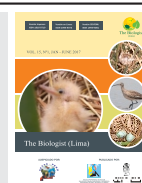




The Biologist (Lima)



ORIGINAL ARTICLE / ARTÍCULO ORIGINAL

NEW PHYLOGENETIC ANALYSIS METHOD USING TAXPLOT, AND ITS APPLICATION ON ACUTE DIARRHEAL DISEASE-CAUSING BACTERIA

NUEVO MÉTODO DE ANÁLISIS FILOGENÉTICO USANDO TAXPLOT, Y SU APLICACIÓN EN BACTERIAS CAUSANTES DE ENFERMEDAD DIARREICA AGUDA

Roberto Adolfo Ubidia-Incio^{1,3} & Jeel Jr. Moya-Salazar^{1,2,*}

¹Facultad de Ciencias y Filosofía, Universidad Peruana Cayetano Heredia, Av. Honorio Delgado 430, Lima 31, Lima, Perú. ²Hospital Nacional Docente Madre Niño San Bartolomé, Av. Alfonso Ugarte 825, Lima 01, Lima, Perú.

³Universidad Nacional Federico Villarreal, Facultad de Ciencias Naturales y Matemáticas. Jr. Río Chepén 290, Lima 10, Perú

*E-mail: jeel.moya.s@upch.pe / roberto.ubidia.i@upch.pe

ABSTRACT

Proteins are important in phylogenetics because they show conservation, not only in their sequence, but also in their functions and the properties of their different substructures due to residue or amino acid family conservation. The analysis of the set of data from a great number of proteins will show us properties that have been the driving force of species diversification. Here, we present a phylogenetic method for constructing cladograms using TaxPlot data. We constructed a cladogram using the UPGMA method using the total hits obtained from all the combinations of TaxPlot previously transformed into percentages of similarity. This cladogram based on the similarity of bacterial proteomes presents a large similarity with the relations that exist for the analyzed organisms based on conserved sequences.

Keywords: database – genome-wide analysis – phylogenetics – protein

RESUMEN

Las proteínas son importantes en la filogenética dado que muestran la conservación, no solo de sus secuencias, sino también de sus funciones y de las propiedades de sus diferentes subestructuras debido a la conservación de las familias de aminoácidos. El análisis de grupos de datos de un gran número de proteínas puede mostrar sus propiedades que han tenido fuerza de manejo de la diversificación de especies. Aquí, presentamos un nuevo método filogenético para la construcción de cladogramas usando la base de datos TaxPlot. Hemos construido un cladograma utilizando el método UPGMA de acuerdo con los aciertos (Hits) totales obtenidos a partir de todas las combinaciones de TaxPlot previamente transformados en porcentajes de similitud. Este cladograma basado en la similitud de los proteomas bacterianos presenta una gran similitud con las relaciones que existen para los organismos analizados en base a secuencias conservadas.

Palabras clave: análisis genómico – base de datos – filogenética – proteína

INTRODUCTION

Genome analysis is a useful tool because it allows us to understand diverse biologic phenomena by showing the conservation of the genes responsible of specific functions, not only one isolate gene but a group of them with a function that has been conserved during the course of evolution (Roman *et al.*, 2000). Comparing genomes of many related organisms allows also seeing and understanding how the emergence of new genes can help to develop new functions inside a pre-established gene system. That is why identifying orthologs is critical to predict protein functions on new strains or species that cause a determined pathology (Roman *et al.*, 2000; Tatusov *et al.*, 2003). Without doubt, the later would benefit the understanding of the pathogenic mechanisms employed by many organisms such as Acute Diarrhea Disease causing bacteria, the way they appeared and diversified. Moreover, in a greater measure, the utility when looking for strategies against them and preventing their infection (Tatusov *et al.*, 2003).

Comparative genomics offers a useful scope for understanding information at a genomic level, for this, the first objective when a genome is sequenced, is to identify the functional regions, both genes and regulatory sequences. A part of that information needs experimental work, but another part could be obtained from comparative analysis *in silico*, which can facilitate the identification of those elements (Tatusov *et al.*, 2003). Thus, the purpose of comparative genomics is, through the comparison of genomes of different species, to understand how those species have evolved, and the functions of their genes and non-coding regions (Thorat & Thakare, 2013).

