Bioinformatics, 2017, 1–4
doi: 10.1093/bioinformatics/btx394
Advance Access Publication Date: 16 June 2017
Applications Note

OXFORD

Sequence analysis

EDEN: evolutionary dynamics within environments

Philipp C. Münch[1,2,3], Bärbel Stecher[2,4] and Alice C. McHardy[1,3,5,6,*]

[1]Computational Biology of Infection Research, Helmholtz Centre for Infection Research, 38124 Brunswick, Germany, [2]Max von Pettenkofer-Institute for Hygiene and Clinical Microbiology, Ludwig-Maximilian University of Munich, 80336 Munich , Germany, [3]German Centre for Infection Research (DZIF), Partner Site Hanover-Brunswick, 38124 Brunswick, Germany, [4]DZIF, Partner Site LMU Munich, 80336 Munich, Germany, [5]Department of Algorithmic Bioinformatics and [6]Cluster of Excellence on Plant Sciences (CEPLAS), Heinrich Heine University Dusseldorf, 40225 Dusseldorf, Germany

*To whom correspondence should be addressed.
Associate Editor: Inanc Birol

Received on January 14, 2017; revised on May 19, 2017; editorial decision on June 9, 2017; accepted on June 15, 2017

Abstract

Summary: Metagenomics revolutionized the field of microbial ecology, giving access to Gb-sized datasets of microbial communities under natural conditions. This enables fine-grained analyses of the functions of community members, studies of their association with phenotypes and environments, as well as of their microevolution and adaptation to changing environmental conditions. However, phylogenetic methods for studying adaptation and evolutionary dynamics are not able to cope with big data. EDEN is the first software for the rapid detection of protein families and regions under positive selection, as well as their associated biological processes, from meta- and pangenome data. It provides an interactive result visualization for detailed comparative analyses.
Availability and implementation: EDEN is available as a Docker installation under the GPL 3.0 license, allowing its use on common operating systems, at http://www.github.com/hzi-bifo/eden.
Contact: alice.mchardy@helmholtz-hzi.de
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Microorganisms can adapt to changing environmental conditions by evolutionary processes such as mutation, lateral gene transfer and recombination (Bendall *et al.*, 2016; Denef and Banfield, 2012; Koonin *et al.*, 2002). Only a small fraction of mutations are thought to be beneficial, and few will be fixed and contribute to the substitution rate measurable in phylogenomic studies (Bendall *et al.*, 2016; Nishant *et al.*, 2009). One of the most widely used measures for quantifying the type and extent of selection acting on a protein family site or sequence at the molecular scale is the d_N/d_S ratio (Ford, 2002; Nielsen, 2005). If the rate of change at nonsynonymous sites (d_N) within a gene family exceeds the rate of change at synonymous sites (d_S), i.e. $d_N/d_S > 1$, positive selection is assumed to operate on the encoded protein. This indicates that adaptation to altered environmental conditions is taking place and that the observed changes increase the fitness of the respective organism. A $d_N/d_S < 1$, on the other hand, is taken as an indicator for negative selection, with changes in the protein sequence or at a site decreasing fitness (Hurst, 2002; Koonin and Rogozin, 2003), such as for instance, changes in catalytic sites that would lead to a loss of function.

Calculating the d_N/d_S ratio for the large-scale sequence datasets that are being generated in metagenomics and comparative microbial genomics is very challenging, due to the run times of commonly used tree inference software and software for quantifying positive selection, such as FastCodeML (Valle *et al.*, 2014), which relies on maximum likelihood methods (Pond and Frost, 2005). With EDEN, we provide a fully automated software package and visualization framework for a rapid meta- or pangenome wide analysis of the evolutionary processes affecting protein families and associated biological processes. EDEN is based on a fast approximate tree inference

and count-based d_N/d_S inference for individual protein families. The software can be applied to compare the selection profiles of bacterial species with different phenotypes or lifestyles, such pathogens versus mutualists, and to study selection from metagenome datasets of microbial communities (Fig. 1b).

2 Implementation

EDEN is provided as a Docker image, which is a virtualization of the application that includes everything needed to run the program (Merkel, 2014) (Fig. 1a). Other than Docker, no further software has been installed. EDEN requires as input (meta-)genome DNA sequences in FASTA format, from which open reading (ORF) frame DNA and protein sequences will be generated with Prodigal (Hyatt *et al.*, 2010) (Fig. 1a). Alternatively, the user can provide two files in FASTA format, corresponding to the DNA and protein sequences for a set of ORFs. Optionally one or multiple Hidden Markov Models (HMMs) can also be provided, which will be used to infer groups of ORFs, as well as a sample grouping table, with sample properties of interest, such as their origin, to enable comparative analyses of multiple input samples in groups representing these properties.

