Validation of predicted anonymous proteins simply using Fisher's exact test

Jean-Michel Claverie[1,*] and Sébastien Santini[1]

[1]Aix Marseille Univ, CNRS, IGS (UMR7256), IMM (FR3479), Luminy, Marseille F-13288, France

*To whom correspondence should be addressed.

Abstract

Motivation: Given its increasing efficiency, accuracy, and decreasing cost, genomes sequencing has become the primary (and often the sole) experimental method to characterize newly discovered organisms, in particular from the microbial world (bacteria, archaea, viruses). This generates an ever increasing number of predicted proteins the existence of which is unwarranted, in particular when they do not share significant similarity with proteins of model organisms. As a last resort, the computation of the selection pressure from pairwise alignments of the corresponding "Open Reading Frames" (ORFs) can validate their existences. However, this approach is error-prone, as not usually associated with a significance test.

Results: We introduce the use of the straightforward Fisher's exact test as a post processing of the results provided by the popular CODEML pairwise sequence comparison software. The respective rates of nucleotide changes at the non-synonymous vs. synonymous position (as determined by CODEML), are turned into entries into a 2x2 contingency table, the probability of which is computed under the Null hypothesis that they should not behave differently (i.e. the ORFs do not encode real proteins). I show that strong negative selection pressures do not always provide a significant argument in favor of the existence of proteins.

Contact: Jean-Michel.Claverie@univ-amu.fr

1 Introduction

Since the first two bacterial genomes were sequenced 25 years ago (Fleischmann *et al.*, 1995; Fraser *et al.*, 1995), partial and whole genome sequencing have become the method of choice in identifying and charactering new microorganisms (bacteria, archaea, unicellular eukaryotes, viruses), revealing the stupendous extent of their diversity. By their simplicity of use and low cost, the most recent 3rd generation sequencing platforms (Goodwin *et al.*, 2016) have made microbial genomics accessible to non-specialists (MacLean *et al.*, 2009; Gwinn *et al.*, 2019), while a few large centers are fully taking advantage of their huge throughput to run biodiversity exploration projects of ever increasing dimensions (Harris *et al.*, 2018; Sunagawa *et al.*, 2020; Chen *et al.* 2021). The most recurring (and unexpected) lesson emerging from the analyses of these enormous datasets is that overall morphological and phylogenetic similarities, as well as similar metabolisms and lifestyles, could hide large differences in gene contents and encoded proteomes. Within microorganisms belonging to a given clade, such as eukaryotic classes, bacterial genera, or virus families, genomes are found to encode a subset of « core » proteins (i.e. with homologues in all members) together with proteins unevenly distributed, some only present in a single species or strain. This dichotomy is best documented for bacteria and viruses for which core genes might only represent a small proportion of the pangenome, i.e. of all the genes occurring at least once among all clade members (Land *et al.*, 2015, Claverie & Abergel, 2018). While there is little doubt that genes encoding homologous proteins in multiple divergent members of a clade are real, the level of certainty is much lowered when they occur only once, or within very close clade members where the corresponding Open Reading Frames (ORF) may occur by chance. Given the A+T richness of STOP codons (TAA, TAG, TGA), random ORFs are also statistically expected to occur at higher frequency in high G+C content genomes, increasing the risk of protein overprediction (Legendre *et al.*, 2019). The uncertainty further increases when the predicted protein is short (typically less than 100 residues) or does not exhibit a significant sequence (or functional motif) similarity in the reference databases (Sayers *et al.*, 2021). Such cases, referred to as "ORFans" represent a large proportion of predicted microbial proteomes (Entwistle *et al.*, 2019) in particular for large viruses (Philippe *et al.*, 2013; Gallot-Lavallée *et al.*, 2017; Legendre *et al.*, 2018; Abergel & Claverie, 2020; Boratto *et al.*, 2020). ORFans may result from *de novo* gene creations from non-coding regions in prokaryotes, eukaryotes, and their viruses (Legendre et al., 2018; McLysaght & Guerzoni, 2015; Schlötterer, 2015; Schmitz & Bornberg-Bauer, 2017; Van Oss & Carvunis, 2019).

