Poster A-16

Conservation of genomic neighbourhood is related to the rate by which intraparalogous proteins evolve in two bacterial species

Authors:
Vasilis J. Promponas (*University of Cyprus, Cyprus*)
Nikolas Papanikolaou (*University of Crete, Greece*)
Ioannis Iliopoulos (*University of Crete, Greece*)

Short Abstract: We have estimated evolutionary rates of proteins encoded by paralogous genes in the complete genome of Salmonella typhimurium LT2 considering gene order conservation. Here we show that linked paralogs have significantly higher sequence identities (84.5%) compared to their Escherichia coli K12 homologous proteins than non-linked paralogs (46.3%).

Long Abstract:

Recent studies have demonstrated that proteins encoded by genes with conserved order, often termed as linked genes, have similar rates of evolution [1]. Furthermore, genomic neighbourhood has been associated with physical interactions [2] and common expression patterns or functional relations [3-4]. Other researchers [5] have shown that in the case of intraparalogs (i.e. paralogous genes encoded in the same genome) gene order conservation may be successfully used as a means to identify paralogous operons in bacterial genomes. We have gone the other way around, trying to answer the question whether those intraparalogs that have maintained their local neighbourhood during the course of evolution evolve faster or not. We define as linked genes those genes sharing at least one neighbour downstream or upstream in the genome (not interrupted by interference of other genes), using as an indicator the relative position of the gene under consideration. Formally, given two genomes A and B, encoding Na and Nb genes respectively, and assuming these genes are ordered a(1), a(2), …, a(Na) and b(1), b(2), …, b(Nb) respectively, two genes a(i), and b(j) (1≤i≤Na, 1≤j≤Nb) are defined as linked when the following two conditions are met:

1.Genes a(i), b(j) exhibit statistically significant sequence similarity

2.At least one of a(i-1), a(i+1) exhibits statistically significant sequence similarity with one of b(j-1), b(j+1).

Along these lines we have chosen to study putative paralogs encoded in the complete genome of Salmonella typhimurium LT2 [6] (obtained from NCBI ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Salmonella_typhimurium_LT2/NC_003197.faa) with respect to their corresponding homologs in a reference bacterial genome. As a reference we have used the genome of Escherichia coli K12 [7] (obtained from NCBI ftp://ftp.ncbi.nih.gov/genomes/Bacteria/Escherichia_coliK12/NC_000913.faa), a gamma proteobacterion relatively close to S. typhimurium, with extensive annotation.

In order to identify putative paralogs in the S. typhimurium LT2 genome we have performed an all-against-all self comparison using BLASTP [8] with a strict e-value threshold 1e-10 (all other options having their default values). Applying a milder cut-off of 1e-6 the results were

essentially identical (data not shown). We have employed additional filtering criteria in order to avoid false positives due to matches in short conserved sequence domains. Specifically, to assign two proteins as paralogous we have required that the best local alignment detected by BLASTP provided at least a 50% coverage of the complete length of both sequences.

Out of the 4425 protein coding genes in the S. typhimurium LT2 genome 1841 (41.6%) have been assigned as paralogs to at least one gene in the same genome following the rules previously described. For this subset of proteins we have determined putative homologs in the E. coli K12 genome, based on statistically significant sequence similarities detected by an all-against-all BLASTP comparison, taking the simple best-hit approach [9] using an e-value threshold of 1e-10.

With the E. coli K12 genome as a reference we have identified a homolog for 1216 S. typhimurium LT2 proteins. The remaining 625 proteins exhibit similarities to E. coli proteins that fall below the threshold and were subsequently excluded from our study. Of the remaining 1216 cases, and without taking the strand into account, 1009 were identified as linked and 207 as non-linked paralogs in S. typhimurium LT2 genome.

For these two classes of paralogs we have calculated the average percent of identities reported in the BLASTP output, as a measure of sequence conservation. For the 1009 pairs of proteins with conserved order the average sequence identity was 84.5% (standard deviation 11.5%), whereas for the 207 non-linked genes the corresponding protein homologs shared sequence identity of 46.3% (standard deviation 22.2%) significantly lower than the one calculated for linked genes (Mann-Whitney W = 186828.5, p-value < 2.2e-16). When alternative more sophisticated divergence metrics were applied, such as Poisson corrected distances, the results were essentially the same.

Our analysis clearly demonstrates that gene order conservation is related to sequence conservation. Paralogs within a genome that have conserved order tend to evolve significantly slower compared to ones that are not linked.

Acknowledgements

VJP wishes to thank University of Cyprus for financial support.

References

1.Williams, E.J. and Hurst, L.D. 2000. The proteins of linked genes evolve at similar rates. Nature 407: 900-903.

2.Dandekar, T., Snel, B., Huynen, M.A. and Bork, P. 1998. Conservation of gene order: a fingerprint of proteins that physically interact. Trends Biochem. Sci. 23: 324-328.

3.Cohen, B.A., Mitra, R.D., Hughes, J. and Church, G.M. 2000. A computational analysis of whole-genome expression data reveals chromosomal domains of gene expression. Nature Genet. 26: 183-186.

4.Lercher M.J., Urrutia A.O. and Hurst, L.D. 2002. Clustering of housekeeping genes provides a unified model of gene order in the human genome. Nature Genet. 31: 180-183.

5.Janga SC and Moreno-Hagelsieb G. 2004. Conservation of adjacency as evidence of paralogous operons. Nucleic Acids Res. 32: 5392-5397.

6.McClelland M, Sanderson KE, Spieth J, Clifton SW, Latreille P, Courtney L, Porwollik S, Ali J, Dante M, Du F, Hou S, Layman D, Leonard S, Nguyen C, Scott K, Holmes A, Grewal N, Mulvaney E, Ryan E, Sun H, Florea L, Miller W, Stoneking T, Nhan M, Waterston R and Wilson RK. 2001. Complete genome sequence of Salmonella enterica serovar Typhimurium LT2. Nature. 413:852-856.

7.Blattner FR, Plunkett G 3rd, Bloch CA, Perna NT, Burland V, Riley M, Collado-Vides J, Glasner JD, Rode CK, Mayhew GF, Gregor J, Davis NW, Kirkpatrick HA, Goeden MA, Rose DJ, Mau B and Shao Y. 1997. The complete genome sequence of Escherichia coli K12. Science. 277:1453-1474.
8.Altschul, S.F., Madden, T. L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res. 25: 3389-3402.
9.de Daruvar, A., Collado-Vides J. and Valencia, A. 2002. Analysis of the cellular functions of Escherichia coli operons and their conservation in Bacillus subtilis. J. Mol. Evol. 55: 211-221.