The bioinformatics tool, TaxPlot, that belongs to NCBI (National Center for Biotechnology Information), allows us to observe the similitude among three organisms based on the proteome of one of them, it refers to the amount of similar proteins from the genomes of three different species, using the first one as a query, from which every protein sequence is taken and compared with pre-computed BLAST results (<http://blast.ncbi.nlm.nih.gov/Blast.cgi>) of the other two species (Wheeler *et al.*, 2001). This type of analysis has been previously used to study

similarities between other organisms, such as archaeas (Makarova *et al.*, 2015). In other cases, these tools have been used to predict genes from recently sequenced genomes, for example on *Serratia plymuthica* AS-12 (Roman *et al.*, 2000, Neupane *et al.*, 2012), *Staphylococcus epidermidis* (Thorat & Thakare, 2013), and others (Heidelberg *et al.*, 2000, Bhagwat & Bhagwat, 2008; Siddaramappa, 2007). This tool allows us to analyze genomes where a comparison has been previously done, so it is not needed to download complete genomes in order to do an analysis. Nevertheless, it has the disadvantage of not being continually maintained, so the number of available genomes is limited and not updated.

Acute Diarrhea Disease is one of the four principal causes of child mortality in the world and one of three that has been related to climate change (Murga *et al.*, 1993; MINSA, 2015). In the Americas, it has been estimated that 112 000 annual deaths are caused by diarrheal diseases (WHO, 1999). This affection causes alterations on the human development in three fundamental aspects: limiting the organic growth, reducing the intellectual capital, and increasing the Disability-Adjusted Life Years (DALY), mainly in children under 2 years.

The principal causes of Acute Diarrhea Disease are the bacteria from the genus *Campylobacter*, *Salmonella*, *Shigella*, *Vibrio*, *Proteus* and *Enterococcus faecalis*; and also *Escherichia coli* O157:H7, *Klebsiella pneumoniae*, *Morganella morgani*, *Aeromonas oxytona*, among others (Cama *et al.*, 1999; WHO, 1999; Seas *et al.*, 2000; Winn *et al.*, 2006).

The objective of the present investigation was to design a novel phylogenetic method using the data from Diarrhea causing disease-bacteria listed in the TaxPlot tool, and to analyze their genomes and establish a proteomic-scale relationship between them.

MATERIALS AND METHODS

The application of this method was done through an analytic, retrospective, cross-sectional study

applied to acute diarrhea disease-causing bacterias.

Species: The species considered for testing this method: *Campylobacter jejuni subsp jejuni* (NCTC 11168=ATCC700819) [Cje], *Shigella dysenteriae* (strain Sd197) [Sdy], *Salmonella enterica subsp enterica* (serovar Typhimurium strain LT2) [Sen], *Vibrio cholerae* (01biovar El Tor strain N16961) [Vch], *E. coli* O157:H7 (strain Sakai) [Eco], *Proteus mirabilis* (strain H14320) [Pmi], *E. faecalis* (strain V583) [Efa], *K. pneumoniae* (strain 342) [Kpn] (Winn *et al.*, 2006; Cama *et al.*, 1999).

Resources for genome analysis: An inter-protein comparative analysis of each species was done with the tool TaxPlot from NCBI (<http://www.ncbi.nlm.nih.gov/sutils/taxik2.cgi?isbact=1>) with the objective of observing the degree of similitude between orthologs proteins of the analyzed species. The relation between each species (“query” species) and the others in groups of two (comparative species) was observed.

Techniques for collecting and processing data: The number of total hits was used; this means, the number of proteins that produced hits for the two comparative species based on the total number of proteins of the query species without taking into account the bias in the similitude towards any of the comparative species. All this initial data was tabulated in a comparative matrix with the variables [Cje], [Sdy], [Sen], [Vch], [Eco], [Pmi], [Efa] and [Kpn], where the relationship between the degrees of similitude was analyzed.

Eight organisms were included; we used TaxPlot as a tool to observe the similitude between the selected bacteria based on their proteome instead of the classical reach in which conserved regions are evaluated. The plots generated by this tool, provide us with total hits data, that are hits obtained from comparing all the proteins from a query organism against the protein from other two species selected, through pre-calculated BLAST scores, those hits were distributed depending on the degree of similarity to one or the other species.

