Bioinformatics, 2016, 1–5
doi: 10.1093/bioinformatics/btw231
Advance Access Publication Date: 5 May 2016
Original Paper

OXFORD

Genome analysis

DOGMA: domain-based transcriptome and proteome quality assessment

Elias Dohmen[1,2,*], Lukas P.M. Kremer[1], Erich Bornberg-Bauer[1] and Carsten Kemena[1,*]

[1]Institute for Evolution and Biodiversity, University of Münster, Münster 48149, Germany and [2]Institute for Bioinformatics and Chemoinformatics, Westphalian University of Applied Sciences, Recklinghausen 45665, Germany

*To whom correspondence should be addressed.

Associate Editor: John Hancock

Received on January 29, 2016; revised on April 20, 2016; accepted on April 21, 2016

Abstract

Motivation: Genome studies have become cheaper and easier than ever before, due to the decreased costs of high-throughput sequencing and the free availability of analysis software. However, the quality of genome or transcriptome assemblies can vary a lot. Therefore, quality assessment of assemblies and annotations are crucial aspects of genome analysis pipelines.
Results: We developed DOGMA, a program for fast and easy quality assessment of transcriptome and proteome data based on conserved protein domains. DOGMA measures the completeness of a given transcriptome or proteome and provides information about domain content for further analysis. DOGMA provides a very fast way to do quality assessment within seconds.
Availability and Implementation: DOGMA is implemented in Python and published under GNU GPL v.3 license. The source code is available on https://ebbgit.uni-muenster.de/domainWorld/DOGMA/.
Contacts: e.dohmen@wwu.de or c.kemena@wwu.de
Supplementary information: Supplementary data are available at *Bioinformatics* online.

1 Introduction

Genomics and transcriptomics have become key technologies in all branches of life sciences over the last years and the huge amount of resulting data calls for proper computational analysis methods. Though many improvements were made in automatized sequence assembly and genome annotation, both processes remain an extremely challenging task, and there are enormous differences in the quality of the resulting data (Fang *et al.*, 2013). These differences result mainly from problems that can occur during the various steps of an assembly and annotation pipeline. Reads can be of low quality, they may not be precisely mapped and/or many small scaffolds might exist, making subsequent steps in the pipeline, such as gene annotation, more complicated.

Even with high quality assemblies, gene annotation is a complex process and also with transcriptome data as supporting information, gene models can be wrong, incomplete or missing altogether from the annotation (Yandell and Ence, 2012). Therefore, quality assessment of the results is important and, accordingly, several methods to determine quality of reads, assemblies, transcriptome and proteome data have been developed. Read quality can be assessed using per-base error-rates or k-mer distributions (Clark *et al.*, 2013; Gurevich *et al.*, 2013). For resulting genome or transcriptome assemblies, programs like CEGMA (Parra *et al.*, 2007) or BUSCO (Simão *et al.*, 2015) have been developed. Both CEGMA and BUSCO use a similar approach: a predefined set of genes, that are conserved over a specific clade, is compared to genes annotated in the genome or transcriptome to be analyzed. The measured gene completeness, the percentage of core genes found in the sample, is assumed to be representative for the whole gene set of the genome and is therefore used as a quality indicator. The CEGMA scores of Ensembl genomes (Cunningham *et al.*, 2014, version 81), for example, are in the range of 46.77% (*Vicugna pacos*) to 97.58% (*Saccharomyces cerevisiae*).

The core set of CEGMA is based on 'euKaryotic clusters of Orthologous Groups' (KOGs), resulting in a set of Core Eukaryotic Genes (CEGs) (Tatusov *et al.*, 2003). BUSCO's core gene set is based on orthologous genes from OrthoDB (Waterhouse *et al.*, 2013). CEGMA and BUSCO can be used for genome quality assessment, while for transcriptomes and proteomes only BUSCO can be used. The protein quality index (PQI) provides a list of quality scores for publicly available proteomes (Zaucha *et al.*, 2015). The quality score consists of different components, such as sequence length, domain content, percentage of undefined amino acids and others. Protein domains are independently evolving structural and functional building blocks of proteins (Marsh and Teichmann, 2010; Moore *et al.*, 2008), known to be well conserved across taxa (Ekman *et al.*, 2005). Usually, domains are represented as Hidden Markov Models (HMMs). As HMMs are more sensitive than other sequence search methods, protein domains can be identified in very diverged sequences with HMMs (Remmert *et al.*, 2011). Because of the better detection of protein domains, they are a good candidate for quality assessment.