If the experimental validation of predicted proteins has become much easier through various mass-spectrometry platforms, it is not available to all of the many laboratories generating genomic data. It also requires the corresponding microorganisms to be isolated and cultivated, thus disregarding the increasing number of metagenomics assemblies (Benler *et al*. 2021). Furthermore, certain proteins might only be expressed (and experimentally detectable) at specific time in the life cycle of an organisms, in certain environmental conditions, or in specific organs. Thus, our capacity of experimentally demonstrating the actual existence of predicted proteins has fallen much behind the overwhelming production rate of genomic data.

To overcome this difficulty it has become customary to compute the selection pressure, i.e. the ratio of the synonymous vs. non-synonymous mutation rates as a way to validate bioinformatic protein predictions (e.g.: Doutre *et al*., 2014: González *et al.*, 2016); Prabh & Rödelsperger, 2016); Christo-Foroux *et al*., 2020).

The concept/calculation of the selection pressure is based on the fact that proteins are usually made for a purpose and that their function, directly derived from their amino-acid sequences, tend to be conserved throughout evolution. Hence, in most cases, the non-synonymous positions of their coding regions will vary much less rapidly than the synonymous ones, the changes of which have lesser consequence on the organism's fitness. The concept of selection pressure, was most widely disseminated via the CODEML program of the PAML package for phylogenetic analysis (Yang, 2007; Jeffares, 2015). The computation requires the disposal of at least two homologous ORF sequences, and involves five straightforward steps: 1) from the comparison with the associated amino-acid sequence, each position in the ORF nucleotide sequence is classified as synonymous or non-synonymous in reference to the degeneracy of the genetic code. Their respective total numbers are denoted N_S and N_{NS}; 2) the two homologous amino-acid sequences are optimally aligned, then codon-wise converted into a nucleotide sequence alignment; 3) the observed nucleotide changes associated to the positions previously mapped as synonymous or non-synonymous are separately counted and are denoted n_S and n_{NS}; 4) one then forms the ratios $d_N= n_{NS}/N_{NS}$ and $d_S= n_S/N_S$, separately quantifying the mutation rates at the two different types of positions; 5) finally, one compute the "selection pressure" as the ratio $\Omega= d_N/d_S$.

The values of Ω are intuitively interpreted as follows: $\Omega < 1$ will correspond to proteins the (beneficial) functions of which resist amino-acid changes, also said to evolve under negative (purifying) selection. This is by far the most frequent situation. In contrast, $\Omega > 1$ correspond to the less frequent cases where changes in protein are positively selected (i.e. adaptive evolution) either to modify or abolish its (detrimental) function. Although conceptually simple, the practical implementation of this analysis comes up against two contradictory constraints. The first is that it must be based on an alignment of impeccable quality, and therefore between two highly similar protein sequences. The second is that the number of substitutions must be sufficiently high, while keeping the probability of multiple substitutions at the same site negligible (which would distort the estimate of d_s and d_{ns}). To our knowledge, the validity range of the method was never rigorously defined in terms of pair-wise sequence divergence (i.e.acceptable value ranges for N_S, N_{NS}, and the d_S or d_N ratios), although CODEML can compute a likelihood value for a large suite of adaptive evolution models (the grasp of which is beyond the reach of most of occasional users). Fortunately, the situation becomes easily tractable if we only wish to use CODEML and compute Ω to evaluate the quality of *ab initio* protein prediction from the pairwise alignment of two homologous ORFs, as presented in the next section.

2 Methods

The concept/computation of the selection pressure is based on the fact that the non-synonymous and synonymous positions of the coding regions of real proteins are expected to diverge at different speeds, thus leading to $\Omega \neq 1$ in most cases. However, in the case of false protein (ORF) predictions, the bioinformatics distinction made between non-synonymous and synonymous positions becomes irrelevant, and both types of positions are no longer expected to display a different mutational behavior. We then expect the computed selection pressure to remain close to one, within the range of random fluctuations.

