DOGMA: Domain-Based Transcriptome and Proteome Quality Assessment

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

Introduction

The quality of genome or transcriptome assemblies can vary a lot[1]. Therefore, quality assessment of assemblies and annotations are crucial aspects of genome analysis pipelines.

DOGMA[2] is a program to assess the quality of a proteome or transcriptome based on conserved protein domains that act as a representative quality indicator for the whole assembly.

Protein domains are independently evolving structural and functional building blocks of proteins, known to be well conserved across taxa[3]. Domain arrangements are specified by the order of protein domains in an amino acid sequence[4].

Quality assessments so far have mostly used gene based approaches, although domains and not genes are the independently evolving units.

In contrast to gene-based methods (used in tools like BUSCO[5]), protein domains as very conserved sequence motifs can be better characterized by e.g. Hidden Markov Models (HMMs) and easier detected even in very diverged sequences[6].

In combination with our newly developed tool RADIANT (RApid DomaIn ANnoTation) for fast annotation of sequences with Pfam domains[7], sequence data can be analyzed with DOGMA very fast, at nearly the same quality as with the original PfamScan.

Materials & Methods

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.

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 (BUSCO[5]).

References

[1] Fang H et al., A daily-updated tree of (sequenced) life as a reference for genome research. Sci. Rep., 3, 2015, 2013

[2] Dohmen E et al., DOGMA: Domain-based transcriptome and proteome quality assessment. Bioinformatics (Oxford, England) page btw231, 2016.

[3] Ekman D et al., Multi-domain proteins in the three kingdoms of life: orphan domains and other unassigned regions. Journal of molecular biology, 348(1):231-43, 2005

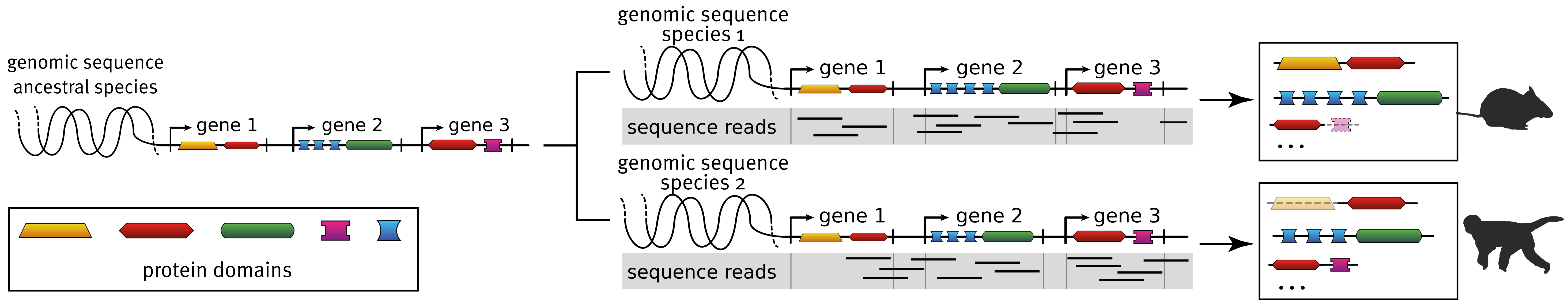
[4] Forslund K and Sonnhammer ELL, Evolution of Protein Domain Architectures. Methods in molecular biology (Clifton, N.J.), 856:187-216, 2012

[5] Simão FA et al., BUSCO: assessing genome assembly and annotation completeness with single-copy orthologs. Bioinfomatics, pages 9-10, 2015

[6] Remmert M et al., HHblits: lightning-fast iterative protein sequence searching by HMM-HMM alignment. Nature Methods, 9(2):173-175, 2011

[7] Finn RD et al., The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 44 (D1):D279–85, 2016