Taxonomy: A phylogenetic classification of the proteins codified in the complete genomes was done through the creation of cladograms, using the UPGMA method (unweighted paired group means analysis) (Kuske *et al.*, 2002), using the data from

each query species, joining the two species with a higher percentage of hits, then calculating the mean of the values of the two joint species against all the other, and repeating this process until every species had been joined (Fig. 1). Finally, a consensus cladogram was created based on the eight cladograms previously created.

A cladogram based on rRNA 16S was made using the software Phylip version 3,696 (University of Washington, <http://evolution.genetics.washington.edu/phylip.html>) and visualized with TreeView version 1,6.6 (University of Glasgow, <http://taxonomy.zoology.gla.ac.uk/rod/treeview.html>) in order to compare the method based on conserved sequences alignment and the method presented here (Fig. 2).

Data Analysis: Data analysis was performed in four basic processes: Codification, tabulation, table construction, and preparation of graphics and cladograms.

Limitations: In principle, the presentation of analyzed species does not include the totality of existent enteropathogens because TaxPlot does not include all of them on its database. Second, there is no bioinformatics tool available for designing cladograms with data from TaxPlot. Lastly, function patterns were not evaluated for enterobacteria. In spite of these limitations, our research contributes to the bioinformatics development applied to clinic microbiology.

RESULTS

With the help of the bioinformatics tool TaxPlot of NCBI we could access to similitude data of all the proteome of the Acute Diarrhea Disease-causing bacteria analyzed in groups of three. We obtained 168 plots where it was possible to observe the similitude between each species selected as query towards other two from the group of analyzed bacteria. From this data, previously transformed into percentages, obtained from the total hits for each comparison, a cladogram was created (Fig. 2). The methodology for elaborating this cladogram is described in the flowchart presented in Figure 1.

