensive overview of the analysis through dynamic tables and graphics that are also packed in a unique zip file for downloading. The option of receiving a link to the results via e-mail is also available for this pipeline.

Performance testing of PECAS

To test PECAS predictions, we compared its sensibility and specificity performance with several other classical secretion analysis tools (i.e., SignalP, TargetP, PSORTb, WoLFPSORT, and Phobius) based on curated datasets for secreted proteins (See Supplementary Figures S1 and S2).

PECAS's sensitivity ranks among the top three of the tools evaluated, being only an average of 3 % less sensitive than the best one for this parameter, Phobius (Table 1). Concerning specificity, only SignalP outperforms PECAS with an average of 2 % higher specificity (Table 1). Therefore, although our tool is not the best performer for any of the two parameters assessed independently, it is the most balanced showing the highest marks for the combination of both quality controls. This enhancement is achieved through an integrative approach due to the inclusion of multiple tools in the primary prediction phase of the workflow, aiming to complement possible weaknesses of each tool with the strengths of the others, thus improving individual prediction capabilities. Furthermore, results from PECAS main analysis pipeline can be refined by subcellular location through PSORTb or WoLFPSORT as additional filter to improve specificity values by removing false positives, but this filter might involve a significant loss of sensitivity, which may not be of interest depending of the analysis.

Conclusion

The web tool PECAS is targeted to a wide field, allowing the inspection of prokaryotic and eukaryotic proteins that are candidates to be secreted through the classical pathway. This tool is expected to have a significant impact in research because of its flexibility, sensitivity, specificity, and the fact that it runs under an intuitive user interface, thus enabling a complete screening for secreted proteins on datasets derived from NGS technologies or FASTA format files. PECAS depicts workflows that comprise an array of analysis tools focused on identification of potential secreted proteins through a one-step submission interface. This way, researchers may be encouraged to inspect old data from a new perspective or even as a serendipity act, to unveil previously overlooked connections between proteins, which may remain unexplored due to the technical complexity of this kind of studies. Therefore, this web tool will aid to decipher clue questions concerning the secretome consequences in cancer development and hormonal regulation, among others, as emphasized by Greening and Simpson (2013).

Acknowledgments We thank J. Ferrero for his advice with Perl/CGI scripting and I. Lázaro for the webpage's illustrations. ARC, AMA, JLL, and the research expenses are supported by the Basque Country Government (Etortek Research Programs 2011/2014) and by the Innovation Technology Dept. of Bizkaia. JAO was supported by research project AGL2011-30495 of the Spanish National Research Plan and received additional support from the Public University of Navarre.

Compliance with the ethical standards

Conflict of interest The authors declare that they have no conflict of interest.

References

Alexa A, Rahnenfuhrer J (2010) topGO: Enrichment analysis for Gene Ontology. R package version 2.16.0

Brown NA, Antoniw J, Hammond-Kosack KE (2012) The predicted secretome of the plant pathogenic fungus *Fusarium graminearum*: a refined comparative analysis. PLoS One 7(4):e33731

Caccia D, Dugo M, Callari M, Bongarzone I (2013) Bioinformatics tools for secretome analysis. Biochim Biophys Acta 1834(11):2442–2453

Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, Madden TL (2009) BLAST+: architecture and applications. BMC Bioinformatics 10:421

Cortazar AR, Aransay AM, Alfaro M, Oguiza JA, Lavín JL (2014) SECRETOOL: integrated secretome analysis tool for fungi. Amino Acids 46(2):471–473

Emanuelsson O, Brunak S, von Heijne G, Nielsen H (2007) Locating proteins in the cell using TargetP, SignalP and related tools. Nat Protocols 2:953–971

Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A (2014) Hetherington In: Holm K, Mistry L, Sonnhammer J, Tate EL, Punta L. Pfam: the protein families database Nucleic Acids Res. 42(Database issue):D222–30

Finn RD, Clements J, Eddy SR (2011) HMMER web server: interactive sequence similarity searching. Nucleic Acids Res 39:29–37

Franceschini A, Szklarczyk D, Frankild S, Kuhn M, Simonovic M, Roth A, Lin J, Minguez P, Bork P, von Mering C, Jensen LJ (2013) STRING v9.1: protein-protein interaction networks, with increased coverage and integration. Nucleic Acids Res 41:D808–D815

Greening DW, Simpson RJ (2013) (Eds). An updated secretome (Special issue). Biochim Biophys Acta.1834 (11)

Horton P, Park KJ, Obayashi T, Fujita N, Harada H, Adams-Collier CJ, Nakai K (2007) WoLF-PSORT: protein localization predictor. Nucleic Acids Res 35:585–587

Juncker AS, Willenbrock H, Von Heijne G, Brunak S, Nielsen H, Krogh A (2003) Prediction of lipoprotein signal peptides in Gram-negative bacteria. Protein Sci 12(8):1652–1662

Krogh A, Larsson B, von Heijne G, Sonnhammer EL (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J Mol Biol 305(3):567–580

Pierleoni A, Martelli PL, Casadio R (2008) PredGPI: a GPI-anchor predictor. BMC Bioinformatics 9:392

Rahman O, Cummings SP, Harrington DJ, Sutcliffe IC (2008) Methods for the bioinformatic identification of bacterial lipoproteins encoded in the genomes of Gram-positive bacteria. World J Microb Biot 24(11):2377–2382

Wymelenberg AV, Sabat G, Martinez D, Rajangam AS, Teeri TT, Gaskell J, Kersten PJ, Cullen D (2005) The Phanerochaete chrysosporium secretome: database predictions and initial mass spectrometry peptide identifications in cellulose-grown medium. J Biotechnol 118(1):17–34

Yu NY, Wagner JR, Laird MR, Melli G, Rey S, Lo R, Dao P, Sahinalp SC, Ester M, Foster LJ, Brinkman FSL (2010) PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. Bioinformatics 26:1608–1615