Protein domains can occur together in the same protein and form in combination a domain arrangement, specified by their order in the amino acid sequence (Forslund and Sonnhammer, 2012).

The domain content is considered in two different measures of the PQI. The first one is that the sample proteome is searched for SUPERFAMILY (Gough *et al.*, 2001) domain arrangements that occur in the CEGs of CEGMA. The second measure is a simple comparison of the number of different domains that occur in the sample proteome compared to the number of domains in the same clade. However, no program is provided and new proteomes can only be evaluated by sending a request to the providers of the PQI.

The domain content is incorporated in the program introduced here as well. Domain arrangements have been successfully used in different contexts such as homology search (Terrapon *et al.*, 2014) or orthology construction (Bitard-Feildel *et al.*, 2015). The advent of analyses using protein domains and their arrangements to study functional roles and evolutionary relationships of proteins and complete families demonstrates the utility of information about absence or presence of specific protein domains and their arrangements (e.g. Forslund and Sonnhammer, 2008; Moore and Bornberg-Bauer, 2012; Sardar *et al.*, 2014; Vogel *et al.*, 2005).

In this work we introduce DOGMA (DOmain-based General Measure for transcriptome and proteome quality Assessment), a program that facilitates the quick assessment of transcriptome and proteome quality. DOGMA scores a sample transcriptome or proteome regarding its completeness of conserved protein domains provided as percentage of a defined core set. Combined with a tool for fast domain annotation (UProC (Meinicke, 2015)), DOGMA is able to rapidly perform quality assessments and provide information about missing domains and domain arrangements.

2 Methods

2.1 Construction of a core set of conserved domains and domain arrangements

A core set of Conserved Domain Arrangements (CDAs) is used in DOGMA to be compared against the transcriptome or proteome of interest (the sample transcriptome/proteome). A CDA can also consist of a single domain. Per default the core set consists of CDAs that are conserved across six eukaryotic model species: (*Arabidopsis thaliana*, *Caenorhabditis elegans*, *Drosophila melanogaster*, *Homo sapiens*, *Mus musculus* and *Saccharomyces cerevisiae*). Beside this eukaryotic core set, DOGMA offers alternatively specialized core sets for mammals and insects (see below and Supplementary Online Material).

To take into account the different requirements of transcriptome and proteome data, DOGMA distinguishes between these two kinds of data by scoring CDAs in different ways. For this reason DOGMA can be run in a proteome and a transcriptome mode. The first step of producing the core set of conserved CDAs is to annotate the domains in the selected core species. We use UProC (Meinicke, 2015) to produce the domain annotation of core as well as sample data. UProC is very fast although this comes with the cost of a higher memory usage. Because UProC does not provide positional information of the domains it is not possible to use order information for the domain arrangements. DOGMA therefore incorporates every possible subset of domains in a single sequence. The size of CDAs DOGMA considers can be selected by the user and is set to 3 by default. This means that domain arrangements comprising more domains are considered as well, but just their subsets up to the size of 3 are counted (e.g. A-B-C-D → A-B-C, A-B-D, A-C-D, B-C-D). We analyzed the effect of the CDA size parameter on the comparison results by using different maximum sizes.

CDAs can occur in different organisms at a variable frequency. The better conserved a CDA is, the more likely it should appear in all core species at nearly the same frequency. Protein isoforms may bias this analysis. For proteome data it is usually very easy to filter out the different isoforms and keep only one version of the protein (e.g. the longest). This approach does allow taking the frequency of occurring CDAs into account in DOGMA (proteome mode). For the construction of the core set DOGMA takes the minimum number of CDAs that were encountered, but only uses CDAs that do not vary by more than two in the number of occurrences among all core species. Protein domain repeats are known to differ significantly in their copy number (Ekman *et al.*, 2007). Therefore, all domains identified as repeats in the Pfam database are handled separately by DOGMA and just the lowest domain count shared by all core species is set as the domain count for a particular CDA. However, to allow a simpler usage on transcriptome data, an isoform cleaning step is not required. Because of the missing isoform cleaning step, frequency information is not considered, but a more simplistic presence/absence evaluation is used (transcriptome mode). Table 1 summarizes the number of CDAs in the different core sets for both modes.