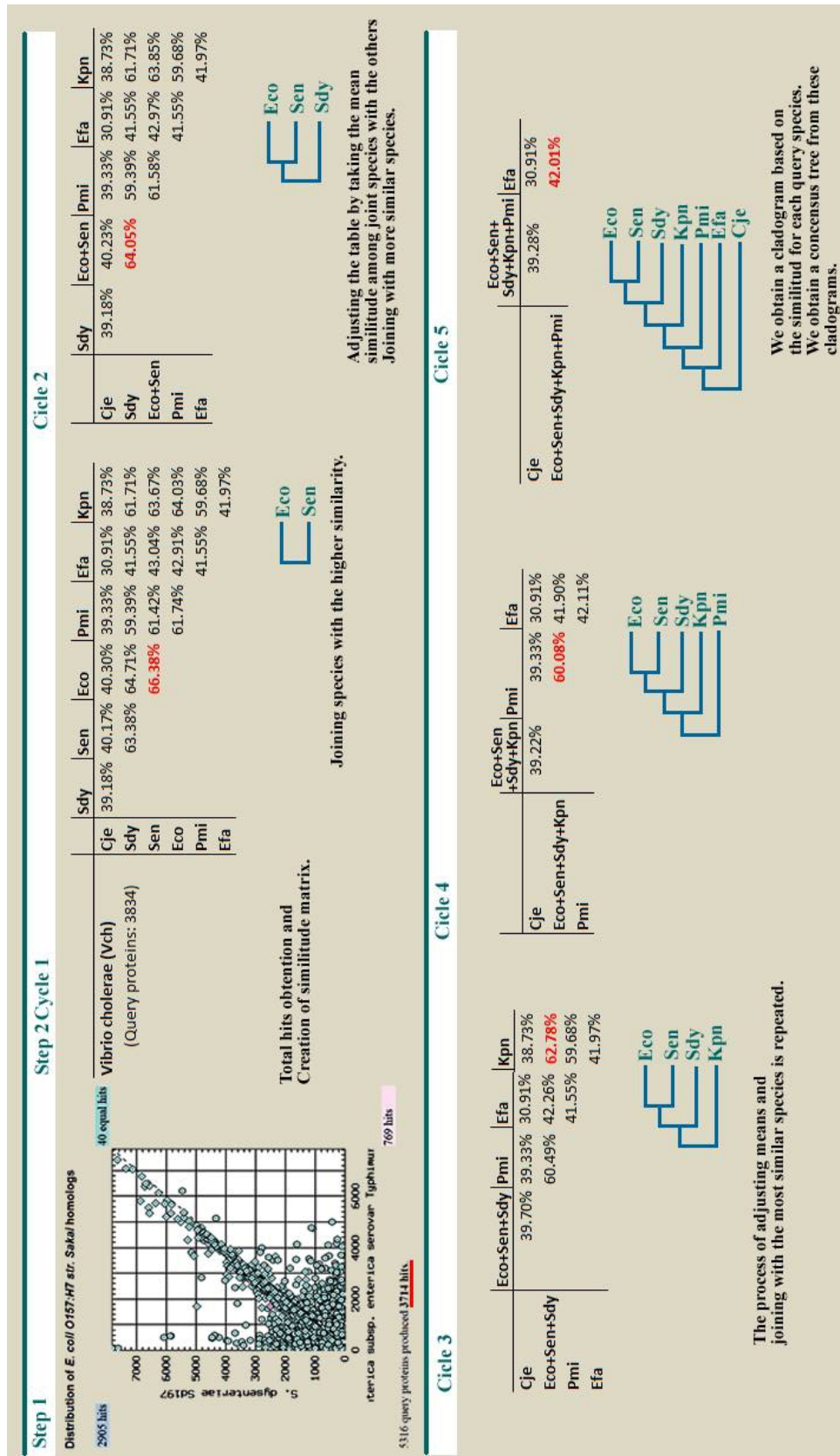


Figure 1. Flowchart for the development of the cladogram based on TaxPlot data for the ADD-causing species.