As the non-synonymous and synonymous positions are two mutually exclusive categories, we can evaluate how much both positions behave differently using Fisher's exact test in the analysis of the 2x2 contingency table computed from the pairwise alignment of two homologous protein predictions, as follows:

	# Substituted	# Non-substituted
Non-synonymous	n_{NS}	$N_{NS}- n_{NS}$
Synonymous	n_S	$N_S- n_S$

The above values are directly available from the standard CODEML output, and the probability (P-value) that both position types behave differently (hence that the ORFS prediction are wrong) can be calculated by the many available implementation of Fisher's test (online or in R, for instance).

3 Results

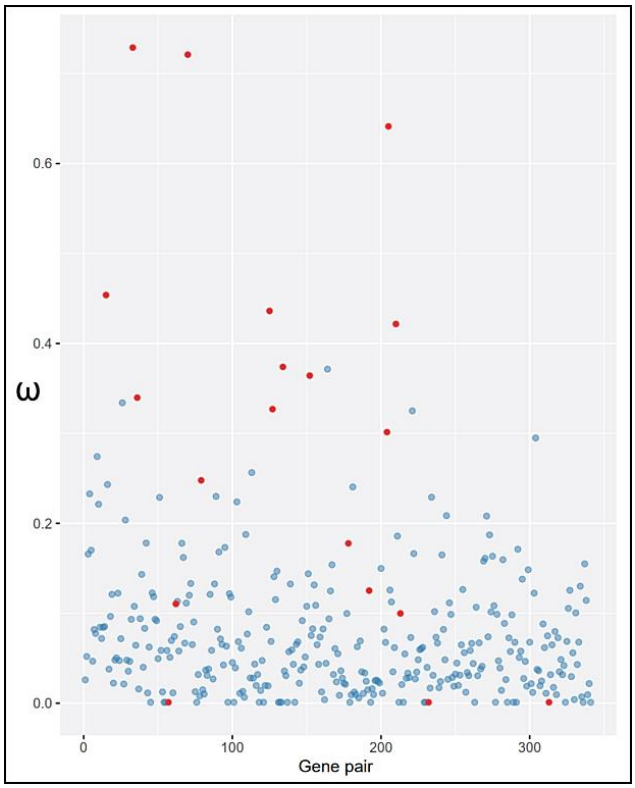
Table 1 present the results of the above analysis for 17 predicted ORFans, only shared by a Pacmanvirus-like strain recently isolated in our laboratory (ORF_B, in preparation) and a previously isolated strain (Andreani *et al*., 2017)(ORF_A, Pacmanvirus A23, NCBI # NC_034383). These predicted proteins (available upon request) share more than 95% identical residues, but do not exhibit other significant similarities in the public databases (Sayers. E.W. *et al.*, 2021). Although $2>d_S>0.01$ in all cases (the usual default pairwise alignment thresholds when using CODEML) (Aylward, 2018), none of these alignments but one exhibit a significant difference (P-value <0.05) in the mutational behavior of the non-synonymous *vs*. synonymous positions. Interestingly, the corresponding P-value (hence the validity of the protein prediction) does not correlate with Ω values [0.001-0.729], some of which (e.g. when Ω <0.4) might be attributed, - when using an unchecked, genome-wide automated analysis protocol -, to *de novo* created genes ongoing purifying selection. On the contrary, the use of Fisher's test on the raw substitution frequencies strongly suggests that most ORFans shared by these two new strains are conserved by chance, and might not be expressed as actual proteins.

The application of the above testing protocol is not limited to ORFans, but can be used in the context of genome-wide analyses to test the reliability of Ω estimations on each pairs of orthologous ORFs (Fig. 1). Interestingly, the use of Fisher's exact test automatically filters out pairwise alignments that do not exhibit enough substitutions because they are too similar, or their alignments too short. The only parameter then remaining to be fixed is the % of identical residues between orthologous proteins that should be greater than 70% to ensure high quality pairwise alignments, and minimize the probability of multiple substitutions at one given site (Jeffares, 2015).

Table 1. Validation of ORFan predictions from two closely related Pacmanvirus strains-like strain.