As UProC is not the standard program of the Pfam database, DOGMA supports additionally PfamScan for domain annotation that is provided by the Pfam database. PfamScan provides positional information of domains in arrangements, but is much slower than UProC. With PfamScan it is possible to take the order of domains in CDAs into account (proteome mode), which might improve the result because the order of domains is often maintained over evolutionary long distances (Kummerfeld and Teichmann, 2009). In this case, DOGMA collapses consecutive duplications of domains in CDAs to a single instance of the domain (e.g. A-B-B-B-C → A-B-C),

Table 1. Number of CDAs in DOGMA core sets with default settings in proteome (prot) and transcriptome (trans) mode (based on the annotation with UProC)

CDA size	Eukaryotes		Mammals		Insects	
	prot	trans	prot	trans	prot	trans
1	965	1975	4204	5010	2893	3578
2	660	1021	11 082	6182	4351	3412
3	392	485	8894	5666	3208	2547

as it has been shown that copy number variations can also occur between closely related species (Ekman *et al.*, 2007).

2.2 Evaluating a transcriptome/proteome using CDAs

Transcriptomes as well as proteomes can be evaluated using UProC or PfamScan annotations. UProC can be directly applied to both kinds of data. PfamScan annotations of transcriptomes, however, need to be produced manually as an initial ORF prediction and translation is necessary but not directly supported in DOGMA. The subsequent evaluation of a sample transcriptome or proteome is executed under the specifications stated above for the corresponding core set construction. Next to the completeness score of the sample transcriptome or proteome, all missing CDAs are listed in the DOGMA output.

2.3 Benchmark and completeness scores in comparison

To validate the suitability of DOGMA's completeness scores for assessing data quality, different datasets were analyzed and the results were compared to the results of existing programs. We used real as well as simulated data. The real dataset consists of proteome and transcriptome data on which DOGMA completeness scores were compared to the scores generated with CEGMA (Parra *et al.*, 2007, version 2.5) and BUSCO (version 1.1b). The real data consists of a proteome set of all 69 species from Ensembl (Cunningham *et al.*, 2014, version 81) and a transcriptome set of 98 tissue samples from 79 species (the same dataset that has been used to evaluate BUSCO (Simão *et al.*, 2015)). For proteome data, the 69 Ensembl proteomes were annotated with pfam domains (Finn *et al.*, 2014, version 28) using UProC (Meinicke, 2015). These sets were then analyzed with DOGMA and the corresponding genomes with CEGMA. BUSCO evaluations were also based on the proteome data. BUSCO was additionally used to analyze the transcriptome data and compared to an annotation of the same data with UProC and subsequent analysis with DOGMA. Core sets for eukaryotes, metazoans, vertebrates, arthropods or fungi were used in BUSCO, while DOGMA utilized core sets for eukaryotes, mammals or insects whenever suitable. CEGMA used optimizations for vertebrates or mammals when suitable. The benchmark data has also been analyzed via DOGMA in a combination with the PfamScan utilities for annotation instead of UProC. The results can be found in the Supplementary Data. The simulated dataset consists of several transcriptomes that have been produced with a more relaxed adapter and quality trimming in each run to have a stepwise decrease in transcriptome quality. The exact process is described in the Supplementary Data.

3 Results and discussion

A comparison of the completeness scores of 69 proteomes calculated with BUSCO, CEGMA and DOGMA is shown in Figure 1. The completeness scores for the proteome data calculated with DOGMA are highly correlated to the ones calculated with BUSCO (Pearson's Product Moment Correlation Test gives a Correlation Coefficient of 0.912 and a *P*-value of 1.113E-027) and CEGMA (Correlation Coefficient: 0.890; *P*-value: 1.543E-024).