The first step in the assessment of evolutionary patterns for the input sequences is their division into groups of ORFs that are subsequently processed together. Groups can either be obtained (i) based on the user-provided grouping table, (ii) by searching for protein family members with hmmsearch against user-provided hidden Markov Models (Eddy, 1998) or (iii) by searching the ORFs using hmmsearch versus the complete TIGRFAM HMM collection (Haft *et al.*, 2003). For each group, then a multiple DNA sequence alignment (MSA) is calculated with MUSCLE (Edgar, 2004). Subsequently, a multiple codon alignment is constructed using PAL2NAL 14 under consideration of the MSA and the protein sequences (Suyama *et al.*, 2006). Based on the codon MSA, a phylogenetic tree is reconstructed with an efficient implementation of the neighbor-joining algorithm using a modified version of Clearcut (Sheneman *et al.*, 2006). Specifically, we control for gaps in the alignment that are mostly of technical origin (due to the alignment of smaller assembled contigs to longer reference sequence), by excluding these from mismatch counts in calculation of the additive pairwise distance matrix. Next, d_N/d_S is calculated using the counting method (Pond and Frost, 2005), which achieves a trade-off between the computational effort and the quality of the estimates. For d_N/d_S calculation, the ancestral amino acid and coding sequences are reconstructed for all internal nodes of each protein family tree, using maximum parsimony as the optimization criterion. The values of d_N and d_S are then inferred from these sequences, considering the least costly of several different mutation paths between codons, as in Tusche *et al.* (2012). The d_N/d_S ratio is then calculated using a lookup table with the probabilities that a change will cause a nonsynonymous change for all possible codon comparisons possible (Nei and Gojobori, 1986).

For calculation of the average d_N/d_S for a considered group, low-confidence positions are excluded by filtering positions from the alignment with a user-defined proportion of gaps. P values are calculated using a one-sided Fisher's exact test, based on the d_N and d_S rates for every sequence group in comparison to the entire sample. The false discovery rate is used to control for multiple testing errors, using the Benjamini and Hochberg procedure (Benjamini and Hochberg, 1995), with α per default set to 0.05. To detect putative epitopes within a given set of homologs that are under positive selection, the P value for the sum of the d_N and d_S rates is calculated using a sliding window approach (with the size of 20 codons as a default) and a one-sided Fisher's exact test over the MSA (Bulgarelli *et al.*, 2015; McCann *et al.*, 2012) (see Supplementary Material for details).

