PROTEINS
STRUCTURE ■ FUNCTION ■ BIOINFORMATICS

The lifestyle of prokaryotic organisms influences the repertoire of promiscuous enzymes

Mario Alberto Martínez-Núñez,[1,2]* Katya Rodríguez-Vázquez,[1] and Ernesto Pérez-Rueda[2,3]

[1] Departamento de Ingeniería de Sistemas Computacionales y Automatización, Instituto de Investigaciones en Matemáticas Aplicadas y en Sistemas, Universidad Nacional Autónoma de Mexico, Ciudad Universitaria, México, D.F., México

[2] Departamento de Ingeniería Celular y Biocatálisis, Instituto de Biotecnología, Universidad Nacional Autónoma de México, Cuernavaca, Morelos, México

[3] Unidad Multidisciplinaria de Docencia e Investigación, Sisal Facultad de Ciencias, Sisal, Yucatán, UNAM, México

ABSTRACT

The metabolism of microbial organisms and its diversity are partly the result of an adaptation process to the characteristics of the environments that they inhabit. In this work, we analyze the influence of lifestyle on the content of promiscuous enzymes in 761 nonredundant bacterial and archaeal genomes. Promiscuous enzymes were defined as those proteins whose catalytic activities are defined by two or more different Enzyme Commission (E.C.) numbers. The genomes analyzed were categorized into four lifestyles for their exhaustive comparisons: free-living, extremophiles, pathogens, and intracellular. From these analyses we found that free-living organisms have larger genomes and an enrichment of promiscuous enzymes. In contrast, intracellular organisms showed smaller genomes and the lesser proportion of promiscuous enzymes. On the basis of our data, we show that the proportion of promiscuous enzymes in an organism is mainly influenced by the lifestyle, where fluctuating environments promote its emergence. Finally, we evidenced that duplication processes occur preferentially in metabolism of free-living and extremophiles species.

Proteins 2015; 00:000–000.

Key words: metabolism; bacteria; archaea; comparative genomics; duplicated enzymes.

INTRODUCTION

The bacterial adaptation to environment fluctuations depends on the cellular adequacy to variables such as temperature, nutrients availability, and salinity, which are in constant change. In this context, diverse cellular components such as transporters, regulatory proteins, and catalytic enzymes converge to contend and adapt against these changes. The role that enzymes play in these processes is crucial because they achieve the biochemical transformations of substrates into useful products, providing the cell with matter and energy necessary to grow. Enzymes have usually been described by catalyzing only one reaction on specific substrates, thus they are commonly considered as "specialists"; however, some enzymes may also be multifunctional.[1,2] Multifunctional enzymes have been defined as proteins playing multiple physiological roles in the cell, and can be classified as moonlighting and promiscuous. Moonlighting enzymes are typically composed by a structural domain that performs the catalytic activity, and a noncatalytic domain, associated to regulation or protein–protein interactions, among others; in counterpart, promiscuous enzymes are only composed of catalytic domains performing several biochemical functions.[3–6] In this context, diverse mechanisms that allow enzymatic promiscuity, such as conformational diversity, alternative substrates positions, and different protonation states, among others have been described.[7] From a genomic perspective, our group has recently reported that around 10% of the total enzymatic repertoire in bacterial and archaeal organisms corresponds to promiscuous enzymes.[8]

Additional Supporting Information may be found in the online version of this article.

Grant sponsor: DGAPA, UNAM; Grant number: IN-204714; Grant sponsor: CONACYT; Grant number: 155116; Grant sponsor: Cátedras CONACyT.

*Correspondence to: Mario Alberto Martínez-Núñez, Centro de Investigación en Biotecnología Aplicada del Instituto Politécnico Nacional (CIBA-IPN), Exhacienda San Juan Molino, Tepetitla de Lardizábal, Tlaxcala, México.
E-mail: maal.martinez@gmail.com

Received 23 March 2015; Revised 1 June 2015; Accepted 21 June 2015
Published online 25 June 2014 in Wiley Online Library (wileyonlinelibrary.com).
DOI: 10.1002/prot.24847