For a benchmark of the runtime compared between DOGMA and BUSCO, DOGMA was used in a combination with UProC (Meinicke, 2015) for domain annotation of the transcriptomes and proteomes. Figure 2 shows the runtime BUSCO and DOGMA need for analysis of the whole set of transcriptomes and proteomes. On average DOGMA was over 40 times faster than BUSCO on the transcriptome set. The maximum speedup achieved was more than 150 times faster. DOGMA was on average over 150 times faster on our proteome dataset, with a maximum improvement of over 400 times.

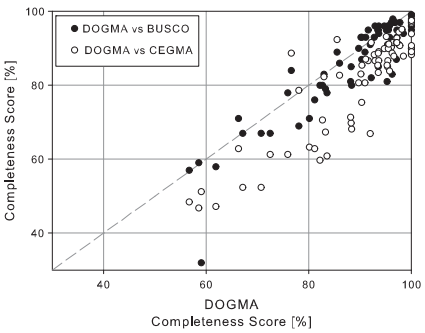

Fig. 1. Completeness scores of 69 proteomes compared between BUSCO, CEGMA and DOGMA. In all three programs the most specific core set available was applied. Pearson's Product Moment Correlation Test gives a Correlation Coefficient of 0.912 and a *P*-value of 1.113E-027 for DOGMA versus BUSCO completeness scores and a Correlation Coefficient of 0.890 with a P-value of 1.543E-024 for DOGMA versus CEGMA scores

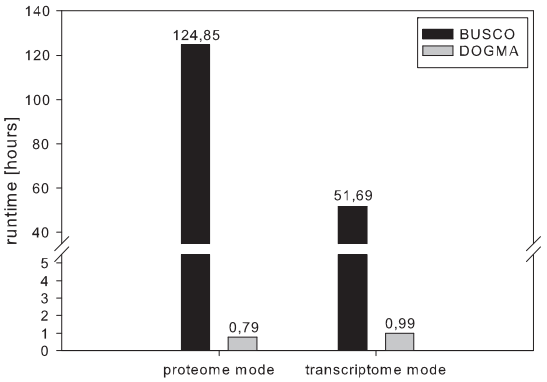

Fig. 2. Benchmark of DOGMA versus BUSCO runtime for analysis of the complete set of 96 transcriptomes and 69 proteomes. The runtime is shown in hours (times for annotation via UProC are included for the DOGMA runtime). The tools ran with one core on a machine with 12 cores and 64 GB RAM in total

When using the PfamScan annotation, completeness scores of DOGMA show a slightly higher correlation to BUSCO scores (see Supplementary Fig. S1) compared to the UProC annotation, but the analysis runs slower than BUSCO (see Supplementary Fig. S2). The completeness scores of the transcriptome dataset calculated by BUSCO and DOGMA are highly correlated as well (Correlation Coefficient: 0.914; *P*-value: 1.293E-038), as can be seen in Figure 3, while DOGMA runs much faster on the same dataset (cp. Fig. 2). A comparison with CEGMA is not possible on these data, because CEGMA has only been designed to work with genomes.

DOGMA completeness scores are similar to completeness scores calculated by BUSCO or CEGMA and represent a suitable measure for the quality of transcriptome and proteome data. Nevertheless, the completeness scores of DOGMA in transcriptome mode are generally higher than those of BUSCO. A similar effect can be seen on the simulated data (Supplementary Fig. S4). With increased error rate the completeness scores of both BUSCO and DOGMA decrease, but this trend is stronger in BUSCO. A reason for this higher scoring could be that protein domains are generally more conserved than their corresponding primary sequences (Fang *et al.*, 2013) and detection of the domains should be easier than sequence based orthology detection (Remmert *et al.*, 2011). Using protein domains would therefore result

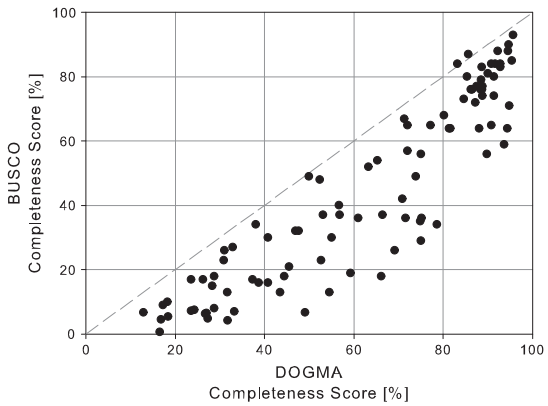

Fig. 3. Completeness scores of 96 transcriptomes compared between BUSCO and DOGMA. In both programs the most specific core set available was applied. Pearson's Product Moment Correlation Test gives a Correlation Coefficient of 0.914 and a *P*-value of 1.293E-038