Protein domains are structually and evolutionary conserved units



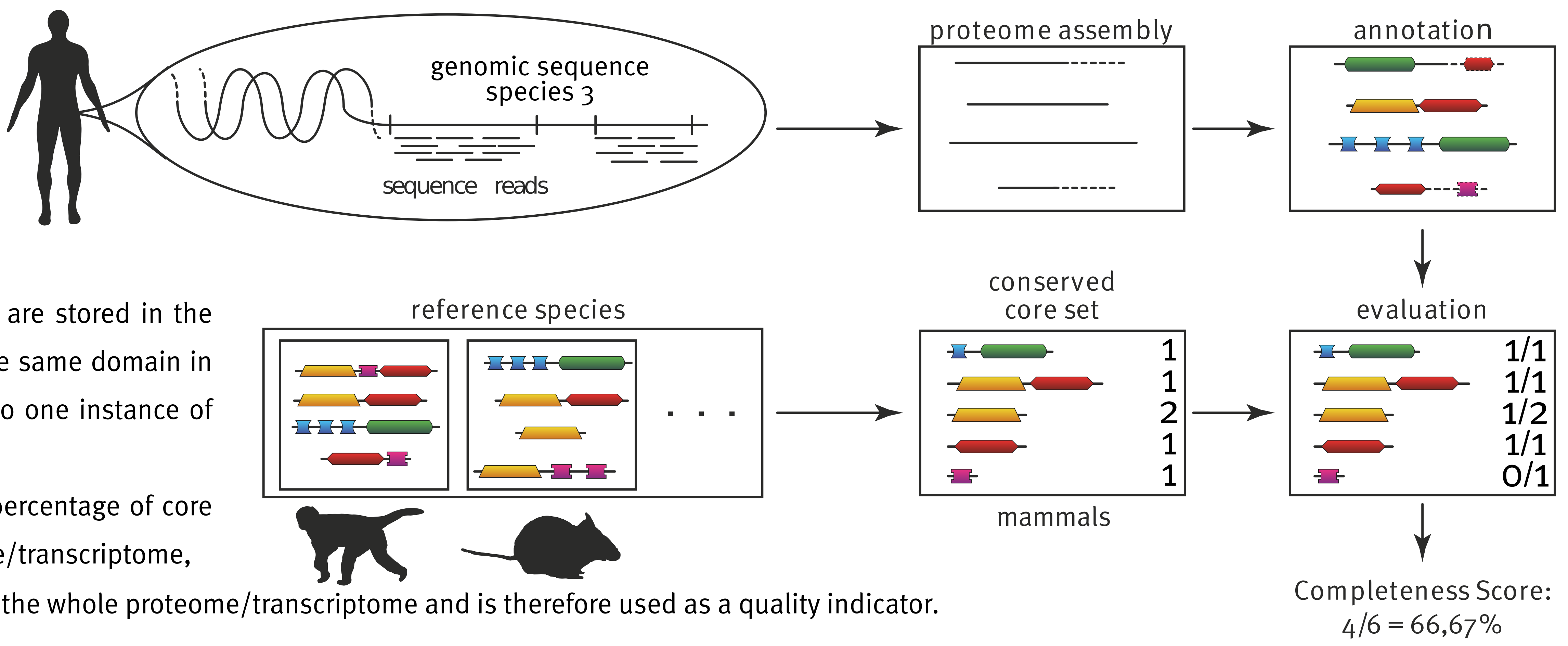
While genomes evolve over time and the gene content changes, a bigger part is conserved and can be found in all related species. Such a core set of well conserved evolutionary units is a good candidate to infer the quality of proteome or transcriptome assemblies.

Domains are encoded in protein coding genes and represent independent evolutionary units within these, while being very conserved and thus easy to detect.