In this work, we evaluated the impact of the environment in shaping and maintaining the prokaryotic enzymatic repertoire, and in particular the one involving the promiscuous enzymes. In this regard, we analyze whether the content of promiscuous enzymes in prokaryotic organisms is influenced by their lifestyle or by the genome size. To achieve this question, an exhaustive analysis of promiscuous enzymes content was accomplished in 761 nonredundant organisms classified into free-living, extremophiles, pathogens, and intracellular groups. On the basis of statistical tests we identified that the distribution of promiscuous enzymes exhibits different patterns according to the lifestyles of organisms. Our results show that free-living organisms are the group with the highest enrichment of promiscuous enzymes, while intracellular and pathogens organisms exhibit the lowest content. In addition, another important result that emerges from our analysis is the lack of correlation between the content of promiscuous enzymes and genome size. Finally, we found that gene duplication processes are more frequent in promiscuous enzymes of free-living microorganisms than in the other lifestyles.

MATERIALS AND METHODS

Genomes and proteomes analyzed

The 761 complete sequencing prokaryotic genomes used in our analysis, including 89 archaeal and 672 bacterial genomes, were downloaded from NCBI ftp server public database section genomes (ftp://ncbi.nlm.nih.gov). These genomes were defined as nonredundant genomes, to exclude any bias associated with overrepresentation of species or strains, as has been previously reported.[8] We considered only genes with open reading frames (ORFs) that encode predicted protein sequences (Supporting Information Table S1).

Identification of enzymes and promiscuous enzymes

For each enzyme we identified the annotation of an Enzyme Commission (E.C.) number using the KEGG database.[9] Then, for each enzyme associated with an E.C. number the presence of both functional and structural domains at sequence level, based on Pfam[10] and Superfamily[11] assignments, respectively, was defined. Our inclusion criterion was quite strict because we were only interested in enzymatic sequences with identified metabolic contexts and functional assignments. Finally, promiscuous enzymes were identified in this set of filtered data as those sequences containing two or more different E.C. numbers. All processes and data management were conducted using *ad hoc* Perl scripts (see Supporting Information Table S2).

Identification of paralogous sequences

Paralogs were defined as protein-coding sequences within a fully sequenced genome with sequence identity $\geq$30%, coverage $\geq$60%, and an *E*-value cutoff of $10e^{-05}$, as similar criteria described by Pushker *et al.*[12] Therefore, for each single proteome, a BlastP[13] all-against-all search was performed, selecting sequences that fulfilled the criteria described above. Once duplicated sequences in each genome were identified, we crosschecked the information with the list of enzymatic sequences of each organism to identify those enzymes that came from duplication events. All sequence analyses were conducted using *ad hoc* Perl scripts.

Performance evaluation

To assess the accuracy of the promiscuous enzymes defined and analyzed in this work, the Carbonell and Faulon's approach was considered and used through their enzyme promiscuity prediction server (http://www.issb.genopole.fr/~faulon/promis.php). At first, we selected a representative sample of enzymes by using the formula $(N^{*}Z^{2*}p^{*}q)/((NE^{2}) + (Z^{2*}p^{*}q))$, with a margin of error of 4.9% and a confidence level of 95% (N = population size, $Z = Z$ score, E = standard error; p and q are the probability to be promiscuous or not, and were assumed to be 0.5). Therefore, we compared our sample (400 of 51,572 promiscuous sequences), and a sample of the dataset kindly provided by Carbonell and Faulon (400 of 19,411 promiscuous enzymes), using the Promis server. In addition, a sample of 400 of 48,846 nonpromiscuous enzymes (negative control), also provided by Carbonell and Faulon, were also evaluated with the Promis server.

This comparison was useful to calculate the following values: (1) true positives (TP): promiscuous enzymes with more than two different E.C. numbers, and functional (PFAM) and structural domains (SUPFAM) (our dataset) and identified by the Promis server, and Carbonell and Faulon's positive set identified by the Promis server; (2) false positives (FP): proteins identified as promiscuous enzymes with the Promis server from the Carbonell and Faulon's negative dataset; (3) false negatives (FN): promiscuous enzymes (our dataset) and Carbonell and Faulon's positive dataset, identified as negative with the Promis server; (4) sensitivity, Sn = TP/(TP + FN), is the fraction of promiscuous enzymes identified with the Promis server; (5) positive predictive value, PPV = TP/(TP + FP), is the fraction of the promiscuous enzymes inferred; (6) Accuracy, Ac = (Sn + PPV)/2, is the PPV and Sn average.