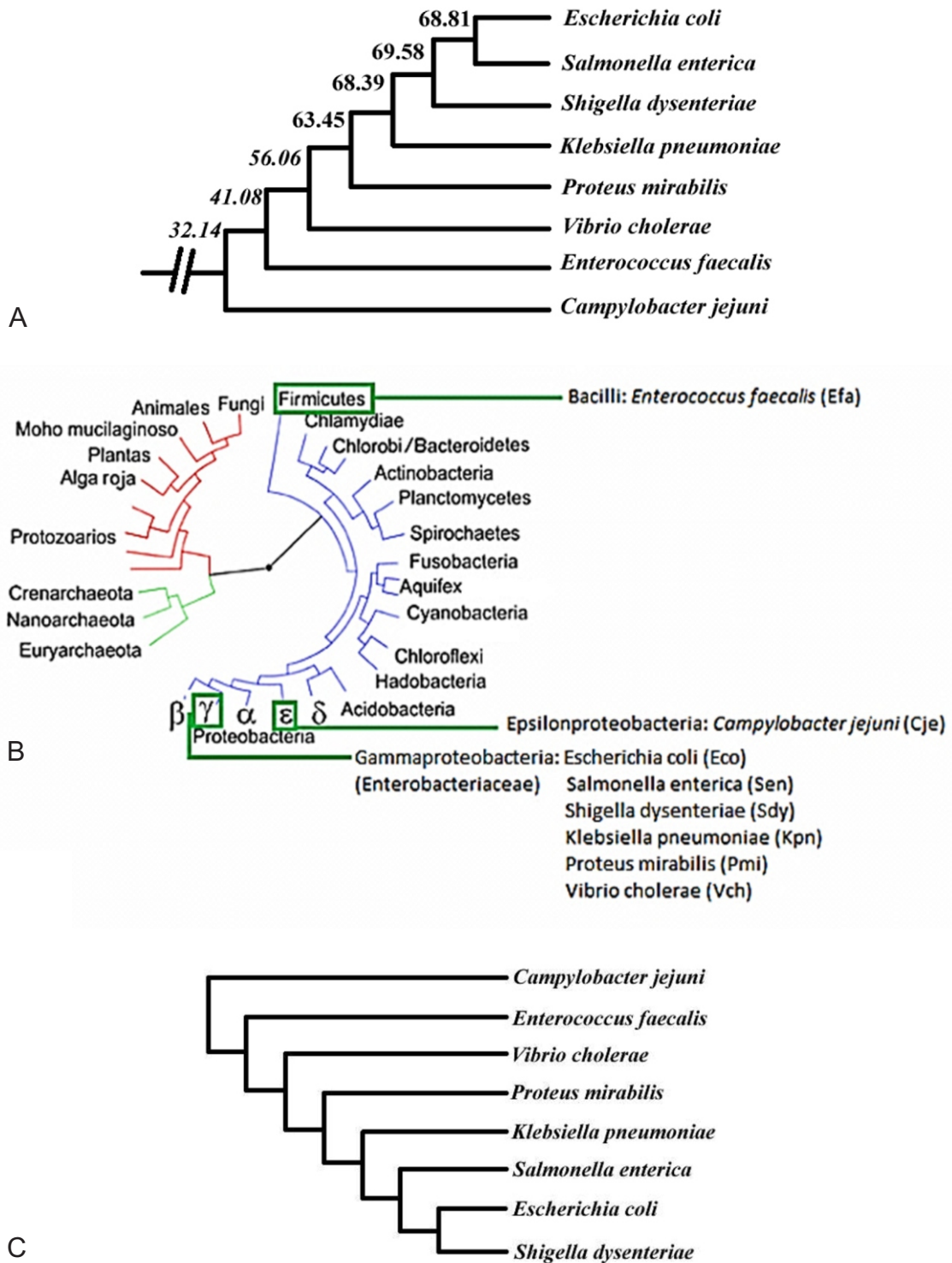


Figure 2. Comparative cladograms. **A.** Consensus tree obtained with the UPGMA method with the total hits data obtained with the bioinformatics tool TaxPlot (numbers indicate the percentage of siilitude). **B.** Phylogenetic tree where the studied groups have been highlighted in green frames, showing the evolutionary divergence [modified from Ciccarelli *et al.* (2006), **C.** Cladogram based on rRNA 16S sequences from the analyzed species using the software Phylip and TreeView.

The cladogram based on the similitude of the proteome of the studied bacteria presents a great similitude with the relations found in these organisms using conserved sequences (Fig. 2) (Moran *et al.*, 2005; Neupane *et al.*, 2012; Moya-Salazar & Ubidia-Incio, 2015). Additionally, it shows the relation that exists between our species based on their proteome, showing as more related species those that have a higher number of similar proteins.

DISCUSSION

The cladogram based in proteome similitude presents a great similitude with the relation that exists for these organisms based on conserved sequences (Fig. 2) (Montiel, 1997; Moran *et al.*, 2005; Chieh-Hua *et al.*, 2011; Neupane *et al.*, 2012). When comparing cladograms (TaxPlot, alignment based on protein universal families, and rRNA 16S), in general, the analyzed bacteria form very similar groups. *Campylobacter jejuni* showed a plesiomorphic character.

It is noticeable the separation between Enterobacteriaceae (*S. enterica*, *E. coli* O157:H7, *S. dysenteriae*, *P. mirabilis* and *K. pneumoniae*) and Vibrionaceae (*V. cholerae*). Also, among this whole group (Gammaproteobacteria), Epsilonproteobacteria (*C. jejuni*) and Firmicutes (*E. faecalis*), where the similitude among enterobacteria ranges between 63,45% and 68,81% with the method based on TaxPlot, nevertheless, when incorporating bacteria from other clades this similitude diminishes drastically, to 56,06% when incorporating *V. cholerae*, 41,08% for *E. faecalis* and 32,14% for *C. jejuni*.

The only exception to the similitude of results between the method based on TaxPlot and the method based on the rRNA 16S sequence is the relation for *S. dysenteriae*, *S. enterica* and *E. coli* O157:H7, where we have *E. coli* and *S. enterica* more related among them than with *S. dysenteriae*. The usual order is *Escherichia coli* more related to *S. dysenteriae* rather than *S. enterica* (Fig. 2).

This method seems to have as limitation the difference in the number of proteins of the query species. *C. jejuni* has the lowest number (1623), in

comparison with over 3000 proteins in the other bacteria, thus the low number of coincidences makes it appear as the farthest, while methods based on protein universal families (Chieh-Hua *et al.*, 2011) shows it is near the other proteobacteria. Being *E. faecalis* the farthest as it belongs to the phylum Firmicutes (Ciccarelli *et al.*, 2006). Curiously, when aligning and obtaining a cladogram from the sequences of rRNA 16S, we obtain the same order between these two bacteria with respect to the result obtained comparing the proteome (TaxPlot, Fig. 2).