in a more complete picture regarding completeness of the results because of fewer missed measurement units in the sample data (CDAs versus single-copy orthologs/conserved genes). Another and on the real data possibly more important point is that the three programs have different ways to optimize the scoring for specific species groups. CEGMA changes the allowed intron sizes when applied with optimizations for vertebrates or mammals. DOGMA and BUSCO on the other hand provide different core sets for different clades, but not exactly for the same clades. A comparison of completeness scores of the same proteomes based on different core sets with DOGMA can be found in Supplementary Figure S3. Core sets are seen as more specific for an organism in comparison to another core set, if they contain just species that belong to a group at a lower taxonomic level. We have shown that core set specificity has an impact on the completeness score (see Supplementary Fig. S3). Additionally, it can be expected that, the more specific a core set is for the sample organism, the more clade specific domain arrangements are included in the analysis and with that the score can be seen as more realistic for the sample organism. Therefore, for best results, it is recommended to choose an appropriate core set for the sample data. For this purpose, next to the provided core sets, DOGMA offers the possibility to test against self-made core sets. That can be useful for functional studies and analyses of evolutionary relationships by testing candidates against related taxa, getting information about missing domains in particular lineages. To ensure that the core set is representative for the whole clade, we performed a leave-one-out test. After the removal of one species from the core set, the completeness scores change only very little (Supplementary Fig. S5).

Additionally, we checked on the Ensembl dataset together with the eukaryotic core set, whether changing the maximum CDA size has strong influence on the completeness score (see Supplementary Figs S6 and S7). The results show that the CDA size has little influence, but the influence seems to be stronger when using DOGMA in combination with UProC. In most cases the completeness score seems to drop with an increase in CDA size, especially for proteomes of lower quality. This can be expected, as likelihood of missing a CDA increases with the increase of arrangement size.

4 Conclusion

The results presented here show that DOGMA achieves similar completeness scores as existing programs, but is able to run much faster when it is used in combination with a fast domain annotation tool such as UProC. Furthermore, the use of protein domains represents a less biased approach because HMMs are more sensitive. Therefore, HMMs are still able to detect domains even if the primary sequence has gained a lot of mutations. When running DOGMA with UProC, the order of domains cannot be taken into account (opposing to using PfamScan for the annotation). However, the completeness scores produced with both methods do not differ much and, due to the huge running time advantage and the possibility to apply UProC directly to transcriptomes, the usage of UProC is generally recommended.

Another advantage compared to other quality assessment methods is that DOGMA offers straightforward additional information about missing protein domains in the dataset that could be used for functional analyses. In conclusion, DOGMA combined with PfamScan annotations should be preferred for functional analyses, because of the additional information about the order of domains in domain arrangements. UProC is recommended for quality assessments due to the extreme running time advantage at nearly the same quality. Finally, the results presented above show that the choice of an appropriate core set is important for the analysis. Both DOGMA and BUSCO provide different core sets. However, with the exception of an eukaryotic core set, the core sets are based on different clades. The test on the eukaryotic core set shows that the score is robust to slight changes in the core set as well as changes in the CDA size. Note that the tools compared here cannot compete in every type of analysis, because they are partly designed for different tasks. DOGMA does not allow to analyze whole genomes like CEGMA and BUSCO. On the other hand CEGMA cannot be used to analyze transcriptome or proteome assemblies which is possible with DOGMA or BUSCO.

We did not compare our program to the PQI because a stand alone program is not available. However, we believe that our approach has several advantages over the domain metrics that are part of PQI. Presumably a simple comparison of the number of domains is not useful because domain types, selected arrangements for specific clades, and variation in frequency for particular domains are important. Since the comparison of domain arrangements for the PQI is based on CEGs, it is limited to eukaryotes and thus to a relatively low number of domains. Moreover, there is no option to choose a more specific core set for the PQI and therefore differences regarding the domain content that might be of interest are lost.