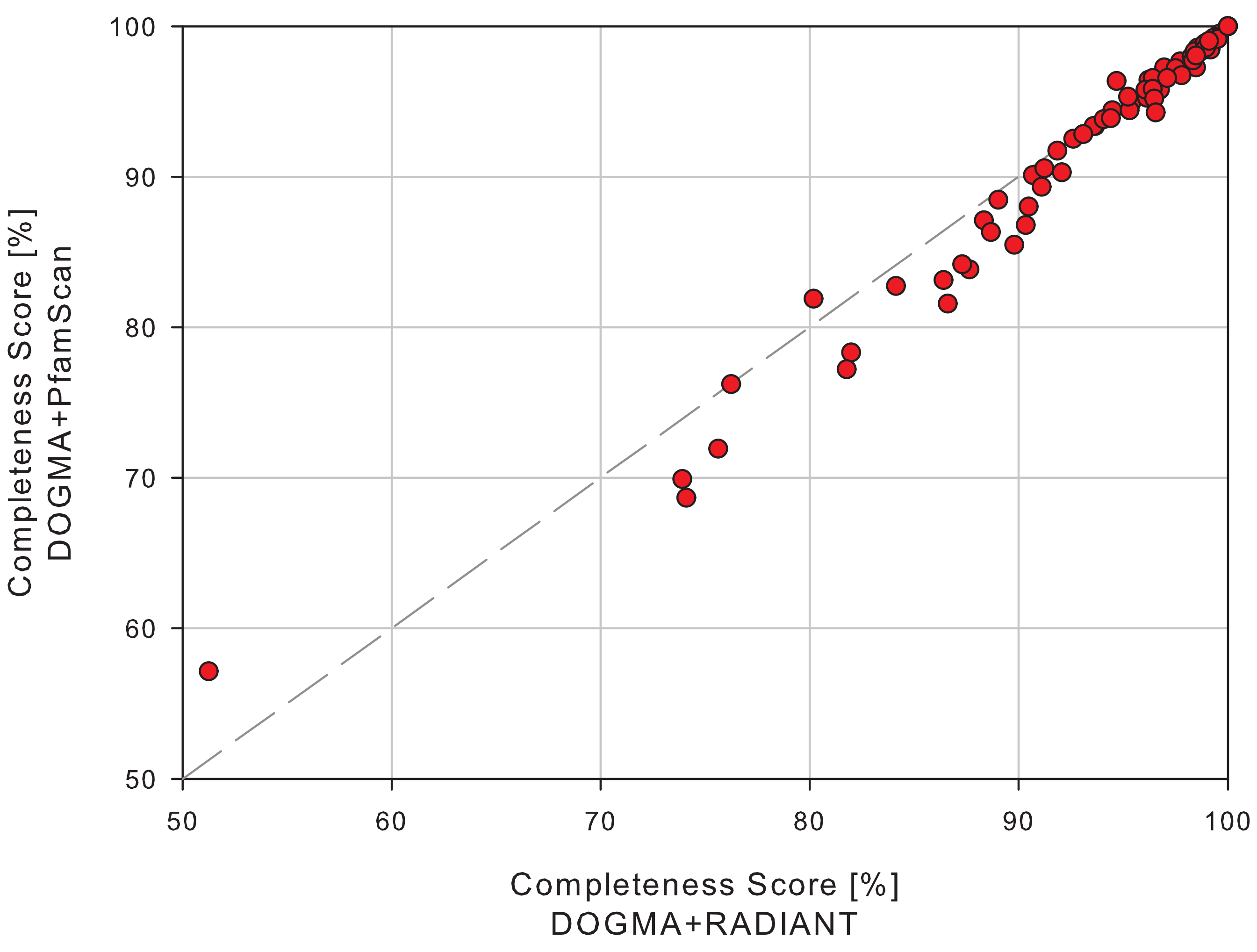
DOGMA compares a proteome or transcriptome to a conserved core set of domain arrangements