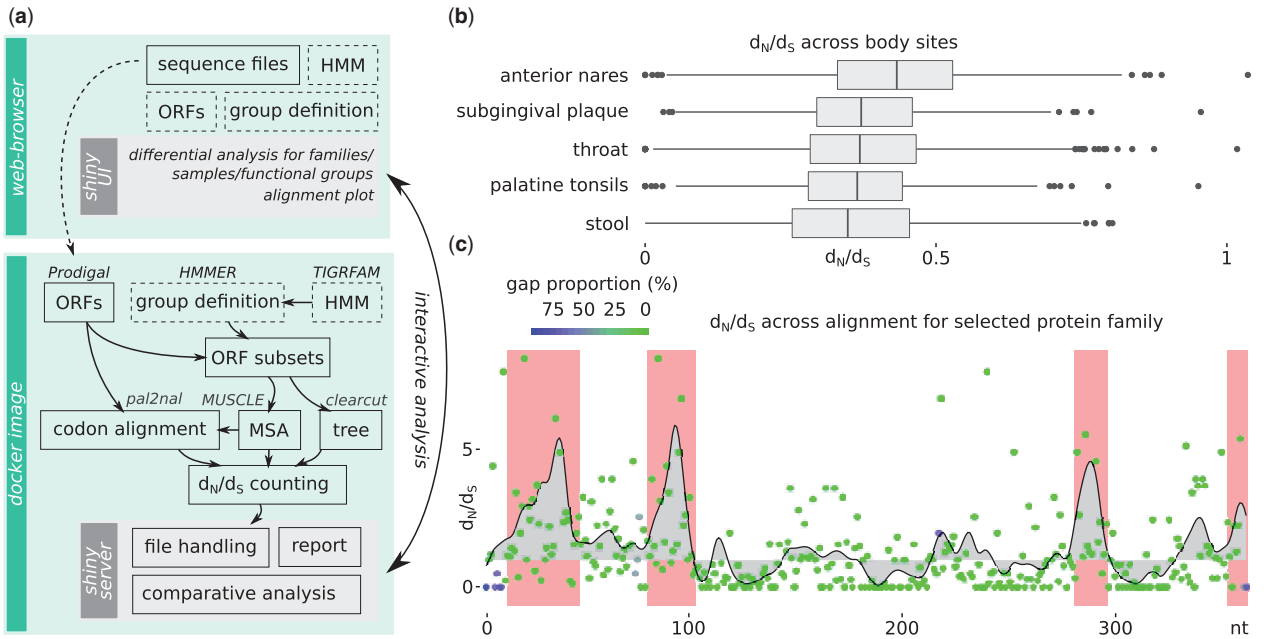

Fig. 1. (**a**) Sample and data processing workflow for d_N/d_S profiling using EDEN. To enable an interactive analysis, an RStudio Shiny server will be started inside the Docker image which is accessible by the user by a web-browser via localhost. Dashed boxes are optional input files. (**b**) Interactive visualization enables comparison of pooled samples. Here selected HMP samples pooled by body site are shown. (**c**) Example output of clusters of residues under positive selection for one gene family. Dots indicate d_N/d_S for a given position in the protein sequence, and their color corresponds to the proportion of gaps in the MSA. Red areas indicate significant clusters of residues under positive selection. Abbreviations: UI, User Interface

As a 'sanity check', we compared the d_N/d_S of EDEN with HyPhy SLAC, which uses a derivative of the Suzuki-Gojobori counting approach, for 50 randomly selected protein families from the HMP dataset and found a high correlation (Pearson's R = 0.873, *P* value = 2.499e-16, Supplementary Fig. S2).

In comparison to FastCodeML (Valle *et al.*, 2014), a run-time optimized version of the codeml program from the PAML package (Yang, 2007), EDEN has a drastically reduced run-time (Supplementary Fig. S2, see Supplementary Material for details). For further detailed analyses with FastCodeML, such as assessing selection for specific clades, the codon alignment and tree calculated by EDEN can be downloaded for individual protein families.

3 Application

We previously used EDEN to study protein families under selection from six assembled metagenome samples (150.000 ORF sequences) of the root microbiota for wild and domesticated barley (*Hordeum vulgare*). This delivered evidence for positive selection acting on protein families linked to pathogenesis, bacteria-phage interactions, secretion and nutrient mobilization in the barley root-associated microbiota (Bulgarelli *et al.*, 2015) and for a higher degree of selection acting on protein families from the root-associated microbiota than on those found in bulk soil. EDEN was also used to compare the selection patterns for protein families from multiple strains of *Colletotrichum* (Hacquard *et al.*, 2016).

We applied EDEN to 66 samples of the HMP project (Consortium *et al.*, 2012) from six body sites (body sites dataset on http://eden.bifo.helmholtz-hzi.de, Fig. 1b). These were sampled from healthy individuals and had similar alpha- and beta diversities, except for the stool samples, which were more diverse (Consortium *et al.*, 2012). The results indicate a positive relationship between d_N/d_S values and the exposure of body sites to the surrounding environment. The highest d_N/d_S values were found for samples from the external portion of the nose (anterior nares), followed by the oral microbiome (subgingival plague, palatine tonsils, throat) and the lowest values for stool. Also in comparison of samples of one body site to all others, a significantly increased d_N/d_S (FDR corrected *P* values < 0.001) was observed (in that order) for the microbiome from the external portion of the nose, subgingival plaque and throat. Interestingly, across all six body sites, most protein families with significant signs of positive selection in comparison to all other protein families from the respective samples (FDR adjusted *P* value <0.01) were annotated with transport and binding functions, suggesting the existence of a functional panselectome. Other than that, 'energy metabolism' was found as a prominent association for the oral and nose associated microbiomes.

We also used EDEN to characterize human gut metagenome samples from (Qin *et al.*, 2010) (BMI dataset on http://eden.bifo.helmholtz-hzi.de). EDEN determined a significantly higher d_N/d_S for the protein coding genes from lean individuals (BMI < 25) compared to overweight (BMI 25-30) and obese individuals (BMI > 30; *P* value < 0.001), suggestive of a higher functional diversity in the guts of lean individuals. For lean individuals compared to obese individuals, this finding was consistent over all functional groups, except for regulatory functions. For these, d_N/d_S was slightly, though not significantly, higher for the obese than for the lean individuals.