Statistical analysis

Nonparametric Kruskal–Wallis and Wilcoxon rank-sum statistical tests, as well as Spearman's test were used to evaluate promiscuous enzymes distributions in each

Table I
Comparison of Enzyme Content of Organisms According to Their Lifestyles

	Free-living	Pathogens	Extremophiles	Intracellular
Number of organisms	367	188	158	48
Average number of genes	3723	2875	2644	1531
Bacteria percentage	91.0	99.5	65.8	98.0
Archaea percentage	8.9	0.5	34.2	2.0
Average number of enzymes	828	691	629	396
Average ratio of enzymes (enzymes/ORFs)	23.5	25.7	24.5	31.5
Average ratio of enzymes (enzymes/ORFs) in equivalents sets	23.5	25.7	24.5	31.5
Average number of promiscuous enzymes	74	59	55	31
Average proportion of promiscuous enzymes in observed groups (promiscuous/enzymes)	9.05	8.30	8.55	7.9
Average proportion of promiscuous enzymes in equivalent groups of 48 elements each one (promiscuous/enzymes)	9.0	8.4	8.5	7.9

lifestyle. Statistical significance was set at $P \leq 0.05$. The implementation of these tests was carried out using the package stats of R programming language for statistical analysis.[14] Data managements were conducted with *ad hoc* Perl scripts.

RESULTS AND DISCUSSION

Distribution of genome sizes importantly varies between lifestyles

A total of 761 nonredundant organisms were classified into four categories according to their lifestyle[15]: free-living organisms with 367 genomes, including nonpathogenic bacteria; pathogens with 188 genomes, which include organisms reported to produce a disease in plants or animals; extremophiles with 158 organisms; and intracellular with 48 organisms, which include obligate intracellular pathogens (Table I and Supporting Information Table S1). The lifestyle annotation was based on information provided in the corresponding literature deposited in NCBI database[16] as well as in BacMap Genome Atlas.[17]

Based on this classification, the analysis and comparison of genome sizes represented in the four lifestyles were accomplished, where the genome size was calculated as the number of ORFs reported by each organism. The average genome size by each lifestyle obtained was 3723 ORFs for free-living organisms, 2875 for pathogens, 2644 for extremophiles, and 1531 for intracellular organisms (Table I and Supporting Information Fig. S1). To determine whether there is a significant difference in the genome size distribution between the four lifestyles a Kruskal–Wallis test was applied, resulting in a difference in the distributions of genome sizes between lifestyles (P values $< 2.2e^{-16}$). In a subsequent step, a paired Wilcoxon test to evaluate individual differences among lifestyles showed that extremophiles and pathogens are the only two classes with similar distributions of genome sizes (P-value $= 0.91$), while the other categories are statistically different from each (P-value $< 1.03e^{-08}$). The organisms classified as intracellular exhibited the shortest genome size, which is consistent with the notion that these organisms have suffered a reduction in their genomes as a mechanism of adaptation to the environment of their host, from whom it receives nutrients from cytoplasm or tissues without the need to synthesize for themselves.[18] In counterpart, the group of free-living organisms was the group with larger genome sizes.

To exclude a bias as a consequence of overrepresentation of sequenced genomes in the observed results or an uneven sampling of genomes with different size ranges, we repeated the analysis considering equivalent subsets of 48 organisms for pathogens, extremophiles, and free-living lifestyles. This number of organisms was selected, because it corresponds to the smallest group of genomes included in the intracellular lifestyle. The process of randomly selecting 48 genomes per lifestyle class was performed 10,000 times each, obtaining the average of each one, and their statistical differences were evaluated. From these analyses, the only difference between equivalent and observed groups was found between pathogens and extremophiles. Wilcoxon's test shows that the distributions of genome sizes in these two equivalent groups are different, while in the observed groups was equal, with larger genomes in pathogens in comparison to extremophiles. This behavior is probably due to a bias in pathogenic organisms, where there are species that also exhibit a free-living stage, such as the bacterium *Bacillus licheniformis*, which is a human pathogen and that exhibits a spore stage in soil[19–21]; or the bacterium *Mycobacterium marinum*, which is a pathogen of fishes and humans and also found in swimming pools, beaches, rivers, and lakes.[19,22] The results of the analysis applied to the equivalent groups confirmed that the intracellular organisms are those with smaller genomes sizes, while free-living organisms are those with larger genomes. In this respect, the increase in genome size is not accompanied by a corresponding increase in the number of metabolic genes, but in the number of regulatory genes, such as transcription factors (TFs).[8]