ORF_A	ORF_B	Ω	d_N	d_S	P value	ORF Validation	n_S	N_S-n_S	n_{NS}	N_{NS}-n_{NS}
A_002	B_5	0.088	0.041	0.465	<E-59	Yes	128	148	40	926
A_111	B_117	0.454	0.021	0.047	0.16	No	5	98	9	413
A_011	B_14	0.729	0.013	0.018	1.00	No	1	62	3	219
A_142	B_152	0.340	0.032	0.093	0.22	No	4	44	5	157
A_164	B_176	0.001	0.000	0.020	0.18	No	1	50	0	228
A_169	B_180	0.110	0.004	0.034	0.09	No	2	60	1	267
A_178	B_190	0.721	0.010	0.013	1.00	No	1	80	3	312
A_193	B_204	0.248	0.013	0.051	0.18	No	2	43	3	237
A_251	B_259	0.436	0.024	0.054	0.31	No	2	36	3	130
A_252	B_260	0.327	0.018	0.056	0.25	No	2	38	3	164
A_260	B_268	0.374	0.013	0.035	0.06	No	7	183	7	529
A_286	B_292	0.364	0.008	0.021	0.42	No	1	52	2	263
A_313	B_318	0.178	0.004	0.023	0.33	No	1	53	1	245
A_327	B_332	0.125	0.004	0.029	0.23	No	1	38	1	272
A_341	B_348	0.641	0.023	0.036	0.59	No	1	33	4	172
A_348	B_355	0.422	0.010	0.024	0.46	No	1	43	2	194
A_057	B_63	0.001	0.000	0.029	0.16	No	1	39	0	210

Fig. 1. Selection pressure associated to all proteins shared by two close Pacmanvirus strains. ORFs associated to Ω values not significantly different from 1 are in red (P-value>0.05). All ORFs with Ω>0.4 are suspected to be false predictions.



Acknowledgements

We acknowledge the use of the PACA-Bioinfo platform.

Funding

The IGS laboratory is partially supported by recurrent funding from CNRS and Aix-Marseille University

Conflict of Interest: none declared.

References

Abergel, C., Claverie, J.M. (2020) Giant viruses. *Curr. Biol.*, **30**, R1108-R1110.

Andreani, J. *et al.* (2017) Pacmanvirus, a New Giant Icosahedral Virus at the Crossroads between Asfarviridae and Faustoviruses. *J. Virol.*, 2017, **91**, e00212-17.

Aylward, F.O. (2018) Introduction to calculating dN/dS ratios with codeml V.2. dx.doi.org/10.17504/protocols.io.qhwdt7e

Benler, S. *et al.* (2021) Thousands of previously unknown phages discovered in whole-community human gut metagenomes. *Microbiome*, **9**, 78.

Boratto, P.V.M. (2020) Yaravirus: A novel 80-nm virus infecting *Acanthamoeba castellanii*. *Proc. Natl. Acad. Sci. U. S. A.*, **117**, 16579-16586.

Chen, I-M. *et al.* (2021) The IMG/M data management and analysis system v.6.0: new tools and advanced capabilities. *Nucleic Acids Res.*, **49**, D751–D763.

Christo-Foroux, E, *et al.* (2020) Characterization of *Mollivirus kamchatka*. the First Modern Representative of the Proposed Molliviridae Family of Giant Viruses. *J. Virol.*, **94**, e01997-19.

Claverie, J.M. and Abergel, C. (2018) *Mimiviridae*: An expanding family of highly diverse large dsDNA viruses infecting a wide phylogenetic range of aquatic eukaryotes. *Viruses*, **10**, 506.

Doutre, G., Philippe, N., Abergel, C., Claverie, J.M. (2014) Genome analysis of the first Marseilleviridae representative from Australia indicates that most of its genes contribute to virus fitness. *J. Virol.*, **88**, 14340-14349.

Entwistle, S., Li, X., Yin, Y. (2019) Orphan Genes Shared by Pathogenic Genomes Are More Associated with Bacterial Pathogenicity. *mSystems*, **4**, e00290-18.