4 Conclusion

EDEN can identify protein families under positive selection from metagenome and pangenome datasets. It reports gene families and regions thereof with a significantly elevated d_N/d_S in comparison to a specified background and allows comparative studies of multiple samples. The results obtained for metagenome samples from different demonstrate how these analyses provide insights into the relationship between the signs of molecular adaptation found in the microbiome to biological processes and their environments.

Author contributions

A.C.M. and P.C.M. conceived and designed the experiments. P.C.M. implemented the software. P.C.M. wrote the manuscript with comments from A.C.M. and B.S. All authors approved the final version of the manuscript.

Acknowledgement

We thank Lars Steinbrück for valuable comments and for providing and implementation of his original code.

Conflict of Interest: none declared.

References

Bendall,M.L. *et al.* (2016) Genome-wide selective sweeps and gene-specific sweeps in natural bacterial populations. *ISME J*, **10**, 1589.

Benjamini,Y. and Hochberg,Y. (1995) Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J. R. Stat. Soc. Ser. B (Methodological)*, **57**, 289–300.

Bulgarelli,D. *et al.* (2015) Structure and function of the bacterial root microbiota in wild and domesticated barley. *Cell Host Microbe*, **17**, 392–403.

Consortium,H.M.P. *et al.* (2012) Structure, function and diversity of the healthy human microbiome. *Nature*, **486**, 207–214.

Denef,V.J. and Banfield,J.F. (2012) In situ evolutionary rate measurements show ecological success of recently emerged bacterial hybrids. *Science*, **336**, 462–466.

Eddy,S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14**, 755–763.

Edgar,R.C. (2004) Muscle: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res.*, **32**, 1792–1797.

Ford,M.J. (2002) Applications of selective neutrality tests to molecular ecology. *Mol. Ecol.*, **11**, 1245–1262.

Hacquard,S. *et al.* (2016) Survival trade-offs in plant roots during colonization by closely related beneficial and pathogenic fungi. *Nat. Commun.*, **7**, 11362.

Haft,D.H. *et al.* (2003) The tigrfams database of protein families. *Nucleic Acids Res.*, **31**, 371–373.

Hurst,L.D. (2002) The ka/ks ratio: diagnosing the form of sequence evolution. *Trends Genet.*, **18**, 486–487.

Hyatt,D. *et al.* (2010) Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**, 1.

Koonin,E.V. and Rogozin,I.B. (2003) Getting positive about selection. *Genome Biol.*, **4**, 331.

Koonin,E.V. *et al.* (2002) Horizontal gene transfer in prokaryotes: quantification and classification.

McCann,H.C. *et al.* (2012) Identification of innate immunity elicitors using molecular signatures of natural selection. *Proc. Natl. Acad. Sci. USA*, **109**, 4215–4220.

Merkel,D. (2014) Docker: lightweight linux containers for consistent development and deployment. *Linux J.*, **2014**, 2.

Nei,M. and Gojobori,T. (1986) Simple methods for estimating the numbers of synonymous and nonsynonymous nucleotide substitutions. *Mol. Biol. Evol.*, **3**, 418–426.

Nielsen,R. (2005) Molecular signatures of natural selection. *Annu. Rev. Genet.*, **39**, 197–218.

Nishant,K.T. *et al*. (2009) Genomic mutation rates: what high-throughput methods can tell us. *Bioessays*, **31**, 912–920.

Pond,S.L.K. and Frost,S.D. (2005) Not so different after all: a comparison of methods for detecting amino acid sites under selection. *Mol. Biol. Evol*., **22**, 1208–1222.

Qin,J. *et al*. (2010) A human gut microbial gene catalogue established by metagenomic sequencing. *Nature*, **464**, 59–65.

Sheneman,L. *et al*. (2006) Clearcut: a fast implementation of relaxed neighbor joining. *Bioinformatics*, **22**, 2823–2824.

Suyama,M. *et al*. (2006) Pal2nal: robust conversion of protein sequence alignments into the corresponding codon alignments. *Nucleic Acids Res*., **34**, W609–W612.

Tusche,C. *et al*. (2012) Detecting patches of protein sites of influenza a viruses under positive selection. *Mol. Biol. Evol*., **29**, 2063–2071.

Valle,M. *et al*. (2014) Optimization strategies for fast detection of positive selection on phylogenetic trees. *Bioinformatics*, btt760.

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