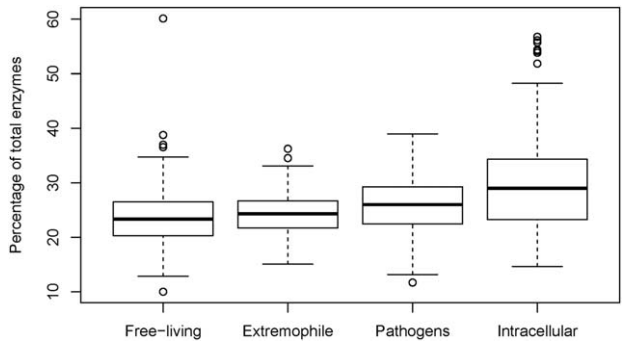

Figure 1

Normalized values of total enzymes by lifestyles. In *x*-axis 761 organisms are classified into four lifestyle categories, *y*-axis represents the proportion of enzymes. The distributions of enzyme fractions are different from each other (Wilcoxon test, *P*-value ≤ 0.01).

The enzymatic content in bacterial and archaeal organisms is influenced by the genome size

To carry out the comparison of content of enzymes by genomes between the different lifestyles, the absolute frequency of enzymes in each organism was obtained and normalized, that is, the ratio between the absolute frequency of enzymes and the genome size. The average of normalized values of the enzymes showed that intracellular organisms are those with the highest percentage (31.56%) of their genome devoted to metabolic functions, followed by pathogens (25.79%), extremophiles (24.51%), and free-living organisms with 23.55% (Fig. 1 and Table I). A Kruskal–Wallis test showed that distributions of normalized values of enzymes are different between lifestyles (*P*-value $= 1.42e^{-8}$). In addition, a paired Wilcoxon test showed that all distributions of the normalized values of enzyme content are different between lifestyles, with a *P*-value ≤ 0.01.

The results reveal that intracellular organisms are those with the highest enzymatic fraction and free-living organisms have the lower proportion of enzymes. Previous analysis concerning genome fraction dedicated to enzymes in bacteria and archaea showed a decrease in the enzymatic content as the genome size increases, that is, a higher proportion of enzymes in genomes containing fewer than 3000 genes than genomes with >6000 genes.[8] Therefore, the above result suggests that free-living organisms exhibit the lowest enzyme content because in this category organisms with larger genomes sizes are present; while in the intracellular category organisms with smaller genomes sizes and higher proportion of enzymes are included. On the basis of previous data we suggest that fluctuating environmental conditions as free-living lifestyle does not favor the increase in the enzyme content, but that it is associated with the genome size, as previously reported. Similar results were obtained when an equivalent dataset was considered, that is, distributions of normalized values of enzymes are different between lifestyles, with a *P*-value $< 2.2e^{-16}$ in the Kruskal–Wallis test, and all distributions of the normalized values of enzymes are different among them, with a *P*-value ≤ 0.003 in the paired Wilcoxon test. These results reinforce the idea that intracellular organisms have a greater percentage of enzyme-encoding genes, while free-living organisms have the lower proportion of their genes devoted to enzymes (Table I).

Free-living organisms exhibit a high enrichment of promiscuous enzymes

To assess whether the fraction of promiscuous enzymes is higher in free-living organisms than in the other lifestyles, an exhaustive analysis concerning these enzymes was conducted. We defined promiscuous enzymes as those proteins with two or more different E.C. numbers, based on the KEGG database. Our results show that free-living organisms have the highest number of promiscuous enzymes with an average of 74 promiscuous enzymes by organism, followed by pathogenic, extremophiles, and intracellular organisms with an average of 59, 55, and 31, respectively (Table I). The ratio between the absolute frequency of promiscuous enzymes and the total number of enzymes in each organism, that is, the normalized values, showed that the group of free-living organisms exhibits the highest proportion of promiscuous enzymes (9.0%), followed by extremophiles (8.5%), pathogens (8.3%), and intracellular organisms (7.9%) (Fig. 2 and Table I). Kruskal–Wallis test shows that distributions of promiscuous enzymes between lifestyles are not equal (*P*-values = 0.001).