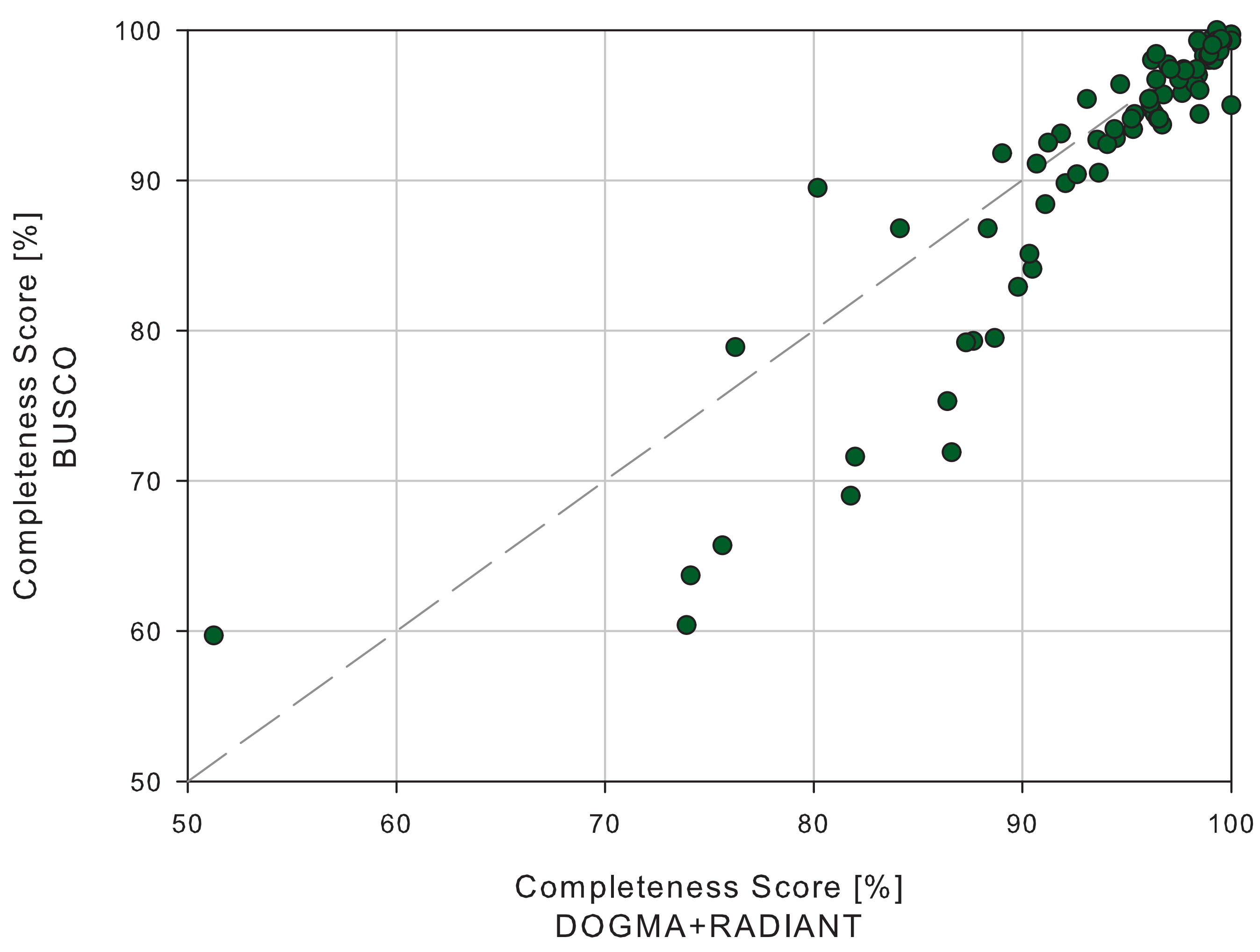
A core set of Conserved Domain Arrangements (CDAs) is used in DOGMA to be compared against the transcriptome or proteome of interest (species 3). CDAs that appear in all reference species and occur at a similar frequency, are stored in the core set. Consecutive repeats of the same domain in an arrangement will be collapsed to one instance of the domain (A-A-A-B → A-B).

The measured completeness, the percentage of core CDAs found in the sample proteome/transcriptome, is assumed to be representative for the whole proteome/transcriptome and is therefore used as a quality indicator.



DOGMA computes with very short runtime completeness scores similar to established methods





The most time consuming step for quality assessment with DOGMA is the domain annotation. For this purpose we developed RADIANT, a program to rapidly annotate Pfam domains in sequence data. Our tests show that DOGMA gives similar completeness scores in combination with RADIANT as in combination with the original PfamScan. Furthermore, the completeness scores are comparable to BUSCO scores. For this comparison, completeness scores were calculated for a data set of 84 proteomes of different quality. The scores of all tools are based on the respective default eukaryotic core set.

Conclusion

- DOGMA achieves similar completeness scores as existing programs
- DOGMA is able to run very fast when it is used in combination with a fast annotation tool such as RADIANT
- the use of protein domains represents a less biased approach because of their high conservation level
- DOGMA offers straightforward information about missing CDAs in the sample proteome/ transcriptome that can be used for functional analyses

DOMAINWORLD

All programs shown here are part of our software suite "DomainWorld", accessible at: http://domainworld.uni-muenster.de/

[1]Molecular Evolution and Bioinformatics Group, Institute for Evolution and Biodiversity, Münster, Germany

[2]Institute for Bioinformatics and Chemoinformatics, Westphalian University of Applied Sciences, Recklinghausen, Germany

WWU MÜNSTER

bornberglab.org
Molecular Evolution and Bioinformatics