The method we present here represents a holistic approximation to the study of phylogenetic relationships between close species that could be contrasted against other more specific approximations in order to observe the evolutionary behavior in organisms that share a common environment or niche. In the case of bacteria, it is known that there have been many events of horizontal gene transfer (HGT) making it more complex the identification of the evolutionary history in these groups as an orthologous gene may be more similar for two organisms that are not necessarily the closest, evolutionarily talking (Akortha & Filgona, 2009; Stecher *et al.*, 2012). Making a comparison using the whole proteome this effect can be diminished in part.

This method reduces the time needed for phylogenetic analysis due to the help provided by the existence of tools as TaxPlot that stores precalculated data, and establishes a phylogenetic reconstruction at a proteomic level; these features work well as no distant organisms are chosen that will diminish the power of this method. These elements of judgment can be transported as exact data employing bioinformatics data that could, as in this case, describe the features of the pathogenic agents who produce a disease.

ACKNOWLEDGMENTS

The Authors are grateful for the contribution of the researchers from the Laboratorios de Investigación y Desarrollo (LID) of Universidad Peruana Cayetano Heredia for their advice, their time, and inculcate in us the vocation for bioinformatics.

BIBLIOGRAPHICS REFERENCES

- Akortha, E.E. & Filgona, J. 2009. Transfer of gentamicin resistance genes among enterobacteriaceae isolated from the outpatients with urinary tract infections attending 3 hospitals in Mubi, Adamawa State. *Scientific Research and Essays*, 4:745-752.
- Bhagwat, A.A.; Bhagwat, M. 2008. Methods and tools for comparative genomics of foodborne pathogens. *Foodborne Pathogens and Disease*, 5:487-497.
- Cama, R.I.; Parashar, U.D.; Taylor, D.N.; Hickey, T.; Figueroa, D.; Ortega, Y.R.; Romero, S.; Perez, J.; Sterling, C.R.; Gentsch, J.R.; Gilman, R.H. & Glass, R.I. 1999. Enteropathogens and other factors associated with severe disease in children with acute watery diarrhea in Lima, Peru. *Journal of Infectious Diseases*, 179: 1139-1144.
- Chieh-Hua, L.; Chun-Yi, L.; Chao, A.H. & Feng-Chi, C. 2011. Changes in transcriptional orientation are associated with increases in evolutionary rates of enterobacterial genes. *BMC Bioinformatics*, 12(Suppl 9): S19.
- Ciccarelli, F.D.; Doerks, T.; von Mering, C.; Creevey, C.J.; Snel, B. & Bork, P. 2006. Toward automatic reconstruction of a highly resolved tree of life. *Science*, 311: 1283-1287.
- Heidelberg, J.F.; Eisen, J.A.; Nelson, W.C.; Clayton, R.A.; Gwinn, M.L.; Dodson, R.J.; Haft, D.H.; Hickey, E.K.; Peterson, J.D.; Umayam, L.; Gill, S.R.; Nelson, K.E.; Read, T.D.; Tettelin, H.; Richardson, D.; Ermolaeva, M.D.; Vamathevan, J.; Bass, S.; Qin, H.; Dragoi, I.; Sellers, P.; McDonald, L.; Utterback, T.; Fleishmann, R.D.; Nierman, W.C.; White, O.; Salzberg, S.L.; Smith, H.O.; Colwell, R.R.; Mekalanos, J.J.; Venter, J.C. & Fraser, C.M. 2000. *DNA sequence of both chromosomes of the cholera pathogen Vibrio cholera*. *Nature*, 406:477-483.
- Kuske, R.C.; Ticknor, L.O.; Miller, M.E.; Dunbar, M.J.; Davis, A.J.; Barns, M.S. & Belnap, J. 2002. Comparison of soil bacterial communities in rhizospheres of three plant species and the interspaces in arid grassland. *Applied and Environmental Microbiology*, 8:1854-1863.
- Makarova, S.K.; Wolf, I.Y. & Koonin, V.E. 2015. Archaeal Clusters of Orthologous Genes (arCOGs): An Update and application for analysis of shared features between Thermococcales, Methanococcales and Methanobacteriales. *Life*, 5: 818-840.
- MINSA. 2015. *Plan de Comunicaciones, Prevención de Enfermedades Diarreicas Agudas (EDA) y Cólera 2014, versión preliminar*. Ministerio de Salud del Perú. Extraído el 07 de Enero del 2015. Desde http://www.minsa.gob.pe/portada/Especial/es/2014/lavadomanos/archivo/Plan_de_comunicaciones-prevencion_de_enfermedades_diarreicas_y_colera.pdf
- Montiel, A.F. 1997. Flora bacteriana habitual. *Boletín Escuela de Medicina U.C. Pontificia Universidad Católica de Chile*, 26:133-139.
- Moran, A.N.; Russell, J.A.; Koga, R. & Fukatsu, T. 2005. Evolutionary Relationships of three new species of Enterobacteriaceae living as symbionts of aphids and other insects. *Applied and Environmental Microbiology*, 71:3302-3310.
- Moya-Salazar, J. & Ubidia-Incio, R. 2015. Genome-scale analysis of protein functions and evolution from acute diarrheal disease-causing bacteria using COG database and Tax Plot. *Revista Latinoamericana de Patología Clínica y Medicina de Laboratorio*, 62:206-219.
- Murga, H.; Huicho, L. & Guevara, G. 1993. Acute diarrhea and *Campylobacter* in Peruvian children: a clinical and epidemiologic approach. *Journal of Tropical Pediatrics*, 39:338-341.
- Neupane, S.; Finlay, R.D.; Alström, S.; Goodwin, L.; Kyrpides, N.C; Lucas, S.; Lapidus, A.; Bruce, D.; Pitluck, S.; Peters, L.; Ovchinnikova, G.; Chertkov, O.; Han, J.; Han, C.; Tapia, R.; Detter, J.C.; Land, M.; Hauser, L.; Cheng, J.F.; Ivanova, N.; Pagani, I.; Klenk, H.P.; Woyke, T. & Högberg, N. 2012. Complete genome sequence of *Serratia plymuthica* strain AS12. *Standards in Genomic Sciences*, 6:165-173.
- Roman, L.T.; Michael, Y.G.; Darren, A.N. & Eugene, V.K. 2000. The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids*