Therefore, to determine which lifestyle presents a significant difference, a paired Wilcoxon's test with the normalized values was conducted. This test suggests that the distribution of promiscuous enzymes in free-living

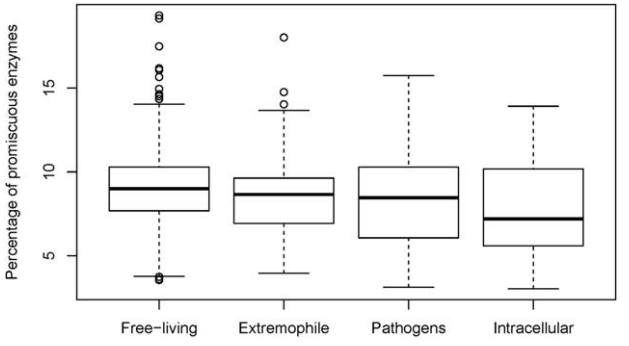

Figure 2

Normalized values of promiscuous enzymes. In *x*-axis organisms are grouped by lifestyle, *y*-axis corresponds to the proportion of promiscuous enzymes. The distributions of enzyme fraction are different from each other (Wilcoxon test, *P*-value ≤ 0.01).

organisms is different from the other three lifestyles (P-value ≤ 0.01), suggesting an enrichment of these multifunctional enzymes in this category. The analysis of the equivalent subsets, as was described previously, shows similar results to those observed for free-living organisms, reinforcing the notion that this lifestyle has the highest proportion of promiscuous enzymes, while results for extremophiles, pathogens, and intracellular organisms were different than those observed in the original groups. Wilcoxon's test shows differences in promiscuous enzymes content among these three equivalent datasets (P-value ≤ 0.006) (Table I). In summary, free-living organisms have the highest content of promiscuous enzymes, while intracellular organisms contain the fewest content. The enrichment of promiscuous enzymes detected in our analyses in free-living organisms may be consequence of an adaptation mechanism to survive in fluctuating ecological environments. In this regard, the presence of a large proportion of promiscuous enzymes would allow the establishment of internal metabolic fluxes that can vary depending on environmental conditions, coupled with lower regulating promiscuous enzymes that enable rapid reprogramming of metabolic response, that is, promiscuous enzymes would be subject to less metabolic regulation than other specialist enzymes.[2,8] Additionally, promiscuous enzymes might endow the organisms with a selective advantage and genome plasticity,[23] which can help to contend against fluctuating ecological niches, such as those faced by free-living microorganisms.

The proportion of promiscuous enzymes is influenced by lifestyle

To analyze whether there is a correlation between promiscuous enzymes content and genome size, similar to the correlation that was identified between total enzymes and genome sizes,[8] a Spearman's test was performed with the normalized values obtained for promiscuous enzymes. The Spearman's coefficient obtained was low, with a value of 0.01 (P-value $= 0.70$), indicating that there is not clear correlation between proportion of promiscuous enzymes and genome size. The normalized values showed that free-living organisms have the high content of promiscuous enzymes among all lifestyles; while intracellular organisms exhibit the less content (Table I), as was discussed in the previous section. The fact that the promiscuous enzymes content is not correlated to genome size leads us to suggest that the proportion of promiscuous enzymes is determined by lifestyle. In this context, lifestyles with fluctuating environments may promote the enrichment of promiscuous enzymes in the organisms regardless of genome size, as a mechanism of ecological adaptation to the environment.