Fleischmann, R.D. *et al.* (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, **269**, 496–512.

Fraser, C.M. *et al.* (1995) The minimal gene complement of *Mycoplasma genitalium*. *Science*, **270**, 397–404.

Gallot-Lavallée, L., Blanc, G., Claverie, J.M. (2017) Comparative Genomics of *Chrysochromulina Ericina* Virus and Other Microalga-Infecting Large DNA Viruses Highlights Their Intricate Evolutionary Relationship with the Established Mimiviridae Family. *J. Virol.*, **91**, e00230-17.

González, C., Lazcano, M., Valdés, J., Holmes, D.S. (2016) Bioinformatic Analyses of Unique (Orphan) Core Genes of the Genus Acidithiobacillus: Functional Inferences and Use As Molecular Probes for Genomic and Metagenomic/Transcriptomic Interrogation. *Front. Microbiol.*, **7**, 2035.

Goodwin, S., McPherson, J.D., McCombie, W.R. (2016) Coming of age: ten years of next-generation sequencing technologies. *Nat. Rev. Genet.*, **17**, 333-351.

Gwinn, M., MacCannell, D., Armstrong, G.L. (2019) Next-Generation Sequencing of Infectious Pathogens. *JAMA*, **321**, 893–894.

Jeffares, D.C. *et al.* (2015) A Beginners Guide to Estimating the Non-synonymous to Synonymous Rate Ratio of all Protein-Coding Genes in a Genome. In Peacock. C. (ed.). *Parasite Genomics Protocols*. Methods in Molecular Biology. Springer Science+Business Media. New York 2015. Vol. 1201. pp. 65–90.

Land, M. et al. (2015) Insights from 20 years of bacterial genome sequencing. *Funct. Integr. Genomics.*, **15**, 141-161.

Legendre, M. *et al.* (2018) Diversity and evolution of the emerging Pandoraviridae family. *Nat. Commun.*, **9**, 2285.

Legendre, M. *et al.* (2019) *Pandoravirus celtis* illustrates the microevolution processes at work in the giant Pandoraviridae Genomes. *Front. Microbiol.*, **10**, 430.

Lewin, H.A. *et al.* (2018) Earth BioGenome Project: Sequencing life for the future of life. *Proc. Natl. Acad. Sci. U. S. A.*, **115**, 4325-4333.

MacLean, D., Jones, J., Studholme, D. (2009) Application of 'next-generation' sequencing technologies to microbial genetics. *Nat. Rev. Microbiol.*, **7**, 96–97.

McLysaght, A., Guerzoni, D. (2015) New genes from non-coding sequence: the role of de novo protein-coding genes in eukaryotic evolutionary innovation. *Philos. Trans. R. Soc. Lond. B Biol. Sci.*, **370**. 20140332.

Philippe, N. *et al.* (2013) Pandoraviruses: amoeba viruses with genomes up to 2.5 Mb reaching that of parasitic eukaryotes. *Science*, **341**, 281-286.

Prabh, N., Rödelsperger, C. (2016) Are orphan genes protein-coding. prediction artifacts. or non-coding RNAs? *BMC Bioinformatics*., **17**, 226.

Sayers, E.W. *et al.* (2021) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res.*, **49**(D1), D10-D17.

Schlötterer, C. (2015) Genes from scratch--the evolutionary fate of de novo genes. *Trends Genet.*, **31**, 215-219.

Sunagawa, S. *et al.* (2020) Tara Oceans: towards global ocean ecosystems biology. *Nat. Rev. Microbiol.*, **18**, 428-445.

Schmitz, J.F., Bornberg-Bauer, E. (2017) Fact or fiction: updates on how protein-coding genes might emerge de novo from previously non-coding DNA. *F1000Res.*, **6**, 57.

Yang, Z. (2007) PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.*, **24**, 1586–1591.

Yang, Z. (2014) Molecular Evolution: A statistical approach. Oxford University Press. pp. 47–55.

Van Oss, S.B., Carvunis, A.R. (2019) De novo gene birth. *PLoS Genet.*, **15**, e1008160.