- Research, 28:33–36.
- Seas, C.; Alarcón, M.; Aragón, J.C.; Beneit, S.; Quiñonez, M.; Guerra, H. & Gotuzzo, E. 2000. Surveillance of bacterial pathogens associated with acute diarrhea in Lima. *International Journal of Infectious Diseases*, 4:96-99.
- Siddaramappa, S.S. 2007. *Comparative and Functional Genomic Studies of Histophilussomni (Haemophilus somnus)* [Engineering Thesis]. Blacksburg, Virginia Polytechnic Institute and State University.
- Stecher, B.; Denzler, R.; Maier, L.; Berneta, F.; Sandersc, M.J.; Pickard, D.J.; Barthel, M.; Westendorf, A.M.; Krogfelt, K.A.; Walker, A.W.; Ackermann, M.; Dobrindt, U.; Thomson, N.R. & Hardt, W.D. 2012. Gut inflammation can boost horizontal gene transfer between pathogenic and commensal Enterobacteriaceae. *Proceedings of the National Academy of Sciences*, 109: 1269–1274.
- Thorat, S.S. & Thakare, V.P. 2013. Analysis of Staphilococcus using comparative genomics. *International Journal of Scientific and Engineering Research*, 4:1775-1779.
- Tatusov, R.L.; Fedorova, N.D.; Jackson, J.D.; Jacobs, A.R.; Kiryutin, B.; Koonin, E.V.; Krylov, D.M.; Mazumder, R.; Mekhedov, S.L.; Nikolskaya, A.N.; Rao, B.S.; Smirnov, S.; Sverdlov, A.V.; Vasudevan, S.; Wolf, Y.I.; Yin, J.J. & Natale, D.A. 2003. The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, 4:41.
- Wheeler, D.; Benson, D. & Rapp, B. 2001. *Using TaxPlot to compare Genomes*. Bethesda, National Center for Biotechnology Information News, 1-2.
- WHO. 1999. *The World Health Report: making a difference*. World Health Organization, Geneva.
- Winn, W.; Allen, S.; Janda, W.; Koneman, E.; Procop, G.; Schreckenberger, P.C. & Woods, G. 2006. *Koneman's Color Atlas and Textbook of Diagnostic Microbiology*, 6th ed. Philadelphia, Lippincott Williams & Wilkins.

Received September 6, 2016.
Accepted January 19, 2017.