In previous reports, when the promiscuous enzymes content was analyzed from the perspective of taxonomic position, that is, considering the microorganisms according to their classification in bacteria and archaea cellular domains using the NCBI database (taxonomy section), archaea displayed the least promiscuous enzymes content, whereas bacteria exhibited the high content.[3,8] However, the taxonomic composition analysis performed in our lifestyles resulted in that archaeal organisms were found in greater proportion (34%) in extremophile class, which is the group with the second highest promiscuous enzymes content (Table I). These data suggest that the taxonomical classification does not determine the fraction of promiscuous enzymes in archaea, but rather the lifestyle. Because of underrepresentation of archaeal genomes in our data, our interpretation should be considered as preliminary and must be further corroborated. In this regard, recent studies have revealed that archaea are globally widespread in decidedly nonextreme environments.[24,25] This may explain why extremophiles are the second lifestyle with the greatest amount of promiscuous enzymes. The documentation that extremophilic organisms are widely distributed in a variety of environments, which exhibit fluctuations in some environmental variables, comes to strengthen the idea that the enrichment of promiscuous enzymes in this lifestyle is a mechanism that arises to adaptation to changing habitats.

Duplication processes are common in free-living and extremophiles organisms

Gene duplication has been described as an important source of raw material for the generation of new functions, and associated with promiscuous enzymes.[23] In this regard, to evaluate duplication processes that may occur in the promiscuous enzymes, we applied a Blastp "all-against-all" search to identify paralogous genes in all the genomes and subsequently they were compared in terms of their lifestyles. First, we normalized the number of duplicated enzymes as the ratio between the absolute frequencies of duplicated enzymes and the total number of enzymes per organism. The average value of duplicated enzymes was 33.6, 30.1, 24.5, and 12.8% for free-living, extremophiles, pathogens, and intracellular organisms, respectively. The proportions of duplicated enzymes are significantly different among lifestyles (Kruskal–Wallis test, P-value $< 2.2e^{-16}$), and a paired Wilcoxon's test resulted in differences between all lifestyles (P-value < 0.05). These results agree with those previously reported for the duplicated enzymes, where a positive correlation with the increment of genome size has been described.[8] Therefore, the free-living organisms with larger genomes also have a greater fraction of duplicated enzymes, while intracellular organisms have a lower proportion of duplicated enzymes as well as the smallest genomes. To analyze duplicated promiscuous enzymes, we calculated the ratio between the absolute frequencies

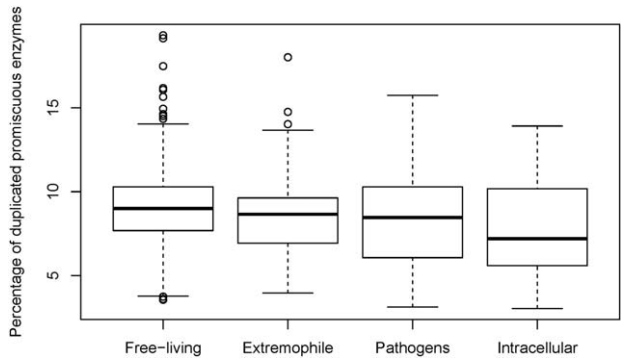

Figure 3

Percentage of duplicated promiscuous enzymes by lifestyle. In *x*-axis organisms are grouped by lifestyle, *y*-axis represents the proportion of duplicated promiscuous enzymes. See text for details.

of duplicated and the total number of promiscuous enzymes per organism. From this, we found an average percentage of 36, 33.7, 25, and 12% for free-living, extremophiles, pathogens, and intracellular, respectively (Fig. 3). Then, we evaluated whether there are significant differences in duplicated promiscuous enzymes content among the four lifestyles. The distributions of normalized values of duplicated promiscuous enzymes are statistically different among lifestyles (Kruskal–Wallis test, *P*-value $< 2.2e^{-16}$). A paired Wilcoxon test was conducted to evaluate which lifestyles have significant differences, identifying that the content of duplicated promiscuous enzymes is different between all lifestyles, with the order previously described (*P*-value ≤ 0.03). We also found that the occurrence of promiscuous duplicated enzymes strongly correlated with genome size, with a Spearman coefficient of 0.77 (*P*-value $< 2.2e^{-16}$), as occurs with total duplicated enzymes. Therefore, there are more promiscuous duplicated enzymes in organisms with large genomes and a small presence in organisms with low number of ORFs, as stated above.