To conclude, DOGMA introduces a new efficient quality assessment, which is based on conserved protein domains and domain arrangements, that is qualitatively comparable to other methods, while DOGMA outperforms other programs on a great scale if combined with a fast annotation tool.

Acknowledgements

We want to thank Sören Perrey and Oliver Niehuis for useful comments and feedback. Erich Bornberg-Bauer's ORCID is 0000-0002-1826-3576. ED was funded by DFG grant BO 2544/7-2.

Conflict of Interest: none declared.

References

Bitard-Feildel,T. *et al.* (2015) Domain similarity based orthology detection. *BMC Bioinformatics*, **16**, 154.

Clark,S.C. *et al.* (2013) ALE: A generic assembly likelihood evaluation framework for assessing the accuracy of genome and metagenome assemblies. *Bioinformatics*, **29**, 435–443.

Cunningham,F. *et al.* (2014) Ensembl 2015. *Nucleic Acids Res.*, **43**, D662–D669.
Ekman,D. *et al.* (2005) Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. *J. Mol. Biol.*, **348**, 231–243.
Ekman,D. *et al.* (2007) Quantification of the elevated rate of domain rearrangements in metazoa. *J. Mol. Biol.*, **372**, 1337–1348.
Fang,H. *et al.* (2013) A daily-updated tree of (sequenced) life as a reference for genome research. *Sci. Rep.*, **3**, 2015.
Finn,R.D. *et al.* (2014) Pfam: the protein families database. *Nucleic Acids Res.*, **42**, D222–D230.
Forslund,K. and Sonnhammer,E.L.L. (2008) Predicting protein function from domain content. *Bioinformatics*, **24**, 1681–1687.
Forslund,K. and Sonnhammer,E.L.L. (2012) Evolution of protein domain architectures. *Methods Mol. Biol. (Clifton, N.J.)*, **856**, 187–216.
Gough,J. *et al.* (2001) Assignment of homology to genome sequences using a library of hidden Markov models that represent all proteins of known structure. *J. Mol. Biol.*, **313**, 903–919.
Gurevich,A. *et al.* (2013) QUAST: quality assessment tool for genome assemblies. *Bioinformatics*, **29**, 1072–1075.
Kummerfeld,S.K. and Teichmann,S. a. (2009) Protein domain organisation: adding order. *BMC Bioinformatics*, **10**, 39.
Marsh,J.A. and Teichmann,S.A. (2010) How do proteins gain new domains? *Genome Biol.*, **11**, 126.
Meinicke,P. (2015) UProC: tools for ultra-fast protein domain classification. *Bioinformatics*, **31**, 1382–1388.
Moore,A.D. and Bornberg-Bauer,E. (2012) The dynamics and evolutionary potential of domain loss and emergence. *Mol. Biol. Evol.*, **29**, 787–796.
Moore,A.D. *et al.* (2008) Arrangements in the modular evolution of proteins. *Trends Biochem. Sci.*, **33**, 444–451.
Parra,G. *et al.* (2007) CEGMA: A pipeline to accurately annotate core genes in eukaryotic genomes. *Bioinformatics*, **23**, 1061–1067.
Remmert,M. *et al.* (2011) HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. *Nat. Methods*, **9**, 173–175.
Sardar,A.J. *et al.* (2014) The evolution of human cells in terms of protein innovation. *Mol. Biol. Evol.*, **31**, 1364–1374.
Simão,F.A. *et al.* (2015) BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. *Bioinformatics*, **31**, 9–10.
Tatusov,R.L. *et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**, 41.
Terrapon,N. *et al.* (2014) Rapid similarity search of proteins using alignments of domain arrangements. *Bioinformatics*, **30**, 274–281.
Vogel,C. *et al.* (2005) The relationship between domain duplication and recombination. *J. Mol. Biol.*, **346**, 355–365.
Waterhouse,R.M. *et al.* (2013) OrthoDB: A hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Res.*, **41**, 358–365.
Yandell,M. and Ence,D. (2012) A beginner's guide to eukaryotic genome annotation. *Nat. Rev. Genet.*, **13**, 329–342.
Zaucha,J. *et al.* (2015) A proteome quality index. *Environ. Microbiol.*, **17**, 4–9.