In this regard, gene duplication process was found in high frequency in free-living and extremophile organisms, wherein the percentage of promiscuous enzymes that arose by duplication in the genome is >30%. In contrast, in intracellular organisms only 12% of promiscuous enzymes have arisen by duplication, around one-third of that observed in free-living and extremophiles lifestyles. From these data, we elucidate that duplication processes provide to free-living and extremophile organisms raw material to improve the functions of proteins to contend and obtain nutrients from changing environments, through mutations and then functional diversification of paralogs.[23] In contrast, in more stable environments, such as intracellular habitats, the gene duplication processes appear to be less significant because these organisms tend to loss gene content as a consequence of their lifestyles. Indeed, it has been described that intracellular organisms have evolved through relaxed selection for many bacterial functions, an underlying mutational deletion bias, restricted rates of horizontal gene transfer inside host cell, deleterious and neutral deletions that will accumulate over time in a ratchet-like manner and result in smaller genomes.[26] The emergence and persistence of paralogs in lifestyles where there are fluctuating environments promotes the improvement and innovation of proteins that can be used in nutrient uptake of different compounds, which may be at low concentrations or in rare forms of assimilation. Finally, the diversification of functions that provide duplication processes promotes environmental adaptation of microorganisms.

CONCLUSIONS

Through the analysis of 761 bacterial species grouped into four different lifestyles, we gained insights concerning how the environment influences the metabolic repertoire of bacteria and archaea. Although our criterion for the identification of enzymes was stringent, because it considers the association of each protein sequence to a functional (Pfam) or structural (Superfamily) domain, the resulting dataset analyzed allowed us to obtain general trends that reflect the metabolic context of microorganisms. In this regard, we found an accuracy of 63% of promiscuous enzymes in our data according to the Promis webserver[27] in contrast to the control datasets described in the same server (69%), suggesting that our definition of promiscuous enzymes is in general enough informative. We observed that different evolutionary forces act on bacterial and archaeal metabolism; on one hand the abundance of total enzymes depends on genome size, correlated positively, while on the other hand, the abundance of promiscuous enzymes is influenced by the lifestyle. In general, between 7 and 9% of bacterial and archaeal metabolism consist of promiscuous enzymes, and these are enriched in free-living organisms, perhaps as an adaptive mechanism, which is favored in species living in fluctuating environments. In addition, we found that duplication processes occur more frequently in organisms inhabiting fluctuating environments, where a third of their promiscuous enzymes have arisen by duplication events. In contrast, organisms inhabiting more stable environments such as intracellular species have a lower proportion of duplicated enzymes and promiscuous enzymes. To sum up, we show that the environment favors the appearance of promiscuous enzymes in species inhabiting fluctuating environments, as well as favors duplication processes that allow functional divergence in enzymes.

ACKNOWLEDGMENTS

The authors thank Enrique Merino and Anny Rodriguez Fersaca for their critical reading of the manuscript. They kindly thank Pablo Carbonnell and Jean-Loup

Faulon for the positive and negative datasets of promiscuous enzymes. They also thank the anonymous reviewers for their helpful insights to improve this work.

REFERENCES

1. Jia B, Cheong GW and Zhang S. Multifunctional enzymes in archaea: promiscuity and moonlight. Extremophiles 2013;17:193–203.
2. Nam H, Lewis NE, Lerman JA, Lee DH, Chang RL, Kim D and Palsson BO. Network context and selection in the evolution to enzyme specificity. Science 2011;337:1101–1104.
3. Cheng XY, Huang WJ, Hu SC, Zhang HL, Wang H, Zhang JX, Lin HH, Chen YZ, Zou Q and Ji ZL. A global characterization and identification of multifunctional enzymes. PLoS One 2012;7:e38979.
4. Huberts DH, van der Klei IJ. Moonlighting proteins: an intriguing mode of multitasking. Biochim Biophys Acta 2010;1803:520–525.
5. Hult K, Berglund P. Enzyme promiscuity: mechanism and applications. Trends Biotechnol 2007;25:231–238.
6. Sengupta S, Ghosh S and Nagaraja V. Moonlighting function of glutamate racemase from *Mycobacterium tuberculosis*: racemization and DNA gyrase inhibition are two independent activities of the enzyme. Microbiology 2008;154:2796–2803.
7. Khersonsky O, Tawfik DS. Enzyme promiscuity: a mechanistic and evolutionary perspective. Annu Rev Biochem 2010;79:471–505.
8. Martínez-Núñez MA, Poot-Hernandez AC, Rodriguez-Vazquez K and Perez-Rueda E. Increments and duplication events of enzymes and transcription factors influence metabolic and regulatory diversity in prokaryotes. PLoS One 2013;8:e69707.
9. Kanehisa M. The KEGG database. Nova Found Symp 2002;247:91–101; discussion 101-103, 119-128, 244-252.
10. Finn RD, Bateman A, Clements J, Coggill P, Eberhardt RY, Eddy SR, Heger A, Hetherington K, Holm L, Mistry J, Sonnhammer EL, Tate J and Punta M. Pfam: the protein families database. Nucleic Acids Res 2014;42:D222–D230.
11. Wilson D, Pethica R, Zhou Y, Talbot C, Vogel C, Madera M, Chothia C and Gough J. SUPERFAMILY–sophisticated comparative genomics, data mining, visualization and phylogeny. Nucleic Acids Res 2009;37:D380–D386.
12. Pushker R, Mira A and Rodriguez-Valera F. Comparative genomics of gene-family size in closely related bacteria. Genome Biol 2004;5:R27.
13. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W and Lipman DJ. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 1997;25:3389–3402.
14. R-Programming Development Core Team. R: a language and environment for statistical computing. Vienna, Austria: R Foundation for Statistical Computing; 2011.
15. Cases I, de Lorenzo V and Ouzounis CA. Transcription regulation and environmental adaptation in bacteria. Trends Microbiol 2003;11:248–253.
16. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Federhen S, Feolo M, Fingerman IM, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T, Maglott DR, Marchler-Bauer A, Miller V, Mizrachi I, Ostell J, Panchenko A, Phan L, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Slotta D, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Wang Y, Wilbur WJ, Yaschenko E and Ye J. Database resources of the National Center for Biotechnology Information. Nucleic Acids Res 2011;39:D38–D51.
17. Stothard P, Van Domselaar G, Shrivastava S, Guo A, O'Neill B, Cruz J, Ellison M and Wishart DS. BacMap: an interactive picture atlas of annotated bacterial genomes. Nucleic Acids Res 2005;33:D317–D320.
18. Moran NA. Microbial minimalism: genome reduction in bacterial pathogens. Cell 2002;108:583–586.
19. Babamahmoodi F, Babamahmoodi A and Nikkhahan B. Review of *Mycobacterium marinum* infection reported from Iran and report of three new cases with sporotrichoid presentation. Iran Red Crescent Med J 2014;16:e10120.
20. Haydushka IA, Markova N, Kirina V and Atanassova M. Recurrent sepsis due to *Bacillus licheniformis*. J Glob Infect Dis 2012;4:82–83.
21. Lovdal IS, From C, Madslien EH, Romundset KC, Klufterud E, Rosnes JT and Granum PE. Role of the gerA operon in L-alanine germination of *Bacillus licheniformis* spores. BMC Microbiol 2012;12:34.
22. Slany M, Jezek P, Fiserova V, Bodnarova M, Stork J, Havelkova M, Kalat F and Pavlik I. *Mycobacterium marinum* infections in humans and tracing of its possible environmental sources. Can J Microbiol 2012;58:39–44.
23. Aharoni A, Gaidukov L, Khersonsky O, McQ GS, Roodveldt C and Tawfik DS. The 'evolvability' of promiscuous protein functions. Nat Genet 2005;37:73–76.
24. Aller JY, Kemp PF. Are Archaea inherently less diverse than bacteria in the same environments? FEMS Microbiol Ecol 2008;65:74–87.
25. Eckburg PB, Lepp PW and Relman DA. Archaea and their potential role in human disease. Infect Immun 2003;71:591–596.
26. Koskiniemi S, Sun S, Berg OG and Andersson DI. Selection-driven gene loss in bacteria. PLoS Genet 2012;8:e1002787.
27. Carbonell P, Faulon JL. Molecular signatures-based prediction of enzyme promiscuity. Bioinformatics 2010;26:2012–2019.